QU'EST-CE QUE

LE TRAVAIL SCIENTIFIQUE DES DONNÉES ?

Big data, little data, no data

CHRISTINE L. BORGMAN

Qu'est-ce que le travail scientifique des données ?

Big data, little data, no data

Christine L. Borgman

Traducteur : Charlotte Matoussowsky

DOI : 10.4000/books.oep.14692

Éditeur : OpenEdition Press

Année d'édition : 2020

Date de mise en ligne : 18 décembre 2020

Collection : Encyclopédie numérique

EAN électronique : 9791036565410

<http://books.openedition.org>

Édition imprimée

EAN (Édition imprimée) : 9791036550386

Nombre de pages : 420

Référence électronique

BORGMAN, Christine L. *Qu'est-ce que le travail scientifique des données ? Big data, little data, no data.*

Nouvelle édition [en ligne]. Marseille : OpenEdition Press, 2020 (généré le 25 juin 2021). Disponible sur Internet : <<http://books.openedition.org/oep/14692>>. ISBN : 9791036565410. DOI : <https://doi.org/10.4000/books.oep.14692>.

© OpenEdition Press, 2020

Creative Commons - Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International
- CC BY-NC-ND 4.0

QU'EST-CE QUE

LE TRAVAIL SCIENTIFIQUE DES DONNÉES ?

La traduction de cet ouvrage a été subventionnée par le ministère
de l'Enseignement supérieur, de la Recherche et de l'Innovation
dans le cadre du Plan national pour la science ouverte

**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION**

*Liberté
Égalité
Fraternité*

CHRISTINE L. BORGMAN

QU'EST-CE QUE

LE TRAVAIL SCIENTIFIQUE DES DONNÉES ?

Big data, little data, no data

Traduction de Charlotte Matoussowsky

En couverture : *Abstract technology big data background concept*

© your123 - stock.adobe.com

Conception graphique : Veronica Holguín

Suivi et coordination éditoriale : Julie Pynte, Amandine Texier et Cédric Gaultier

Traduction : Charlotte Matoussowsky

Correction : Solenne Louis

Mise en page grâce à Métopes. Méthodes et outils pour l'édition structurée.

Cet ouvrage est la traduction française de *Big Data, Little Data, No Data. Scholarship in the Networked World*, paru en 2015 chez MIT Press.

Collection « Encyclopédie numérique », 2020

Cet ouvrage est en ligne en accès ouvert :

<http://books.openedition.org/oep/14692>

Texte : Licence Creative Commons Attribution – Pas d'utilisation commerciale

Pas de modification 4.0 International

<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.fr/>

ISBN papier : 979-10-365-5038-6

ISBN électronique : 979-10-365-6541-0

Note de la traductrice

Big Data, Little Data, No Data est l'œuvre d'une professeure américaine, rédigée dans un contexte essentiellement anglophone. Sa traduction – comme toute traduction – a nécessité un travail d'adaptation de la part de l'équipe éditoriale d'OpenEdition afin de transmettre au lectorat francophone les idées de ce texte dans une langue la plus précise et la plus claire possible.

Ce travail résulte d'une somme de choix, qu'il serait fastidieux de tous expliquer ici : au besoin, certains seront explicités au fil du texte au moyen d'une note de la traductrice (NdT).

Il nous paraît cependant utile d'expliquer ici comment nous avons retranscrit en français le concept fondamental de *data scholarship*, afin de ne pas ennuyer le lectorat avec de trop nombreuses explicitations et notes au fil du texte. Cette explication vient se substituer à un court passage que nous avons supprimé du texte original¹ qui, parce qu'il expliquait des concepts anglophones en des termes anglophones, ne pouvait être traduit littéralement sans trahir soit la pensée de Christine L. Borgman, soit la langue française.

Le terme *data scholarship* est composé des mots *data*, « données », et *scholarship*. Ce dernier désigne l'activité de la ou du *scholar*, c'est-à-dire du savant ou de la savante en quête d'érudition, ainsi que le savoir qui résulte de cette activité. Quasi-synonyme de *research*, il lui est toutefois préféré dans le contexte des sciences humaines. Par ailleurs, la *scholarship* ne se limite pas strictement aux expériences scientifiques, à la recherche d'informations et aux autres activités permettant de construire son propre savoir, mais peut aussi englober la *scholarly communication*, c'est-à-dire la communication formelle et informelle entre scientifiques, ou communication savante, qui comprend entre autres l'édition scientifique, l'évaluation par les pairs (*peer review*), les conférences ou encore les discussions informelles. En français, le terme « recherche », riche de connotations variées touchant aussi bien au travail scientifique qu'à la carrière, nous a paru le plus adéquat à rendre cette notion.

Cependant, la *data scholarship* n'est pas une recherche sur les données, au sens où celles-ci seraient le point d'intérêt des scientifiques, mais plutôt une recherche au moyen des données. Elle est un ensemble de pratiques intégrées dans le travail

1. La lectrice ou le lecteur curieux pourra retrouver le paragraphe concerné au chapitre 3, p. 31-32 de l'édition originale parue chez MIT Press.

quotidien des chercheurs et chercheuses, souvent sans qu'ils en aient conscience et sans que ce labeur soit remarqué, ce qui pousse Christine L. Borgman à le qualifier de « travail invisible ». Ce réseau de significations nous a conduits, avec l'accord de l'autrice, à traduire *data scholarship* par « travail scientifique des données », indiquant par là la transformation opérée sur les données par les acteurs de la recherche.

Par convention, nous avons choisi d'utiliser dans cet ouvrage les termes « sciences exactes » (que nous traduisons de l'anglais américain *sciences*) pour désigner les sciences naturelles, expérimentales et formelles, par opposition aux sciences sociales d'une part et aux sciences humaines d'autre part. L'idée n'est évidemment pas de suggérer que ces dernières seraient « inexactes ». Nous aurions pu recourir à l'acronyme STM (sciences, techniques/technologies et médecine) mais il nous a semblé recouvrir une réalité un peu différente.

Sauf mention contraire, les citations anglophones ont été traduites par nos soins pour le présent ouvrage.

Enfin, parce que Christine L. Borgman a veillé à évoquer aussi bien les femmes que les hommes de science, nous nous sommes efforcés d'en faire autant tout au long de la traduction.

Charlotte Matoussowsky

Preface to the French translation of *Big Data, Little Data, No Data*

The origin of this translation into French was the invitation from Marin Dacos and Françoise Genova to keynote the 2018 Paris conference launching the National Open Science Plan for France, subtitled “from Strategy to Action.” This stimulating three-day event, organized in partnership with the Research Data Alliance, drew participants from France, Europe, and the U.S. to explore emerging issues around open access to publications, data, software, and other scientific research objects. Top-down approaches to open science, in the form of government science policies, met bottom-up approaches in the form of community practices. The ambitious National Plan combines country-specific actions with European initiatives: “France is committed to making scientific research results open to all – researchers, companies, citizens.”

In the two years since that launch event, open science continues to evolve in concept, policy, and practice. Notions of “openness,” whether referring to publications, data, software, or other entities, are as amorphous as ever. What is open to whom, when, why, and under what circumstances varies widely. Open access (OA) to publications is the usual starting point for open science, as is the case in the plan for France. OA publishing comes in many flavors, however. Some plans shift the costs to government funding agencies, some to universities, some to authors, and some to publishers. Other flavors promote preprint servers and institutional repositories as complements to paid subscriptions. Government initiatives such as *Plan S*, due to take effect in 2021, are controversial, resulting in complex compromises among stakeholders (Kwon, 2018; Noorden, 2020; *Plan S*, 2019).

The second pillar of most open science plans, including that of France, is open data. Access to research data, a central theme of this book, is yet more complex than OA publishing. Scholarly publishing has a long history, dating back millennia for books and centuries for journal articles. Data also have long histories, but more as process than as scholarly products to be exchanged. Research data can be embodied in artifacts, but they also can be abstractions or simulations. Almost any entity can be used as evidence of some phenomena. One person’s signal is another’s noise. Humans are in the loop throughout the entire lifecycle of data, from creation to curation to decay or disposal (Borgman, 2019).

In the time since this book's original publication in 2015, stakeholders have come to acknowledge the messiness of openness (Aspesi and Brand, 2020). While most disciplines now accept open access publishing, the means and adoption rates vary by domain, funding source, country, and other factors. More stakeholders recognize the messiness of scientific data, with advances in research on philosophical, epistemological, social, technical, and cultural aspects of data practices (Borgman, 2019; Drucker, 2014; Lane *et al.*, 2020; Leonelli, 2019a, 2019b; Pasquetto *et al.*, 2019; Rosenberg, 2018). The FAIR principles, first promulgated in 2016 (Wilkinson *et al.*, 2016), established a framework for disseminating data openly. These principles were quickly adopted into European science policy (European Union Publications Office, 2018). Practical, on-the-ground efforts to implement FAIR reveal the aspirational nature of the enterprise. For data to be *Findable*, communities must agree on methods for description, search, and retrieval. For data to be *Accessible* and *Interoperable*, stakeholders must agree on technical and legal frameworks. For data to be *Reusable*, originators must provide access to adequate documentation, and often to associated software, instrumentation, and other technologies. Reusing data is a goal more achievable than reproducible or replicable science, all of which remain contested concepts. The FAIR principles are subject to temporal factors. The longer the time from origin, the more difficult data are to find, access, interoperate, or reuse. Data creators retain significant advantages in the ability to reuse research data (Pasquetto *et al.*, 2017, 2019).

Several new journals devoted to interdisciplinary investigations of data have launched in the last five years, such as the *Harvard Data Science Review*, *Scientific Data*, and the *Journal of Data and Information Science*, plus countless special issues of discipline-specific journals. Interest in software preservation, software citation, and data citation continues to grow (Bouquin *et al.*, 2020; Davenport *et al.*, 2020; Smith *et al.*, 2016; *Software Heritage Foundation*, 2019; Wofford *et al.*, 2020). These venues, plus conferences, government reports, and funding initiatives serve to broaden the conversation about research data. Theoretical, technical, and practice topics abound, such as costs and benefits of data preservation, ethics and values of providing access to human subjects data, incentives and disincentives to share or reuse data, tradeoffs between launching new missions and preserving the data of current missions, how practices vary within and between domains, and whether data are best managed by universities, disciplinary repositories, government agencies, or commercial ventures, to name a few.

In parallel with the growth of research into data science is the expanding array of career tracks in data science, data management, and other areas of data practice. Universities in France, Europe, North America, Australasia, Asia, and elsewhere are investing in data science programs at the undergraduate and post-graduate levels. Some universities expect students in all fields to take at least one data science course as a

core requirement. These courses and degrees vary in theoretical, computational, and practical orientation. Many provide general knowledge in computing, statistics, digital scholarship, or related areas. Some are discipline-specific, such as bioinformatics. Yet other programs are professional, such as those in business, management, librarianship, and information sciences. Academic libraries and research institutions are hiring data management specialists. These data professionals play important roles in curating data, developing data archives, aiding researchers in managing their own data, and strategic planning for open science.

Perhaps the most significant advance in this time frame is broader recognition of the knowledge infrastructures (KI) in which scholarship occurs. As explained in Chapter 1 herein, KI encompass human, social, technical, policy, and institutional components of intellectual work and the many interactions between them. Infrastructures develop, evolve, and adapt in complex ways over long periods of time. To consider any component in isolation, whether data, publications, or individual systems, is to see but one feature of the elephant. Silos can emerge and become isolated. Components designed to interoperate seamlessly can become brittle, leading to catastrophic breakdowns. Robust networks can be self-healing, at least for awhile. Infrastructures are inherently fragile, requiring continual maintenance and regular repair. Durability is an accomplishment. Research data are especially fragile because they rarely stand alone. Lacking context, structure, or documentation, a dataset may be little more than a string of numbers. For data to be findable, accessible, interoperable, and reusable, myriad stakeholders must invest in KI components such as metadata, formats, archives, software preservation, data curation, skilled labor. In turn, these stakeholders must work together in ways that keep these KI components working together. As the array of players expands to include public, private, government, and niche partners, KI complexity increases. Infrastructures are notoriously invisible until they break down (Borgman *et al.*, 2020, 2016; Edwards, 2010; Edwards *et al.*, 2013; Mayernik *et al.*, 2017; Scroggins *et al.*, 2020; Scroggins & Pasquetto, 2020; Star & Ruhleder, 1996).

This new edition of *Big Data, Little Data, No Data: Scholarship in the Networked World* appears at a critical juncture in scholarship, open science, and knowledge infrastructures. Charlotte Matoussowsky is an exceptionally conscientious translator, peppering me with thoughtful questions throughout the process. May this volume provoke new conversations to advance our understanding, use, and reuse of research data for the generations to come.

Christine L. Borgman
UCLA, Los Angeles
November 13, 2020

Préface à l'édition française

La présente traduction française trouve son origine dans une invitation à intervenir au colloque inaugural du Plan national pour la science ouverte à Paris en 2018, sous-titré « De la stratégie à l'action », que m'avaient adressée Marin Dacos et Françoise Genova. Ces trois jours de conférences stimulantes, organisés en partenariat avec la Research Data Alliance, ont attiré des participants venus de France, d'Europe et des États-Unis pour discuter des problématiques émergentes autour de l'accès ouvert des publications, des données, des logiciels et des autres objets de recherche. Les logiques descendantes, à savoir les politiques scientifiques gouvernementales, s'y sont confrontées aux démarches ascendantes des pratiques disciplinaires. L'ambitieux Plan national pour la science ouverte combine des actions propres à la France et des initiatives européennes : « La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens ».

Dans les deux années qui ont suivi, la science ouverte a continué d'évoluer au point de vue conceptuel, stratégique et pratique. La notion d'« ouverture », qu'elle se réfère aux publications, aux données, aux logiciels ou autres entités, est plus indéfinissable que jamais. Qu'est-ce qui est ouvert, à qui, quand, pourquoi et à quelles conditions : tout cela est très variable. L'*open access* ou accès ouvert (AO) aux publications constitue souvent le point de départ des démarches en faveur de la science ouverte, comme en France avec le Plan national. L'édition en AO connaît cependant de nombreuses variantes. Certains projets en font supporter le coût par des organismes de financement étatiques, d'autres par les universités, d'autres par les auteurs et autrices et d'autres par les maisons d'édition. D'autres variantes promeuvent les serveurs de prépublications (*preprint*) et les référentiels institutionnels en complément d'abonnements payants. Les initiatives gouvernementales, telles que le *Plan S* qui doit entrer en vigueur en 2021, sont controversées et aboutissent à des compromis complexes entre les parties prenantes (Kwon, 2018 ; Noorden, 2020 ; *Plan S*, 2019).

Le deuxième pilier de la plupart des projets de science ouverte, y compris français, est l'*open data*, ou les données ouvertes. L'accès aux données de recherche, qui constitue un des principaux thèmes du présent ouvrage, est plus complexe encore que l'édition en AO. La publication scientifique est l'héritière d'une longue histoire, vieille de plusieurs millénaires pour les livres et de plusieurs siècles pour les articles de revue. Les données aussi existent depuis longtemps, mais ont longtemps été un processus plutôt qu'un produit scientifique échangeable. Les données de la recherche peuvent être des artefacts matériels, mais aussi des abstractions ou des

simulations. Pratiquement n'importe quelle entité peut être utilisée pour mettre en évidence des phénomènes. Ce qui représente un signe pour les uns est du bruit pour les autres. Les êtres humains interviennent tout au long du cycle de vie des données, de leur création à leur abandon ou leur destruction en passant par leur conservation (Borgman, 2019).

Lors de la première publication de ce livre en 2015, les parties prenantes avaient fini par reconnaître le caractère désordonné de l'ouverture (Aspesi et Brand, 2020). Si la plupart des disciplines acceptent aujourd'hui l'édition en accès ouvert, les moyens et les rythmes d'adoption varient en fonction des domaines, des financements, des pays et d'autres facteurs. La majorité des acteurs admettent que le désordre est aussi une caractéristique des données scientifiques; la recherche sur les pratiques en la matière a d'ailleurs fait de grands progrès sur le plan philosophique, épistémologique, social, technique et culturel (Borgman, 2019; Drucker, 2014; Lane *et al.*, 2020; Leonelli, 2019a, 2019b; Pasquetto *et al.*, 2019; Rosenberg, 2018). Les principes FAIR, promulgués en 2016 (Wilkinson *et al.*, 2016), ont instauré un cadre pour la diffusion ouverte des données. Ils ont été rapidement intégrés dans les politiques scientifiques européennes (European Union Publications Office, 2018). Les efforts concrets pour mettre en œuvre FAIR sur le terrain ont révélé l'ambition de ce projet. Pour que les données soient *trouvables* (*Findable*), les communautés disciplinaires doivent convenir de méthodes de description et de recherche. Pour qu'elles soient *Accessibles* et *Interopérables*, les parties prenantes doivent s'accorder sur le cadre technique et légal. Pour qu'elles soient *Réutilisables*, leurs créateurs et créatrices doivent fournir une documentation adéquate, ainsi qu'un accès aux logiciels, à l'instrumentation et aux autres technologies qui leur sont souvent associées. La réutilisation des données est un but plus accessible que la science reproductible ou répliquable, mais toutes restent des concepts contestés. Les principes FAIR sont soumis à des facteurs temporels. Plus long est le laps de temps qui s'est écoulé depuis l'origine des données et plus il est difficile de les trouver, d'y accéder, de les rendre interopérables ou de les réutiliser. Les créateurs et créatrices de données conservent un avantage significatif en matière de réutilisation des données de la recherche (Pasquetto *et al.*, 2017, 2019).

Ces cinq dernières années, de nouvelles revues consacrées à la recherche interdisciplinaire sur les données ont été créées, comme *Harvard Data Science Review*, *Scientific Data* et *Journal of Data and Information Science*, sans compter les innombrables numéros spéciaux de revues disciplinaires. L'intérêt pour la préservation des logiciels, la citation de logiciels et la citation de données ne cesse de grandir (Bouquin *et al.*, 2020; Davenport *et al.*, 2020; Smith *et al.*, 2016; Software Heritage Foundation, 2019; Wofford *et al.*, 2020). Ces médias, auxquels s'ajoutent les colloques, les rapports gouvernementaux et les projets de financement, permettent

d'élargir le débat autour des données de la recherche. Les sujets abondent, qu'ils soient théoriques, techniques ou pratiques, comme les coûts et bénéfices de la préservation des données, l'éthique et la valeur de l'accès aux données sur les sujets humains, les facteurs incitatifs et dissuasifs au partage ou à la réutilisation des données, les compromis à trouver entre le lancement de nouveaux projets et la préservation des données des projets existants, la variété des pratiques au sein des domaines et entre eux et la gestion des données par les universités, les référentiels disciplinaires, les administrations et les sociétés commerciales, pour ne citer que quelques acteurs.

Parallèlement à l'essor de la recherche en sciences des données, les possibilités de carrières dans la science, la gestion et d'autres domaines des pratiques en matière de données se sont élargies. Les universités de France, d'Europe, d'Amérique du Nord, d'Australasie, d'Asie et d'ailleurs investissent dans des programmes en sciences des données aux niveaux licence et master. Certaines universités exigent de leurs étudiantes et étudiants qu'ils suivent au moins un cours en datalogie dans leur cursus. L'orientation théorique, informatique et pratique de ces cours et diplômes est variée. Beaucoup enseignent des connaissances générales en informatique, statistique, recherche numérique ou autres. Certains sont propres à une discipline, comme la bio-informatique. D'autres sont professionnalisants, comme ceux dispensés dans les cursus de commerce, de management, de documentation et de sciences de l'information. Les bibliothèques universitaires et les instituts de recherche embauchent des spécialistes de la gestion de données. Ces professions jouent un rôle important dans la conservation, le développement d'archives, l'assistance aux chercheurs et chercheuses dans la gestion de leurs propres données et la planification stratégique de la science ouverte.

L'avancée la plus significative de cette période est sans doute la reconnaissance dont se sont mises à bénéficier les infrastructures de la connaissance (IC) où se construit la recherche. Comme nous l'expliquons dans le chapitre 1 de ce livre, les IC recouvrent les composantes humaines, sociales, techniques, stratégiques et institutionnelles du travail intellectuel et leurs nombreuses interactions. Les infrastructures se développent, évoluent et s'adaptent de façon complexe sur de longues périodes. Considérer l'une de ces composantes isolément, que ce soient les données, les publications ou les systèmes individuels, revient à ne voir qu'une partie du corps de l'éléphant. Des silos peuvent émerger et se retrouver isolés. Des composantes conçues pour interagir avec fluidité peuvent se fragiliser et mener à des pannes catastrophiques. Des réseaux robustes peuvent se réparer eux-mêmes, au moins pour un temps. Les infrastructures sont par essence fragiles et nécessitent un entretien continu et des réparations régulières. La durabilité est un exploit. Les données de la

recherche sont particulièrement fragiles, parce qu'elles se suffisent rarement à elles-mêmes. Faute de contexte, de structure ou de documentation, un jeu de données peut n'être guère qu'une série de chiffres. Pour que les données soient trouvables, accessibles, interopérables et réutilisables, des myriades d'acteurs doivent investir dans les composantes des IC, telles que les métadonnées, les formats, les archives, la préservation de logiciels, la conservation de données, la main-d'œuvre qualifiée... Ces acteurs doivent ensuite travailler ensemble pour que ces composantes continuent de fonctionner à l'unisson. À mesure que les parties prenantes se diversifient pour inclure des partenaires publics, privés, étatiques et de niche, la complexité des IC augmente. On le sait : les infrastructures sont invisibles jusqu'au moment où elles ne marchent plus (Borgman *et al.*, 2020, 2016; Edwards, 2010; Edwards *et al.*, 2013; Mayernik *et al.*, 2017; Scroggins *et al.*, 2020; Scroggins et Pasquetto, 2020; Star et Ruhleder, 1996).

Cette nouvelle édition de *Qu'est-ce que le travail scientifique des données ? Big data, Little data, No data* paraît alors que la recherche, la science ouverte et les infrastructures de la connaissance se trouvent à un carrefour. Charlotte Matoussowsky est une traductrice particulièrement consciencieuse, qui a émaillé nos échanges de questions judicieuses tout au long du processus. Puisse cet ouvrage susciter de nouveaux débats qui feront progresser notre compréhension, notre utilisation et notre réutilisation des données de la recherche pour les générations futures.

Christine L. Borgman
UCLA, Los Angeles
13 novembre 2020

Préface

Les *big data* suscitent aujourd'hui un vif intérêt, mais les *little data* sont tout aussi essentielles à la recherche scientifique. Plus le volume des données augmente, plus notre capacité à examiner des constatations individuelles diminue. L'observateur doit sans cesse s'écarter des phénomènes qu'il étudie, il lui faut de nouveaux outils et perspectives. Néanmoins, les mégadonnées ne sont pas nécessairement de meilleures données. Plus l'observateur s'éloigne de son point de départ, plus on éprouvera de difficultés à déterminer la signification de ses observations : comment elles ont été recueillies, manipulées, réduites et transformées, avec quels présupposés et quels objectifs. Les scientifiques préfèrent souvent travailler sur un petit nombre de données que l'on peut inspecter de près. Parfois, elles et ils ne disposent d'aucune donnée, car ces dernières n'ont pas été découvertes ou ne peuvent pas l'être.

Les données de la recherche sont à la fois bien plus que de simples marchandises à exploiter et bien moins. Plans de gestion, critères de diffusion et autres politiques pleines de bonnes intentions que proposent les organismes de financement, les revues et les instituts de recherche tiennent rarement compte de la diversité des données et des pratiques disciplinaires. Rares sont les politiques qui s'efforcent de définir la notion de « donnée » autrement que par une liste d'exemples. Plus rares encore sont celles qui répondent aux motivations intrinsèques et extrinsèques divergentes des acteurs de la recherche. Les données peuvent revêtir simultanément des dimensions très variées selon les individus. Elles peuvent être un atout à maîtriser, à accumuler, à négocier, à combiner, à explorer et, éventuellement, à divulguer. Elles peuvent constituer une charge qu'il faut gérer, protéger ou détruire. Elles peuvent être sensibles ou confidentielles et leur diffusion peut comporter de grands risques. Leur valeur peut ne pas apparaître immédiatement ou n'être comprise que bien plus tard. Certaines données valent la peine d'être conservées indéfiniment tandis que d'autres n'ont qu'une utilité éphémère. En effet, en quelques heures ou quelques mois, certains types d'observations perdent leur intérêt sous l'effet des avancées technologiques et scientifiques.

Pour comprendre le rôle des données dans la recherche scientifique, il faut d'abord admettre qu'elles ne sont pas des objets naturels, possédant une essence propre. En réalité, elles sont des représentations d'observations, d'objets ou d'autres entités qui servent à mettre en évidence des phénomènes à des fins de recherche. Ces représentations varient au fil du temps et en fonction des scientifiques et des circonstances. Dans l'ensemble des sciences exactes, humaines et sociales, la communauté

scientifique crée des données, les utilise, les analyse et les interprète, souvent sans s'accorder sur ce qu'elles sont. Concevoir quelque chose comme une donnée est, en soi, un acte scientifique. La recherche repose sur la preuve, sur l'interprétation et sur l'argumentation. Les données sont ainsi un moyen, dont la fin est le plus souvent l'article de revue, l'ouvrage, la communication de colloque ou une autre production digne de reconnaissance universitaire. Le chercheur ou la chercheuse travaille rarement dans l'idée de les réutiliser.

Galilée émaillait ses carnets de croquis. Les astronomes du ^{xix}^e siècle réalisaient des clichés sur des plaques de verre. Aujourd'hui, ces scientifiques captent des photons à l'aide d'instruments numériques. Si l'on peut mettre en correspondance des images du ciel étoilé photographiées par des appareils grand public avec celles prises par des missions spatiales, c'est parce que les astronomes ont convenu de représentations communes pour décrire et mapper les données. En effet, l'astronomie a investi massivement dans des normes, des outils et des archives afin de pouvoir rassembler des observations recueillies sur plusieurs siècles. L'infrastructure des connaissances astronomiques est cependant loin d'être achevée ou complètement automatisée. Les professionnels de l'information jouent un rôle clé dans l'organisation de l'accès aux données, qu'elles soient astronomiques ou autres.

Parce que les relations entre publications et données sont multiples, les données de la recherche gagnent à être examinées dans le cadre de la communication savante. Leur production peut se faire délibérément et sur le long terme, créant ainsi une mine de ressources dont la valeur va en augmentant. Elle peut se faire *ad hoc*, voire au petit bonheur, en se saisissant des indicateurs des phénomènes présents au moment de la collecte. Même si le protocole de recherche est parfaitement défini, qu'il s'agisse d'astronomie, de sociologie ou d'ethnographie, le recueil des données peut s'avérer aléatoire ; les découvertes faites à un stade influent sur le choix des données au stade suivant. Quel que soit le domaine, il faut, pour devenir scientifique, apprendre à évaluer les données, à déterminer leur fiabilité et leur validité et à s'adapter aux conditions de son laboratoire, de son terrain ou de ses archives. Les publications visant à exposer des découvertes situent celles-ci dans le contexte de la discipline, en se fondant sur l'expertise du lectorat. On y présente les informations utiles à la compréhension de l'argumentation, des méthodes et des conclusions. En revanche, les détails nécessaires à la reproduction de l'étude sont souvent omis, car on part du principe que la cible est familière des méthodes du domaine. Les notions de réplification et de reproductibilité, bien qu'elles constituent un argument courant en faveur de la publication de données, ne sont pertinentes que dans certaines disciplines et ardues à mettre en œuvre même en leur sein. Le plus difficile reste de déterminer quelles productions scientifiques méritent d'être préservées.

Les politiques de gestion, de diffusion et de partage des données occultent la complexité des rôles qu'elles occupent dans la recherche et ignorent largement la diversité des pratiques intradisciplinaires et interdisciplinaires. La notion de donnée varie grandement au sein des sciences exactes, humaines et sociales et à l'intérieur de chaque spécialité. Dans la plupart des disciplines, on apprend la gestion des données plus qu'on ne l'enseigne, ce qui conduit à recourir à des solutions *ad hoc*. Les chercheurs et chercheuses éprouvent d'ailleurs souvent de grandes difficultés à réutiliser leurs propres données. Les rendre exploitables par des inconnus est plus épineux encore. Le partage de données n'est la norme que dans de rares disciplines, car il est très délicat à mettre en œuvre et nécessite d'importants investissements dans l'infrastructure de la connaissance.

Le présent ouvrage est destiné au vaste lectorat que sont les parties prenantes des données de la recherche, parmi lesquelles les scientifiques, les chercheurs et chercheuses, les chefs et cheffes d'établissement, les organismes de financement, les maisons d'édition, les bibliothèques, les archives et les responsables politiques. La première partie dresse un panorama des données et de la recherche scientifique tout au long de quatre chapitres, conduisant à une analyse des notions de donnée, de recherche et d'infrastructure de la connaissance, ainsi que de la diversité des pratiques scientifiques. Dans la deuxième partie, nous examinons l'usage des données dans le cadre des sciences exactes, des sciences sociales et des sciences humaines à travers trois chapitres constitués d'études de cas, dont la structure parallèle permettra la comparaison. La dernière partie traite en trois chapitres des politiques et des pratiques en matière de données; nous nous y penchons sur l'origine des nombreux écueils auxquels se heurte le travail scientifique des données, parmi lesquels la diffusion, le partage et la réutilisation, l'attribution, le crédit et la découverte, et, enfin, les données à conserver et pourquoi.

L'histoire de la recherche et celle des données sont étroitement liées depuis longtemps. Ni l'une ni l'autre ne sont des notions nouvelles. Ce qui est nouveau, c'est la volonté d'extraire les données des processus scientifiques pour les exploiter à d'autres fins. Les coûts, les bénéfices, les risques et les récompenses associés à l'utilisation des données de la recherche sont en train d'être redistribués entre des acteurs en concurrence. L'objet de ce livre est de susciter un débat bien plus riche et plus éclairé entre ces parties prenantes. C'est l'avenir de la recherche qui est en jeu.

Christine L. Borgman
Los Angeles, Californie
Mai 2014

I. Données et recherche

I. Provocations

« La valeur des données réside dans leur usage. »
National Research Council, *Bits of Power*

Introduction

En 1963, dans une formule qui a fait date, Derek de Solla Price a opposé science légère et science lourde (ou mégascience), en anglais *little science* et *big science*. Weinberg (1961) avait récemment créé ce dernier terme pour désigner les grands travaux qu'une société humaine entreprend afin de réaliser ses aspirations. Les monuments du ^{xx}e siècle auxquels Weinberg faisait référence étaient d'énormes fusées, des accélérateurs de haute énergie et des réacteurs de recherche à large débit. C'étaient des « symboles de notre temps » comparables aux pyramides d'Égypte, au château de Versailles ou à Notre-Dame de Paris. C'était l'ère de Spoutnik, une époque où l'on déversait de vastes sommes d'argent dans l'entreprise scientifique. Price et Weinberg interrogeaient la trajectoire de la *big science*, se demandant quelles étaient les valeurs relatives des sciences lourdes et légères (Price), si la science lourde valait les sommes qu'on y investissait et même si celle-ci ne nuisait pas à la science en général (Weinberg).

Le *big data* jouit aujourd'hui du même statut hypertrophié que la *big science* il y a cinquante ans. Il fait la une des magazines *Science*, *Nature*, *The Economist* et *Wired*, ainsi que du *Wall Street Journal* et du *New York Times*, sans compter de nombreuses autres publications plus ou moins connues. Tout comme la mégascience devait nous révéler les secrets de l'univers, les mégadonnées sont censées mettre au jour les trésors enfouis du train binaire de la vie. Le *big data* est le pétrole des affaires d'aujourd'hui (Mayer-Schonberger et Cukier, 2013), le ciment des collaborations (Borgman, 2007) et une source de friction entre universitaires (Edwards *et al.*, 2011 ; Edwards, 2010).

Pourtant, les données ne coulent pas comme le pétrole, ne fixent pas comme le ciment et n'allument pas de feu par friction comme les allumettes. Leur valeur réside dans leur usage, d'où la rédaction du rapport *Bits of Power* (National Research Council, 1997). La question qu'il faut aujourd'hui poser est : « qu'est-ce qu'une donnée ? ». Le consensus nous dit qu'une définition unique ne saurait suffire. Les données revêtent toutes sortes de valeurs, qui ne peuvent parfois apparaître que longtemps après leur collecte, leur classement ou leur perte. La valeur des données varie largement en fonction des lieux, des époques et des contextes. Détenir les bonnes données est, en général, préférable

à en posséder beaucoup. Les mégadonnées accaparent l'attention, alors que les petits ruisseaux de données peuvent être tout aussi précieux. Il n'est que trop fréquent de ne pas détenir de données du tout, soit parce qu'il n'en existe aucune qui soit pertinente, soit parce qu'elles existent, mais qu'on ne peut les trouver, soit parce qu'elles existent, mais qu'elles ne sont pas accessibles en raison d'un contrôle exclusif, d'embargos, de barrières techniques, de détériorations dues à une mauvaise conservation, soit, tout simplement, parce que ceux qui possèdent ces données ne peuvent pas, ou ne veulent pas, les diffuser.

Les données prolifèrent, sous forme numérique comme physique. Le *big data* permet de penser et de poser de nouvelles questions à grande échelle. Pour la première fois, les scientifiques peuvent interroger des ensembles de données où $n = \text{tous}$ (Edwards *et al.*, 2013 ; Mayer-Schonberger et Cukier, 2013 ; Schroeder, 2014). Néanmoins, les données numériques sont aussi bien plus fragiles que les sources physiques, dont certaines ont traversé les siècles. Contrairement au papier, au papyrus et à la peinture, les données numériques ne peuvent être interprétées sans le dispositif technique qui a permis de les créer. Les matériels et les logiciels évoluent rapidement, rendant les fichiers numériques illisibles s'ils ne sont pas migrés vers leurs nouvelles versions au fur et à mesure. Une documentation doit les accompagner, non seulement pour les lignes et les colonnes d'un tableur, mais aussi dans les procédures grâce auxquelles on les a obtenus. De même, des spécimens, des diapositives et des échantillons peuvent parfois n'être interprétables qu'en se référant à leur documentation. Si l'on n'investit pas expressément dans la conservation des données, la plupart d'entre elles disparaîtront rapidement.

C'est la puissance des données combinées à leur fragilité qui en fait un sujet d'étude fascinant dans la communication savante. Prises isolément, les données n'ont ni valeur ni sens. Elles peuvent constituer un atout ou une charge, ou les deux. Elles existent au sein d'une infrastructure de la connaissance, c'est-à-dire d'un écosystème de personnes, de pratiques, de technologies, d'institutions, d'objets matériels et de relations. Tous les éléments de l'infrastructure sont en mouvement, avec des fluctuations dans les parties prenantes, les technologies, les politiques et les jeux de pouvoir. Les enjeux sont importants non seulement pour les savantes et savants d'aujourd'hui et de demain, mais aussi pour celles et ceux qui souhaitent utiliser les connaissances qu'ils créent.

Big Data, Little Data

Par son sous-titre français – *Big Data, Little Data, No Data* –, le présent ouvrage se réclame de Price et du souci de tous les domaines de recherche pour la conservation et la maîtrise de leurs ressources intellectuelles. Les données sont des moyens, des résultats et des atouts de la recherche. Elles sont omniprésentes et pourtant souvent éphémères. La question « qu'est-ce qu'une donnée ? » devient souvent « quand est-ce une donnée ? », car admettre qu'un phénomène peut être traité comme une donnée est, en soi, un acte scientifique (Borgman, 2007, 2012a ; Bowker *et al.*, 2010 ; Star et Bowker, 2002).

On trouvera une définition nominale du mot anglais « *data* » (donnée) dans l'*Oxford English Dictionary* : 1) « information ; donnée ; ensemble de données » ; 2) « informations (surtout numériques) liées entre elles et considérées collectivement, généralement obtenues par un travail scientifique et utilisées à des fins de référence, d'analyse ou de calcul » ; 3) « quantités, caractères ou symboles traités par ordinateur, considérés collectivement. Également (en contexte non technique) : information sous forme numérique ». Ces définitions, restrictives et circulaires, ne parviennent ni à saisir la richesse et la variété des données de la recherche, ni à dévoiler leurs propres prémisses épistémologiques et ontologiques. C'est pourquoi le chapitre 2 est consacré à expliquer le concept de donnée.

Les données possèdent des caractéristiques qui, conjuguées à des tendances sociales et techniques plus larges, contribuent de plus en plus à la reconnaissance de leur utilité, de leur valeur et des problèmes qu'elles posent dans la communication savante.

Masse

Derek de Solla Price (1963) a établi que les principales distinctions entre science légère et science lourde (ou mégascience) sont d'ordre qualitatif. De son point de vue, la science lourde était dominée par des collègues invisibles qui établissaient des relations communautaires, échangeaient des informations en privé et géraient les activités professionnelles de leur discipline (Crane, 1970 ; Furner, 2003b ; Lievrouw, 2010). La science légère est, quant à elle, menée à une échelle plus modeste, dans des communautés de recherche plus restreintes ; le consensus sur les questions et les méthodes de recherche y est moindre, de même que l'infrastructure. Depuis les observations de Price, la manière de pratiquer la science sous toutes ses formes a considérablement changé. Parce qu'il était l'un des premiers historiens contemporains de la science, son point de vue était fortement influencé

par l'essor de l'activité de recherche dans l'après-guerre (Furner, 2003a, 2003b). Les pratiques dominantes dans la recherche aujourd'hui, qui font un usage intensif des données et des calculs, étaient à peine visibles lorsque Price est décédé en 1981. Néanmoins, son intuition quant à la distinction qualitative entre science légère et science lourde est toujours valable à l'ère du *big data*.

Les concepts de *big data* et *little data* ne sont que grossièrement analogues à ceux de *big science* et *little science*. Price ne distinguait pas ces derniers en fonction de la taille des projets, mais de la maturité de la science en tant qu'entreprise. La science contemporaine, ou *big science* pour reprendre ses termes, se caractérise par des efforts collaboratifs internationaux et par des collèges invisibles de chercheurs et de chercheuses qui se connaissent et échangent des informations de manière formelle et informelle. La *little science*, ce sont les trois cents ans de travaux indépendants à petite échelle qui permettent de développer la théorie et la méthode pour comprendre les problématiques de la recherche. Cette science légère, qu'on appelle souvent *small science*, se caractérise par des méthodes et des données hétérogènes, ainsi que par une analyse et un contrôle locaux (Borgman *et al.*, 2007a ; Cragin *et al.*, 2010 ; Taper et Lele, 2004). Comme le remarquait Price, les disciplines légères peuvent devenir lourdes, bien que beaucoup conserveront un caractère modeste.

La distinction entre *big data* et *little data* est problématique parce que quelque chose peut être massif (*big*) de nombreuses façons. Le terme *big data* n'est entré dans le dictionnaire *Oxford* qu'en 2013 : « données de très grande taille, généralement au point que leur manipulation et leur gestion présentent des difficultés logistiques significatives ; [également] la branche informatique concernant ce type de données ». D'autres définitions des mégadonnées parlent d'échelle relative plutôt que de taille absolue. Dans le cadre des applications commerciales et administratives, Mayer-Schonberger et Cukier (2013) envisagent le *big data* sous l'angle d'informations qui peuvent être extraites à grande échelle, alors qu'elles ne pourraient pas l'être à une échelle plus modeste. Dans le cadre universitaire, le *big data* est la recherche sur un phénomène donné rendue possible par le recours aux données à une échelle ou dans un périmètre sans précédent (Meyer et Schroeder, 2014 ; Schroeder, 2014).

Des données massives ou non selon ce qu'on peut en faire, les informations qu'elles peuvent nous révéler et l'échelle d'analyse adaptée au phénomène étudié, qu'il s'agisse de comportement du consommateur ou de découverte de médicaments. Une ancienne définition, qui distingue en quoi les données peuvent être massives, reste encore utile : volume, variété, vitesse ou une combinaison de ces dimensions (Laney, 2001). L'augmentation substantielle de l'une ou l'autre peut modifier l'échelle de la recherche.

L'omniprésence des données contribue aussi à leur caractère massif. Maintenant que nos quotidiens sont outillés de technologies de l'information, on enregistre les traces du comportement humain en toute facilité. Il y a vingt ans à peine, l'accès aux télécommunications se mesurait par la proportion de foyers équipés d'une ligne téléphonique. Aujourd'hui, chaque individu possède potentiellement plusieurs outils de communication, chacun doté de son propre identifiant. Même dans les pays en voie de développement, la fourniture numérique d'information est possible grâce à la croissance exponentielle des technologies de communication mobile. Ces appareils omniprésents sont cependant bien plus que de simples téléphones. Ils peuvent détecter, communiquer et calculer ; ils peuvent enregistrer et diffuser du texte, des images, du son et des vidéos. Ces traces peuvent être marquées de coordonnées de temps et de lieu, créant des relevés d'activité continus. Les bâtiments, les véhicules et les lieux publics sont eux aussi équipés de technologies similaires. On peut combiner ces traces pour élaborer de fructueux modèles d'activité sociale. Les données et leurs usages potentiels prolifèrent bien plus vite que le droit à la vie privée ou les politiques informationnelles, qui peinent à suivre.

L'essor du concept de *donnée* dans les médias et dans le discours académique reflète l'omniprésence des sources de données et la masse de données désormais disponibles sous forme numérique. Après avoir été longtemps prédite, la masse critique a été atteinte dans les sciences exactes, en médecine, en affaires et au-delà. Dans le jargon des affaires, on dirait que le *big data* a atteint le point de bascule, le seuil de popularité qui permet à une idée de se répandre rapidement (Gladwell, 2002). Dans tous les secteurs, les données numériques sont plus faciles à générer, miner et diffuser que jamais.

La possibilité de poser de nouvelles questions, de dégager de nouvelles tendances et d'enregistrer des phénomènes autrefois insaisissables a créé une industrie nouvelle, parfois compatible avec les préoccupations de la recherche ou non.

Ouverture

Les tendances aux modèles ouverts ou libres dans les domaines du logiciel, de l'administration, des normes, des publications, des données, des services et de la production collaborative de connaissances ont transformé les relations entre les acteurs de tous ces secteurs (Benkler, 2007 ; Hess et Ostrom, 2007a ; Kelty, 2008 ; Raymond, 2001). L'ouverture favoriserait la circulation de l'information, la modularité des systèmes et des services et l'interopérabilité. Cependant, l'ouverture comporte également des coûts économiques et sociaux, comme le montre le mouvement du « logiciel libre ». Pour reprendre la distinction de

Richard Stallman (2002), l'ouverture s'apparente davantage à la « *free speech* » (liberté d'expression) qu'à la « *free beer* » (bière gratuite).

On considère généralement que l'édition en accès ouvert remonte à la déclaration de Budapest en 2002, laquelle puise aux expériences d'édition électronique des années 1970 (Budapest Open Access Initiative, 2002 ; Naylor et Geller, 1995). L'accès ouvert aux données possède des racines plus anciennes encore. Le système du Centre mondial des données a été créé dans les années 1950 pour archiver et diffuser les données recueillies au sein des programmes d'observation de l'année internationale de géophysique en 1957-1958 (Korsmo, 2010 ; Shapley et Hart, 1982). CODATA a été fondé en 1966 par le Conseil international pour la science afin de promouvoir la coopération dans le domaine de la gestion et de l'utilisation des données (Lide et Wood, 2012). En 2007, l'Organisation de coopération et de développement économiques (OCDE) a codifié les principes de l'accès aux données de la recherche issues de financements publics (Organisation de coopération et de développement économiques, 2007). Depuis, les rapports d'orientation sur l'accès aux données de la recherche continuent de foisonner (Arzberger *et al.*, 2004 ; National Research Council, 1997 ; Esanu et Uhler, 2004 ; Mathae et Uhler, 2012 ; Pienta *et al.*, 2010 ; Wood *et al.*, 2010). Nous examinons plus avant l'édition en accès ouvert et l'*open data* au chapitre 3.

L'accès ouvert est en partie une réponse à la marchandisation des ressources informationnelles. Bien que cette tendance ait pour origine des changements de politiques dans les secteurs de la propriété intellectuelle et de l'économie de l'information, l'atteinte de la masse critique a ouvert de nouveaux marchés. Antécédents médicaux, comportement du consommateur, médias sociaux, recherche d'information, édition universitaire et génétique font partie des domaines où la concentration de données est suffisante pour créer et déplacer des marchés. Certaines de ces données s'échangent exclusivement dans le secteur privé, mais beaucoup suscitent des intérêts académiques et commerciaux. Les données de la recherche universitaire peuvent avoir une valeur commerciale, tout comme les données commerciales peuvent revêtir un intérêt pour la recherche ; cela mène à de nouveaux partenariats et à de nouvelles tensions (Lessig, 2004 ; Mayer-Schonberger et Cukier, 2013 ; Schiller, 2007 ; Weinberger, 2012).

L'accès ouvert, conjugué à la marchandisation des données, conduit à des mutations dans les politiques de recherche. Pouvoirs publics, organismes de financement et revues encouragent ou enjoignent désormais les chercheurs à diffuser leurs données (Finch, 2012 ; National Science Foundation, 2010b ; National Institutes of Health, 2003 ; Research Councils UK, 2012a). L'accès ouvert aux publications et

aux données accélère la circulation des contenus académiques dans de nombreux domaines et participe aux tensions entre parties prenantes.

La circulation de l'information dépend plus que jamais de l'infrastructure technologique. Les réseaux de télécommunications augmentent leur capacité et leur pénétration avec et sans fil. Les investissements technologiques ne diminuent pas dans l'offre et la demande d'informations, d'outils et de services. Cependant, ils ne conduisent pas directement à des améliorations de l'échange d'informations. En effet, les infrastructures techniques sont aussi des cibles pour l'espionnage, qu'il soit industriel, politique ou académique. La vie privée, la confidentialité, l'anonymat et la maîtrise des ressources intellectuelles sont en jeu. Déplacer des données scientifiques ou autres à travers des réseaux requiert un équilibre délicat entre sécurité, droits, protection, interopérabilité et politiques.

La longue traîne

« La longue traîne » est une façon courante de caractériser la disponibilité et l'utilisation des données dans des domaines de recherche ou des secteurs économiques. L'expression est d'abord apparue en anglais sous la plume de Chris Anderson (2004) dans un article du magazine *Wired*, où il comparait le commerce de marchandises dans des boutiques physiques à celui des boutiques en ligne. La distribution statistique – une loi de puissance – est bien connue (figure 1.1). Dans le modèle d'Anderson, environ 15 % de la distribution se trouve dans la tête de la courbe et les 85 % restants sont distribués tout au long de la queue. Dans le cas de la recherche universitaire, un petit nombre de laboratoires travaillent avec de très grands volumes de données, un certain nombre avec très peu de données et la majorité se situe entre les deux. Tout à droite de la courbe, un grand nombre de scientifiques mènent leurs recherches avec des quantités minimales de données (Foster *et al.*, 2013).

La longue traîne est un raccourci commode pour montrer la gamme des volumes de données utilisées dans un domaine ou un laboratoire donné. Elle met également bien en valeur le fait que seules quelques disciplines – telles que l'astronomie, la physique et la génétique parmi les sciences exactes, la macroéconomie dans les sciences sociales et certains champs des humanités numériques – travaillent avec de très grands volumes absolus de données. En somme, les volumes de données sont inégalement répartis entre les disciplines.

Figure 1.1. La loi de puissance communément appelée la longue traîne

Crédit : Jillian C. Wallis.

La métaphore de la longue traîne a pour faiblesse de suggérer que les pratiques en matière de données de n'importe quel individu ou discipline peuvent être placées sur un graphique en deux dimensions. En réalité, les activités scientifiques sont influencées par d'innombrables facteurs en sus du volume de données manipulées. L'objet de recherche détermine généralement le choix des méthodes et des données, mais l'inverse peut aussi être vrai. En effet, la disponibilité de données peut permettre de poser d'autres questions et d'appliquer d'autres méthodes. Le choix des données dépend également d'autres ressources à la disposition des chercheurs et des chercheuses en tant qu'individus, comme la théorie, l'expertise, les laboratoires, l'équipement, les réseaux techniques et sociaux, les sites de recherche, le personnel et d'autres formes d'investissements en capital.

On peut toutefois poser une généralité quant à la distribution en longue traîne des données dans la recherche : les données exploitées par un petit nombre d'universitaires à la tête de la courbe tendent à être volumineuses, mais peu variées. Les sciences lourdes qui produisent de grands volumes de données doivent s'accorder sur des instruments (par exemple des télescopes ou des séquenceurs d'ADN) et des formats communs (par exemple des métadonnées ou des structures de base de données). Ces données sont relativement homogènes pour ce qui est du contenu et de la structure. La capacité de normaliser les structures de données facilite le développement d'infrastructures, d'outils et de services partagés. À l'inverse, plus une spécialité scientifique se trouve en queue de distribution et plus ses pratiques sont caractéristiques de la *small science* ou de la recherche légère, plus ses contenus, ses structures et ses représentations sont variés. Les scientifiques travaillant (seuls ou en équipe) dans ces spécialités peuvent adapter leurs méthodes de recherche, leur collecte de données, leur équipement et leur analyse du problème étudié bien plus facilement que celles et ceux officiant dans la recherche lourde, qui dépendent

de télescopes spatiaux, d'accélérateurs linéaires de particules ou de projets de numérisation massive pour obtenir leurs données. Le revers de cette souplesse est l'absence de normes sur lesquelles baser une infrastructure commune, ainsi que le manque de masse critique pour développer et pérenniser des ressources de données.

Aujourd'hui, la majorité du travail scientifique et du travail savant en général est menée par des individus ou de petits laboratoires disposant habituellement de financements minimes (Heidorn, 2008). Certains de ces laboratoires participent à des collaborations mégascientifiques internationales. Ils peuvent produire ou analyser des mégadonnées et les échanger par le biais de dépôts de données communautaires (National Science Board, 2005 ; Olson *et al.*, 2008). Cependant, beaucoup de ces individus et de ces laboratoires mènent des recherches avant tout exploratoires, locales, variées et qui ne disposent pas de ressources communautaires partagées.

No Data

Aujourd'hui que la communauté scientifique, la population étudiante, les pouvoirs publics, les entreprises et le grand public partent du principe que des données accessibles existent sur n'importe quel sujet ou presque, l'absence de données devient d'autant plus visible. Les volumes, la vitesse et la variété des données disponibles pour répondre aux questions de recherche varient grandement d'un domaine à l'autre. Souvent, les champs de recherche disposant de données en abondance les mettent en commun, ce qui favorise la mutualisation des méthodes, des outils et des infrastructures. Dans un contexte de profusion de données qu'aucun individu ou laboratoire ne saurait analyser seul, la mise en commun permet de les explorer, de les combiner et de bénéficier de plus de regards qu'il n'aurait été possible autrement. Dans les domaines où les données sont rares, à l'inverse, celles-ci constituent des « biens précieux »¹ (Sawyer, 2008, p. 361) qui peuvent déterminer le choix des méthodes et de la théorie. Comme la métaphore de la longue traîne, la dichotomie entre domaines riches en données et domaines pauvres en données simplifie à l'excès la complexité des ressources utilisées dans tout travail de recherche. Nous n'exposerons ici que quelques-unes des raisons possibles de l'absence ou de la rareté de données disponibles dans une question ou un projet de recherche déterminé.

1. Sauf mention contraire, toutes les citations ont été traduites par Charlotte Matoussowsky.

Des données indisponibles

Dans la plupart des disciplines, les scientifiques sont encouragés à créer de nouvelles données. Il est bien plus facile d'obtenir un financement pour étudier quelque chose de nouveau par des observations, des expériences, des enquêtes, des modèles, de l'ethnographie ou d'autres moyens que pour réanalyser des données existantes. Les universitaires obtiennent un avantage compétitif en se penchant sur des sujets où aucune donnée n'existe. À l'inverse, l'astronomie, l'étude des médias sociaux, la modélisation de villes et de climats et les recherches *in silico* dans les sciences de la vie sont des exemples de domaines où les scientifiques recherchent des données réutilisables.

Des données pertinentes peuvent exister, mais se trouver entre les mains d'organismes qui n'ont aucune obligation de les publier, ou qui en sont empêchés par la loi. Il peut s'agir de documents professionnels, de procédés brevetés, d'archives muséales, de dossiers scolaires et d'innombrables autres formes d'information potentiellement utiles à la recherche. Certaines de ces données peuvent être mises à disposition sous licence ou à certaines conditions, telles que l'anonymisation. La tendance à l'ouverture des données dans la recherche, l'administration et les affaires a eu pour conséquence de rendre accessibles des données autrefois considérées comme exclusives.

Les données provenant d'essais cliniques, de médicaments ou d'autres interventions médicales sont particulièrement litigieuses. Elles peuvent en effet avoir une grande valeur financière et concurrentielle. Elles jouent également un rôle essentiel dans les soins cliniques. Les patients souhaitent accéder à ces données et découvertes, car elles servent l'intérêt public. La diffusion sélective et la présentation de données d'essais cliniques sont devenues une question publique. Bien que nous ne les étudions pas en profondeur dans le présent ouvrage, les données biomédicales telles que les résultats d'essais cliniques se trouvent en première ligne des évolutions des politiques en matière d'accès ouvert et des mutations des relations entre parties prenantes (De Angelis *et al.*, 2005 ; Edwards *et al.*, 2009 ; Fisher, 2006 ; Goldacre, 2012 ; Hrynaskiewicz et Altman, 2009 ; Kaiser, 2008 ; Laine *et al.*, 2007 ; Lehman et Loder, 2012 ; Marshall, 2011 ; Prayle *et al.*, 2012 ; Ross *et al.*, 2012 ; Wieseler *et al.*, 2012).

Les données personnelles utilisées dans les sciences humaines et sociales, que nous évoquons au chapitre 5, peuvent également être particulièrement sensibles et non diffusables. Les données pouvant être anonymisées à un degré raisonnable, comme celles issues des enquêtes démographiques, sont les plus susceptibles d'être mises à disposition pour être réutilisées. Les données ethnographiques ou

issues d'enquêtes qualitatives sont rarement mises à disposition d'universitaires autres que les enquêteurs et enquêteuses eux-mêmes.

Des données non publiées

Si l'accès ouvert est connu depuis longtemps dans certains domaines de la recherche, la diffusion des données ne suscite pas pour autant une approbation universelle. Alors que la non-diffusion est considérée comme une faute professionnelle dans certaines disciplines, dans d'autres, c'est l'inverse, comme expliqué au chapitre 8. En chimie, par exemple, la collecte et la conservation des données en vue de leur réutilisation sont assimilées à de la simple « philatélie » (Lagoze et Velden, 2009a, 2009b). Les données peuvent constituer de précieux atouts à échanger, marchander et faire miroiter dans des négociations avec des partenaires scientifiques ou financiers. Une fois que les données sont publiées, les laboratoires perdent le contrôle de qui les exploite, comment, quand et pourquoi. Souvent, les chercheurs et chercheuses craignent que leurs données soient détournées, mal utilisées ou mal interprétées, ce qui pourrait rejaillir sur leur propre travail (Hilgartner et Brandt-Rauf, 1994).

À la suite d'évolutions récentes, les demandes de financement doivent désormais comporter un plan de gestion des données, ce qui constitue un premier pas vers la diffusion. Cependant, rares sont les politiques à rendre l'accès ouvert obligatoire. Il est plutôt demandé aux scientifiques de spécifier le type de données qu'elles et ils vont collecter, comment elles et ils comptent les gérer et comment les rendre accessibles à d'autres. De même, quelques revues, de plus en plus nombreuses, demandent que les données citées dans leurs articles soient diffusées. La diffusion peut se faire selon différents mécanismes, par exemple le dépôt dans des archives locales ou des référentiels institutionnels, l'ajout en annexe d'articles de revues, la publication sur des sites web locaux ou encore la diffusion à la demande (Alsheikh-Ali *et al.*, 2011 ; Wallis *et al.*, 2013).

Dans certains domaines, les laboratoires disposent de périodes d'embargo, ou période d'utilisation exclusive, pour contrôler leurs données avant de les diffuser. Cette période dure généralement de quelques mois à quelques années. Elle est censée être suffisamment longue pour analyser les données et publier les découvertes réalisées, mais assez courte pour encourager leur diffusion auprès de la communauté scientifique. Lorsque les organismes de financement ou les revues exigent des universitaires la diffusion de leurs données, c'est généralement au moment de la publication des résultats ou après, sur demande. Il est rarement exigé des scientifiques de diffuser leurs données avant la publication

de leurs travaux, à moins que la période d'embargo n'ait été dépassée ou que d'autres règles s'appliquent, comme c'est le cas pour les études cliniques.

Dans les domaines pauvres en données décrits par Steve Sawyer (2008), garder ses données par devers soi est une pratique communément admise. Les personnels de recherche en sciences humaines, par exemple, peuvent vouloir préserver leur accès à des manuscrits, des lettres ou d'autres sources rares le plus longtemps possible. Dans les sciences sociales, ils protégeront leurs contenus, leurs sites et les données associées. Dans les sciences de la vie et les sciences physiques, ce seront les sites, les espèces, les observations et les expériences. Un pays peut refuser l'accès à des sites archéologiques, à du patrimoine et à d'autres ressources, qu'il autorisera uniquement aux savantes et savants autochtones et à leurs partenaires. Les scientifiques venus de pays pauvres peuvent, quelle que soit leur discipline, protéger les trésors qu'elles et ils auront rapportés d'un rare voyage à l'étranger.

Dans de nombreux domaines, les chercheurs et chercheuses peuvent explorer des jeux de données ou d'autres ressources tout au long de leur carrière, sans jamais « en avoir terminé » avec leurs données. Certains ensembles deviennent plus précieux à mesure que le temps passe, par exemple des observations cumulées sur une espèce ou un phénomène. Les notes, les enregistrements et les contenus produits par un ou une scientifique peuvent s'avérer précieux pour d'autres, mais ne devenir disponibles qu'à la fin de sa carrière, si elle ou il les offre à des archives.

Des données inutilisables

Documenter des données pour son propre usage est déjà difficile. Rédiger une documentation qui permette à d'autres de les découvrir, les extraire, les interpréter et les réutiliser l'est bien plus. Rendre des données utiles à d'autres est un effort, dont les motivations dépendent de facteurs sociaux, techniques, politiques, économiques et contextuels innombrables, comme abordé aux chapitres 8 et 9.

Diffuser des données et les rendre exploitables sont deux questions différentes. L'information nécessaire à l'interprétation des données dépend entièrement de la problématique et du domaine de recherche des scientifiques qui les réutiliseront, ainsi que de leur expertise et de leurs ressources, comme expliqué au chapitre 4 et dans les études de cas. Pour interpréter les données, il est généralement nécessaire de disposer de guides de codification, de modèles et de descriptions détaillées des méthodes de collecte, de nettoyage et d'analyse. De plus, les jeux de données numériques ne peuvent être ouverts qu'avec des logiciels particuliers, comme des outils statistiques, du code spécifique à un instrument ou des logiciels conçus pour des

applications propres au domaine, de l'art à la zoologie. Beaucoup sont des logiciels propriétaires. Par ailleurs, des renseignements sur les origines et les transformations des données peuvent être essentiels à leur réexploitation. Plus la réutilisation envisagée est loin du point d'origine, que ce soit en termes de temps, de théorie, de discipline ou de toute autre mesure de distance, plus il est difficile d'interpréter un jeu de données ou de jauger sa valeur.

Si les données ne sont pas documentées à temps, tant que des spécialistes sont disponibles pour les décrire, elles peuvent rapidement perdre toute utilité. De même, les jeux de données sont prompts à se désynchroniser des versions logicielles et matérielles utilisées pour les créer et les analyser.

Quelles données valent la peine d'être préservées, pourquoi, pour qui, par qui et pour combien de temps ? La responsabilité de leur conservation incombe-t-elle d'abord aux personnels de recherche, aux communautés scientifiques, aux universités, aux organismes de financement ou à d'autres acteurs ? Ces questions, que nous explorons au chapitre 10, sont au cœur de la problématique de la conservation des données.

Provocations

Manifestement, la question des *données* est bien plus complexe que ce que la presse grand public ou les politiques officielles laissent entendre. Même restreinte à la recherche et à l'érudition, elle reste vaste et compliquée. La littérature sur les données de la recherche augmente rapidement, certes, mais chaque article, communication, livre blanc, rapport et manifeste ne traite que d'un seul aspect de ce problème éphémère. Le présent ouvrage est la première monographie à envisager tout « l'éléphant » des données aux points de vue social, technique et stratégique et à se fonder sur des exemples tirés de différentes disciplines universitaires. Il fait suite à une exploration plus générale de la recherche à l'ère numérique (Borgman, 2007) pour traiter de l'expansion inédite de l'intérêt pour les données apparue dans l'intervalle, comme le montrent les tendances évoquées plus haut.

Aussi ambitieux que soit le projet de ce livre, d'autres questions demeurent quant à la nature, aux rôles et aux usages des données dans la recherche. Théorie, preuve et pratique sont étroitement mêlées. Nous identifions les points d'intersection et les démenons lorsque c'est possible. Les thèses que nous présentons ici explorent les problèmes d'actualité en partant d'un profond souci pour les enjeux auxquels la recherche universitaire et l'activité académique sont confrontées. Le propos

adopte la forme de six provocations visant à susciter un débat plus approfondi entre les nombreux acteurs de l'entreprise scientifique :

1. La reproductibilité, le partage et la réutilisation des données sont des questions débattues depuis des décennies, voire des siècles. Nous intéresser à qui possède les données de la recherche, les contrôle, y a accès et les pérennise permettra de déterminer comment exploiter leur valeur et qui pourra le faire.

2. Transmettre des connaissances au fil du temps et dans différents contextes est difficile. Certaines formes et représentations de données sont partagées aisément d'une discipline à une autre, d'un contexte à un autre et d'une époque à une autre, mais beaucoup d'autres ne peuvent pas l'être. Il est nécessaire de comprendre quelles fonctionnalités sont importantes ou non afin d'inspirer les pratiques et politiques scientifiques et guider les investissements dans les infrastructures de la connaissance.

3. En dépit de la prolifération des formes et des genres, les fonctions des publications scientifiques restent stables. Les données servent des objectifs différents des articles, des ouvrages et des communications. Traiter les données comme des publications risque de renforcer le poids des intérêts catégoriels au détriment de l'expérimentation de nouveaux modèles de communication savante. Il convient d'envisager les fonctions des données dans la recherche du point de vue de différentes parties prenantes.

4. Les travaux de recherche se diffusent plus largement grâce à des mouvements tels que l'édition en accès ouvert, l'ouverture des données et le logiciel libre. Les finalités différentes des données et des publications dans la recherche influent sur les incitations à diffuser, ainsi que sur les moyens et les pratiques de diffusion. L'ouverture de l'accès aux données a des répercussions encore mal comprises sur les personnels de recherche, les bibliothèques, les universités, les organismes de financement, les maisons d'édition et les autres acteurs.

5. Les infrastructures de la connaissance évoluent afin de prendre en compte l'ouverture des données, leur usage intensif dans la recherche, les nouvelles technologies, les médias sociaux et les transformations des pratiques et des politiques. Certaines parties prenantes y gagnent tandis que d'autres y perdent. Coûts, bénéfices, risques et responsabilités sont redistribués. De nouvelles formes d'expertise sont nécessaires, mais leur application varie selon les contextes et les domaines de recherche.

6. Les infrastructures de la connaissance se développent et s'adaptent au fil des générations de savantes et savants. Elles ont besoin d'une vision à long terme pour leur conception et leurs politiques, or le financement de la recherche se fait sur des cycles courts. Des investissements substantiels dans l'infrastructure sont nécessaires afin d'acquérir, de pérenniser et d'exploiter les données de la recherche aujourd'hui et demain. Ces investissements seront controversés, car les choix faits à présent décideront de quelles données et autres ressources informationnelles nous disposerons dans le futur.

Ces provocations sont approfondies en dix chapitres, organisés en trois parties. Les quatre premiers chapitres donnent une vision d'ensemble des données et de la recherche afin d'établir les fondements des six provocations. La deuxième partie consiste en trois études de cas, en autant de chapitres. Nous y étudions de multiples exemples tirés des sciences exactes, des sciences sociales et des sciences humaines, chacun venant corroborer les provocations. Les politiques et les pratiques en matière de données constituent l'objet de la troisième partie, qui offre une analyse comparative de l'accessibilité et de l'utilisation des données à travers les disciplines et les contextes (chapitre 8), explore les questions du crédit, de l'attribution et de la découvrabilité (chapitre 9) et se termine en demandant ce qu'il faut conserver et pourquoi (chapitre 10). La dernière partie dégage les implications de ces provocations pour la pratique scientifique et les politiques en matière de recherche et ouvre vers de nouvelles questions.

Conclusion

La santé d'une communauté intellectuelle repose sur la poursuite obstinée de connaissance de spécialistes, en physique nucléaire comme en culture tibétaine. Le défi consiste à développer des infrastructures de la connaissance qui servent la diversité des idées, des questions, des méthodes et des ressources que chacun apporte à l'érudition. Des universitaires de tous domaines conviennent que l'enthousiasme pour le *big data* constitue un risque pour la recherche à petite échelle (Alberts, 2012 ; Berlekamp, 2012 ; Gamazon, 2012 ; Meyer, 2009 ; Sawyer, 2008 ; Siminovitch, 2012). Le présent ouvrage ouvre la boîte noire des « données », y examinant les idées, les théories, les pratiques, les politiques, les valeurs, les incitations et les motivations intrinsèques. En soi, les données ne sont pas un sujet particulièrement enthousiasmant. Il s'agit cependant d'un cadre extrêmement utile pour observer les bouleversements de la pratique scientifique et comment ils s'inscrivent dans la technologie, l'éducation et les politiques. Nos six provocations visent à élargir et approfondir le débat sur les données de la recherche

en y faisant participer toutes les parties prenantes. Aucune d'elles n'appelle une réponse simple ou claire. Beaucoup de ces questions se négocieront isolément, souvent au coup par coup. Elles influenceront toutes sur la manière dont la recherche est menée aujourd'hui et demain. Tel est le cœur de cet ouvrage.

2. Qu'est-ce qu'une donnée ?

Introduction

Même si le concept est populaire depuis peu, le terme « *data* » (« donnée ») n'a rien de nouveau en anglais. L'*Oxford English Dictionary* le fait remonter à 1646 dans un sens théologique, où il était généralement utilisé au pluriel. L'analyse de l'usage de ce mot dans ECCO (Eighteenth Century Collections Online) (Gale Cengage Learning, 2013) par Daniel Rosenberg (2013) a montré une augmentation régulière des mentions à partir du ^{xvii}^e siècle. Les premières occurrences sont en latin ; le mot *data* entre dans la langue anglaise par le biais des mathématiques et de la théologie. Tout au long du ^{xviii}^e siècle, on débat de l'usage au singulier ou au pluriel. On évoquait les *data* soit en tant que 1) postulats formant le fondement d'un argument, soit en tant que 2) faits, en particulier ceux tirés des Écritures. Rosenberg a découvert que ce n'est qu'à la fin du ^{xviii}^e siècle que le mot s'est mis à désigner des faits sous forme de preuve scientifique recueillis grâce à des expériences, des observations et d'autres recherches. Son étude du corpus en ligne Google Books montre une croissance continue du terme dans la littérature du ^{xx}^e siècle, mais ces dernières analyses sont moins concluantes que celles menées dans ECCO.

L'analyse historique de Rosenberg, certes confinée à l'usage en langue anglaise, conclut que le mot *data* reste une expression rhétorique sans essence propre. Les données ne sont ni une vérité ni une réalité. Elles peuvent être des faits, des sources de preuve ou des postulats qui sont utilisés pour affirmer une vérité ou une réalité. La division tripartite entre donnée, information et connaissance (Machlup et Mansfield, 1983) simplifie à l'excès les relations entre ces notions complexes. Meadows (2001, p. 3) note qu'il « y a toujours une part d'arbitraire dans ce que nous considérons comme une donnée de base ». La remarque de Michael Buckland, comme quoi les données seraient des « preuves présumées », est le plus à même de saisir l'ambiguïté du terme (Buckland, 1991, communication personnelle, 2006 ; Edwards *et al.*, 2007).

Comme noté au chapitre 1, la question « qu'est-ce qu'une donnée ? » se traite mieux en demandant « quand est-ce une donnée ? ». Les questions véritablement intéressantes sur le rôle des données dans la recherche portent sur les processus qui transforment une chose en donnée. Comment les individus, les laboratoires et les communautés créent-ils, sélectionnent-ils et utilisent-ils les données ? Dans ces décisions, quels facteurs sont associés aux données en tant que telles ? Lesquels sont liés à des questions ou à des méthodes de recherche ? Lesquels dépendent

de la façon dont les données sont représentées ? Comment ces considérations varient-elles selon les champs, les disciplines et les problématiques de recherche ? En quoi dépendent-elles des relations entre l'individu et les données, du créateur au conservateur ? Comment la notion de donnée évolue-t-elle au long d'un projet de recherche ou durant la vie des données ? Comment toutes ces questions se transforment-elles alors que de plus en plus de données – ou de signaux pouvant être traités comme des données – sont disponibles sous forme numérique ?

Les données ne sont pas des objets purs ou naturels possédant une essence propre. Elles existent au sein d'un contexte dont elles tirent leur sens ; ce sens provient aussi du point de vue de l'observateur. Comme nous l'avons esquissé dans la deuxième provocation, le degré auquel ces contextes et ces significations peuvent être représentés influe sur la transférabilité des données. Ce chapitre examine les tentatives de définition de la donnée en termes théoriques et opératoires pour se conclure par une définition de travail que nous utiliserons tout au long du présent ouvrage.

Définitions et terminologie

La littérature scientifique, les politiques officielles et la presse grand public sont pleines de débats sur les données, sans qu'elles se donnent vraiment la peine de définir les termes utilisés. Comme le remarque Rosenberg (2013), même l'histoire des sciences et l'épistémologie ne mentionnent les données qu'en passant (Blair, 2010 ; Daston, 1988 ; Poovey, 1998 ; Porter, 1995). D'autres travaux fondateurs sur l'élaboration de sens dans les sciences se penchent sur les faits, les représentations, les inscriptions et les publications sans vraiment s'intéresser aux données en tant que telles (Bowker, 2005 ; Latour et Woolgar, 1986 ; Latour, 1987, 1988, 1993). Dans les sciences humaines, le mot est rarement prononcé, bien que les chercheurs et chercheuses usent de faits, de chiffres, de lettres, de symboles et d'autres entités qui seraient considérées comme des données dans les sciences exactes et sociales. Maintenant que les humanités s'appuient davantage sur les collections numériques, empruntent plus d'outils à d'autres disciplines et développent leurs propres méthodes analytiques des objets numériques, leur concept de donnée devient plus explicite (Borgman, 2009).

Les données sont une forme d'information ; ce dernier concept est encore plus vaste et plus difficile à définir. Les problèmes épistémologiques et ontologiques abondent et donnent lieu à de nombreux livres consacrés à expliquer l'information et la connaissance (Blair, 2010 ; Brown et Duguid, 2000 ; Burke, 2000, 2012 ; Day, 2001 ; Ingwersen et Jarvelin, 2005 ; Liu, 2004 ; Meadows, 2001 ; Svenonius, 2000).

Buckland (1991) distingue l'information comme processus, l'information comme connaissance et l'information comme chose. Donald Case (2002, 2012) en a recueilli des dizaines de définitions et les a regroupées en fonction de leur manière de gérer l'incertitude, la dimension physique, la structure et le processus, l'intention et la vérité. Jonathan Furner (2004a) a appliqué trois critères pour sélectionner des définitions de l'information : cohérence, parcimonie et utilité. Plus tard, il a identifié trois familles de conceptions de l'information qui sont utiles dans un large cadre : les définitions sémiotiques, sociocognitives et épistémiques (Furner, 2010).

Le concept de donnée mériterait à lui seul un ouvrage entier. Cependant, une approche plus restreinte suffira pour analyser les données dans le contexte de la communication savante. Cet exposé se limitera aux définitions, théories et notions utiles pour explorer les similitudes et les différences dans la création, l'utilisation et l'appréhension des données dans les communautés scientifiques.

Définitions par l'exemple

On définit le plus souvent les données en en donnant des exemples, comme des faits, des chiffres, des lettres et des symboles (National Research Council, 1999). Une liste d'exemples ne constitue cependant pas véritablement une définition, car elle n'établit pas de limites claires entre ce que le concept englobe et ce qu'il exclut. La définition proposée par Peter Fox et Ray Harris (2013, p. 10) est typique : « Les “données” comprennent, au minimum, les observations numériques, le suivi scientifique, les données issues de capteurs, les métadonnées, les sorties de modèle et les scénarios, les données comportementales qualitatives ou observées, les visualisations et les données statistiques recueillies à des fins administratives ou commerciales. Les données sont généralement considérées comme des ressources pour la recherche ».

Dans le domaine des politiques, Paul Uhlir et Daniel Cohen (2011) incluent une vaste gamme d'attributs dans leurs exemples de données :

Le terme « donnée » tel qu'il est utilisé dans ce document doit être compris au sens large. Il désigne, outre des manifestations numériques de la littérature (y compris du texte, du son, des images fixes, des images mouvantes, des modèles, des jeux ou des simulations), des formes de données et de bases de données qui nécessitent le recours à des matériels et programmes informatiques pour être utilisables, comme divers types de données de laboratoires, par exemple des données spectrographiques, des séquençages de génomes et des données de microscopie électronique ; des données d'observation, comme des données de télédétection, géospatiales ou socioéconomiques ; ou

d'autres formes de données, qu'elles soient générées ou compilées, par des êtres humains ou par des machines.

La définition d'Uhlir et Cohen admet que les données peuvent être créées par des personnes ou par des machines et reconnaît leurs relations avec les ordinateurs, les modèles et les logiciels. Néanmoins, une telle liste ne peut constituer, au mieux, qu'un point de départ de ce que les données peuvent être pour une personne donnée, à une fin donnée, à un moment donné.

Jorge Luis Borges (1999) a expliqué d'une manière particulièrement charmante pourquoi une définition sous forme de liste est insatisfaisante. Dans son essai de 1942, il présente une classification des animaux tirée d'une prétendue encyclopédie chinoise intitulée *Le marché céleste des connaissances bénévoles* : « a) appartenant à l'Empereur, b) embaumés, c) apprivoisés, d) cochons de lait, e) sirènes, f) fabuleux, g) chiens en liberté, h) inclus dans la présente classification, i) qui s'agitent comme des fous, j) innombrables, k) dessinés avec un très fin pinceau de poils de chameau, l) et cætera, m) qui viennent de casser la cruche, n) qui de loin semblent des mouches »¹. La pensée de Foucault (1994), de Lakoff (1987) et de bien d'autres philosophes et intellectuels a été influencée par la diatribe subtile de Borges sur les mécanismes de classification.

Définitions opérationnelles

On trouve les définitions les plus concrètes des « données » dans des contextes opérationnels. Les institutions chargées de gérer de vastes recueils devraient expliciter quelles entités elles prennent en charge et comment ; cependant, rares sont les définitions qui dressent des frontières claires entre ce qui constitue une donnée et ce qui n'en est pas.

Le modèle de référence Open Archival Information System (OAIS) (Consultative Committee for Space Data Systems, 2012)² représente l'un des principes les plus connus d'archivage de données. Ce document de consensus sur les bonnes pratiques a été créé pour les sciences spatiales, mais a été largement adopté dans les sciences exactes et sociales comme principes directeurs pour l'archivage de données. Le modèle OAIS utilise le mot « donnée » comme complément du nom : ensemble de données, unité de donnée, format de donnée, base de données, entité de données,

1. J. L. Borges, *Œuvres complètes*, t. I, P. Bénichou, S. Bénichou-Roubaud, J.-P. Bernès et R. Cailliois (traduction), Paris, Gallimard (La Pléiade), 1957, p. 749.

2. Pour la version française, voir Consultative Committee for Space Data Systems, 2017, p. 1-10.

etc., tout en définissant les données proprement dites en termes généraux, à l'aide d'exemples :

Données : une représentation réinterprétable formalisée de l'information, adaptée à la communication, à l'interprétation ou au traitement. Exemple : une séquence de bits, un tableau de nombres, les caractères d'une page, un enregistrement de paroles ou un échantillon de roche lunaire. (Consultative Committee for Space Data Systems, 2017, p. 1-10)³

Le modèle OAIS distingue donnée et information comme suit :

Information : toute connaissance pouvant être échangée. Lors de l'échange, elle est représentée par des données. Exemple : une séquence de bits (les données) accompagnée d'une description permettant d'interpréter cette séquence de bits comme des nombres représentant des mesures de températures en degrés Celsius (Information de représentation). (*ibid.*, p. 1-12)

La Data Documentation Initiative (DDI) est une série de standards de métadonnées pour gérer des données tout au long de leur cycle de vie (Data Documentation Initiative, 2012). La DDI est largement employée, notamment dans les sciences sociales, pour la description de données, mais pas pour définir la donnée en soi. Les spécifications des métadonnées DDI, qui sont exprimées en XML, peuvent s'appliquer à tout objet numérique que la DDI considère comme une donnée.

L'Inter-University Consortium for Political and Social Research (ICPSR) compte parmi les cocréateurs de la DDI. L'ICPSR est un centre international de pointe qui archive des données de recherche en sciences sociales depuis le début des années 1960. L'ICPSR laisse les contributeurs et contributrices déterminer ce qu'ils considèrent être leurs données. Il conseille les déposants potentiels en ces termes :

Outre les données quantitatives, l'ICPSR accepte les données de recherche qualitative (y compris des transcriptions et des médias audiovisuels) à des fins de conservation et de diffusion. L'ICPSR a pour objectif la préservation numérique et encourage les chercheurs et chercheuses à déposer leurs données en formats émergents, tels que des sites web, des données géospatiales, des données biomédicales et des vidéos numériques. (Inter-University Consortium for Political and Social Research 2012, p. 4)

3. Référence originale : Consultative Committee for Space Data System, 2012, p.1-12.

Ainsi, même des institutions recueillant et conservant de larges volumes de données peuvent n'imposer aucune définition précise de ce qu'elles acceptent ou non. La donnée reste un concept ambigu, ce qui permet aux archives de s'adapter aux nouvelles formes de données au fur et à mesure de leur apparition.

Définitions catégorielles

Dans des contextes opérationnels et généraux de recherche, on peut distinguer des types de données en les regroupant de façon pertinente. Les archives peuvent, par exemple, classer des données selon leur degré de traitement. Les spécialistes des politiques scientifiques les regrouperont par origine, par valeur ou en fonction d'autres facteurs.

Degrés de traitement

Les niveaux de traitement définis par l'Earth Observing System Data Information System (EOS DIS) de la NASA figurent parmi les catégories de données les plus tranchées. Les données de même origine sont distinguées en fonction de leur traitement, comme montré à la figure 2.1 (NASA's Earth Observing System Data and Information System, 2013).

Ces distinctions fines servent un objectif opérationnel. Les données EOS DIS commencent au niveau 0, qui correspond à des « données brutes à résolution intégrale de l'instrument ». Les produits de niveau 0 ont déjà été nettoyés afin de retirer les artefacts de communication ; ils ne sont donc pas des signaux issus directement de l'instrument. Le niveau suivant, 1A, est à résolution intégrale et s'est vu ajouter des métadonnées pour indiquer des références temporelles, les paramètres de l'instrument et d'autres informations. Au niveau 1B, les données sont divisées par capteurs pour les instruments avec lesquels c'est possible. Les niveaux 2, 3 et 4 sont traités davantage afin d'ajouter d'autres métadonnées, d'harmoniser les produits avec les référentiels espace-temps et d'agréger les données en modèles. Comme indiqué à la figure 2.1, on dispose pour tous les instruments de données de niveau 1 minimum, la plupart arrivent au niveau 2 ou 3 et certains instruments voient leurs données traitées au niveau 4.

Le niveau de traitement de données comme celles des instruments de la NASA dépend de nombreux facteurs, tels que les capacités des instruments et les usages auxquels les données sont destinées. La plupart des scientifiques souhaitent des données de niveau 4 afin de pouvoir les comparer à d'autres modèles de phénomènes. Ces produits sont en effet les plus aisés à comparer d'un instrument à l'autre et d'une mission à l'autre. Certains scientifiques demandent des données de niveau 0 ou même plus brutes encore, en conservant les artefacts de

communication, afin de les nettoyer eux-mêmes. S'il s'agit de tester une théorie, ils ou elles peuvent souhaiter décider eux-mêmes des valeurs aberrantes, de l'échantillonnage, des valeurs manquantes, de la prise en compte de la météorologie et des anomalies techniques, etc. S'il s'agit de repérer des tendances encore complètement inconnues, comme la recherche d'intelligences extraterrestres (SETI), ils ou elles voudront disposer des ensembles de signaux les plus bruts et les plus complets possible (Anderson *et al.*, 2002 ; Sullivan *et al.*, 1997).

Figure 2.1. Niveaux de traitement EOS DIS de la NASA

Les données EOS DIS sont traitées à différents niveaux, allant de 0 à 4. Les produits de niveau 0 constituent des données brutes à résolution intégrale. Aux niveaux supérieurs, les données sont converties dans des formats plus utiles avec de meilleurs paramètres. Tout instrument EOS doit disposer de données de niveau 1. La plupart voient leurs données traitées aux niveaux 2 et 3 et beaucoup atteignent le niveau 4.

Niveau de description des données	
Niveau 0	Données brutes de l'instrument ou de la charge utile reconstituées sous une résolution intégrale, où tous les artefacts de communication (par exemple, synchronisation des cadres, en-têtes de communications, doublons) ont été supprimés. Dans la plupart des cas, l'EOS Data and Operations System [EDOS] fournit ces données aux data centers sous forme de jeux de données de production pour qu'ils soient traités par le Science Data Processing Segment [SDPS] ou par un Science Information Processing Segment [SIPS] et ainsi obtenir des produits de niveaux supérieurs.
Niveau 1A	Données brutes de l'instrument reconstituées sous une résolution intégrale, référencées dans le temps et annotées avec des informations auxiliaires, y compris les coefficients d'étalonnage radiométrique et géométrique et les paramètres de géocodage (par exemple, l'éphéméride de la plateforme) calculés et annexés, mais non appliqués, aux données de niveau 0.
Niveau 1B	Données de niveau 1A ayant été converties aux unités du capteur (tous les instruments ne disposent pas de données sources de niveau 1B).
Niveau 2	Variables géophysiques dérivées à la même résolution et au même endroit que les données sources de niveau 1.
Niveau 3	Variables appliquées sur des référentiels espace-temps uniformes, généralement de manière complète et cohérente.
Niveau 4	Données de sorties de modèle ou résultats de l'analyse de données de niveau inférieur (par exemple, variables dérivées de mesures multiples).

Ces niveaux de traitement ont des répercussions significatives sur la conservation et la maintenance des données pour un usage futur. Celles-ci peuvent nécessiter une gestion à chaque niveau, en particulier les données d'observation, comme les missions de la NASA en produisent, qui ne peuvent être répliquées. Si les données ne sont conservées qu'aux niveaux les plus bas, il peut être nécessaire de prévoir des algorithmes de traitement et de la documentation pour les faire passer aux niveaux supérieurs. Dans de nombreux domaines de la physique, de la chimie et de la biologie, les données instrumentales les plus brutes sont trop volumineuses pour être conservées ; c'est pourquoi les efforts de conservation sont dirigés vers les produits les plus élaborés, qui constituent les résultats d'un projet. Le logiciel de pipeline utilisé pour nettoyer, calibrer et réduire les données d'observation est constamment révisé à mesure qu'évoluent les instruments, les technologies de calcul et les questions de recherche et à mesure que des erreurs sont découvertes et que s'améliorent les méthodes analytiques. Les flux de données d'un instrument peuvent être traités à plusieurs reprises, ce qui conduit à des diffusions multiples. Le contrôle de version constitue donc une part essentielle de la gestion de vastes archives de données d'observation.

Origine et valeur de préservation

Bien qu'ils aient été développés à des fins et dans un système particulier, les niveaux de traitement de la NASA sont utilisés pour catégoriser les données dans d'autres environnements opérationnels. Dans le contexte des politiques scientifiques, des groupements bien plus généraux sont cependant nécessaires. La catégorisation édictée par le National Science Board (NSB) aux États-Unis vise à représenter les données utilisées dans les sciences exactes, les sciences sociales et le secteur technologique. Bien que les sciences humaines, les arts, la médecine et la santé sortent du domaine de compétence du NSB, leurs catégories de données sont aussi pensées pour ces domaines. L'origine des données peut influencer sur les décisions opérationnelles quant à celles valant la peine d'être conservées et pour combien de temps (National Science Board, 2005).

« Les données d'observation », la première des trois catégories du NSB, résultent de la reconnaissance, de la notation ou de l'enregistrement de faits ou d'occurrences de phénomènes, généralement à l'aide d'instruments. Dans les sciences exactes, il peut par exemple s'agir d'observations météorologiques, botaniques et zoologiques, effectuées par satellite, par réseau de capteurs ou par un stylo dans un carnet. Dans les sciences sociales, on pensera à des indicateurs économiques ou à des entretiens, issus de rapports d'entreprises, d'entrevues en ligne ou d'ethnographies. Toute observation peut être associée à des lieux et moments spécifiques ou en impliquer plusieurs (par exemple dans les études transversales

et longitudinales). Les données d'observation sont considérées comme les plus importantes à préserver, car elles sont les moins répliquables.

Les « données computationnelles » sont le produit de modèles, simulations ou processus informatiques. Bien qu'elles soient surtout présentes dans les sciences physiques et les sciences de la vie, on les trouve aussi dans les sciences humaines et sociales. Le physicien ou la physicienne modélise l'univers, l'économiste modélise des interactions interpersonnelles et les marchés et le ou la latiniste modélise des cités et des sites antiques. Pour réutiliser un modèle informatique, une documentation détaillée sur le matériel, les logiciels, les données d'entrée et les étapes intermédiaires peut s'avérer nécessaire. Parfois, les entrées du modèle sont conservées ; parfois, ce sont les sorties. Dans certains cas, seuls les algorithmes sont préservés, au motif que le modèle peut être appliqué de nouveau si nécessaire.

Les « données expérimentales », la troisième catégorie, résultent de procédures en conditions contrôlées pour établir des hypothèses ou mettre à l'épreuve de nouvelles lois. Il peut s'agir, par exemple, de résultats de recherche en chimie obtenus dans un laboratoire expérimental, d'expériences physiques menées dans un accélérateur de particules ou d'expériences psychologiques contrôlées en laboratoire ou sur un terrain. Si l'expérience est conçue pour être répliquable, il peut être plus facile de reproduire ces données que de les conserver. Si les conditions de l'expérience ne peuvent être répliquées, il peut être nécessaire de préserver les données.

Le rapport *Long-Lived Data* met l'accent sur les implications politiques de ces trois catégories de données, qui nécessitent chacune des dispositions particulières pour leur conservation. Il distingue aussi les niveaux de données au sein de chacun des trois types d'origine. Les données peuvent être recueillies sous forme « brute » et affinées dans des versions successives. Dans de nombreux cas, la conservation des données sous plusieurs formes se justifie (National Science Board, 2005, p. 19-20). Le rapport reconnaît que les frontières entre ces catégories sont poreuses. Par exemple, des données d'observation peuvent être employées dans des expériences et des modèles informatiques, ou bien les résultats d'expériences et de modèles peuvent servir à affiner des méthodes de collecte d'observations. Edward (2010) étudie l'interaction entre données d'observation et modèles et expose ainsi le processus, long d'un siècle, qui a rendu mobiles les données de la recherche climatique.

Plusieurs types de traces (*records*) sont associés aux données d'observation, aux données expérimentales et aux données informatiques, comme des documents historiques, des rapports de terrain et des notes manuscrites. *Record* est un autre

de ces termes fondamentaux, mais rarement définis, en dépit de son large usage en droit, en archivistique, en gestion de données et plus généralement en anglais. Selon l'*Oxford English Dictionary*, un *record* est l'attestation d'un fait et connote un témoignage, une trace ou une preuve. Dans les expressions figées *on record* ou *of record*, le sens dominant est le fait d'être conservé comme connaissance ou information. Dans cette acception, le mot *record* est très ancien, remontant au ^{xiv}^e siècle.

Les « traces » (*records*) constituent avantageusement une quatrième catégorie d'origine, car elles comprennent des formes de données qui se classent malaisément parmi les données d'observation, expérimentales ou informatiques ou les résultats de l'une de ces catégories. Des traces de n'importe quel phénomène ou activité humaine peuvent être traitées comme des données à des fins de recherche.

Ces traces peuvent être des documents concernant l'administration, les affaires, les activités publiques ou privées ; des livres et d'autres textes ; des archives ; des documents sous forme d'enregistrements audio et vidéo, de plaques de verre, de papyrus, d'écriture cunéiforme, de bambous, etc. Des traces faisant autorité ont en commun avec les observations de ne pouvoir être répliquées et sont donc précieuses.

Ensembles de données

Les tentatives de catégorisation des recueils numériques de données révèlent leurs origines et leur valeur pour les communautés scientifiques. Le même rapport du National Science Board a établi trois catégories fonctionnelles qui sont largement utilisées dans les sciences (Cragin et Shankar, 2006 ; National Science Board, 2005). Il s'agit, des moins formalisées aux plus formalisées, des ensembles de données de recherche, des ensembles de données ressources ou communautaires et des ensembles de données de référence. De mêmes données peuvent se retrouver dans plusieurs recueils, mais être représentées différemment dans chacune. On peut établir des distinctions plus fines, par exemple entre ensembles physiques et numériques, entre ensembles numériques et numérisés, entre données substitutives et intégrales, entre images statiques et représentations interrogeables et entre chaînes de caractères consultables et contenus enrichis. Nous verrons ces distinctions au chapitre 7, où elles seront étudiées dans le contexte des sciences humaines.

Les « ensembles de données de recherche », la première des trois catégories du NSB, sont les résultats d'un ou plusieurs projets de recherche. Ces données ont bénéficié d'un traitement et d'une conservation limités et leur format et leur structure peuvent ne pas être conformes aux normes scientifiques, quand elles existent. Ces ensembles sont généralement constitués par et pour un groupe de recherche ; ils peuvent ne

pas être sauvegardés à la fin d'un projet. Ces recueils existent par milliers. Ils peuvent concerner, par exemple, « les flux sur les surfaces de neige », le génome d'une levure ou d'autres domaines spécifiques ; ils sont importants pour de petites communautés de recherche (National Science Board, 2005, annexe D).

Les ensembles de données de recherche qui répondent à un besoin permanent peuvent devenir des « ensembles de données ressources » ou « communautaires ». Ces recueils peuvent établir des normes pour leur communauté, soit en en adoptant, soit en en développant de nouvelles. Les ensembles de données ressources bénéficient parfois de subventions directes, mais sans garantie qu'elles seront financées au-delà des priorités immédiates de la communauté ou de l'organisme de financement. Les exemples de cette catégorie vont de la PlasmoDB, qui concerne le génome d'un parasite responsable du paludisme, à l'Ocean Drilling Program, soutenu par la National Science Foundation états-unienne et vingt-deux partenaires internationaux.

Les « ensembles de données de référence », la troisième catégorie, sont ceux qui servent à de vastes communautés, se conforment à des normes solides et sont pérennes. Ils bénéficient de budgets confortables, de communautés diverses et décentralisées et de structures de gouvernance bien établies. On trouvera dans cette catégorie de grandes banques de données internationales qui constituent des ressources scientifiques essentielles, comme la Protein Data Bank, la base de données astronomiques SIMBAD et les jeux de données de référence des collections de l'ICPSR (Protein Data Bank, 2011 ; Genova, 2013 ; National Science Board, 2005 ; Inter-University Consortium for Political and Social Research, 2013).

Ces trois catégories d'ensembles permettent d'évaluer le degré d'investissement d'une communauté dans ses données et le niveau de partage qui s'y produit. Les systèmes de données communautaires font partie des sept types de laboratoires identifiés par Nathan Bos et ses collègues (Bos *et al.*, 2007 ; Olson *et al.*, 2008).

Distinctions conceptuelles

Aussi tranchées qu'elles puissent paraître, les distinctions entre catégories comportent toujours une part d'arbitraire. Toute catégorie et tout nom de catégorie résultent de décisions quant aux critères et à la dénomination. Même les indicateurs les plus concrets, comme la température, la taille et la localisation géospatiale sont des inventions humaines. De même, les systèmes de mesure en pieds et pouces, en mètres et en grammes, en degrés centigrades et Fahrenheit sont l'aboutissement de siècles de négociations. Les constantes fondamentales des poids et mesures sont

soumises à des révisions continues par un organisme international de normalisation (Busch, 2013 ; Lampland et Star, 2009 ; Lide et Wood, 2012 ; Meadows, 2001).

À leur tour, les poids et mesures trouvent des applications multiples. Les balances pour mesurer le poids atomique sont bien plus précises que celles que l'on trouve en supermarché. Les normes de qualité que les pouvoirs publics appliquent à l'eau potable diffèrent considérablement de celles que les surfeurs utilisent pour l'océan où ils nagent. La taille d'une personne se mesure avec des degrés de précision différents dans le cabinet d'un médecin ou lors d'une compétition sportive. La palette des circonstances qui gouvernent la catégorisation des données de la recherche est plus vaste encore.

Sciences exactes et sciences sociales

Si la NASA distingue explicitement les données brutes des données traitées à des fins opérationnelles, le terme « brut » n'en est pas moins relatif, comme d'autres l'ont remarqué (Bowker, 2005, 2013 ; Gitelman, 2013). Le caractère « brut » d'une donnée dépend de là où l'investigation a débuté. Les scientifiques qui combinent des produits de niveau 4 issus de diverses missions de la NASA pourront considérer ceux-ci comme des données brutes, qui constituent leur point de départ. À l'extrême opposé, on cherche à retracer l'origine de données à partir de leur état au moment où l'instrument a détecté le signal pour la première fois. Les instruments sont conçus et fabriqués pour détecter des phénomènes donnés dans des conditions données. Ces choix de conception et de fabrication déterminent à leur tour ce qui peut être capté. Identifier la forme la plus brute possible d'une donnée peut donc s'avérer une régression à l'infini constituée de choix épistémologiques sur la connaissance recherchée.

Dans le cas des chercheurs et chercheuses en sciences sociales qui recueillent des observations sous forme d'enquêtes et d'entretiens, les données brutes peuvent être les questionnaires remplis par les personnes interrogées ou par les enquêteurs et enquêtrices. Souvent, ces formulaires fourmillent de réponses incomplètes ou incompréhensibles. Ils contiennent aussi des erreurs, comme lorsque la personne interrogée se trompe dans l'ordre d'une échelle ou indique une date de naissance improbable. On détecte de telles erreurs en comparant des questions censées susciter des réponses semblables et en réduisant les variables à des séries de chiffres. Dans d'autres cas, les personnes interrogées peuvent avoir répondu au hasard, soit par embarras, soit par malice.

Nettoyer ces types de données tient autant de l'art que de la science et suppose une expertise méthodologique et statistique considérable (Babbie, 2013 ; Shadish *et al.*, 2002). Une fois propres, elles deviennent la base d'analyses dont on tirera des

conclusions de recherche. Il arrive que les décisions quant à la gestion des données absentes, la saisie des valeurs manquantes, la suppression des valeurs aberrantes, la transformation des variables et les autres tâches courantes de nettoyage et d'analyse des données ne soient guère consignées. Elles ont pourtant des répercussions profondes sur les découvertes, l'interprétation, la réutilisation et la réplication (Blocker et Meng, 2013 ; Meng, 2011).

Sciences humaines

Le sens du mot « donnée » est particulièrement ambigu dans les sciences humaines (Borgman, 2009 ; Unsworth *et al.*, 2006). L'équivalent le plus proche des catégories « brut » et « traité » employées dans les sciences exactes et sociales est la distinction entre sources primaires et secondaires. Dans le langage courant, une « source primaire » est un objet ou un document original, par exemple une sculpture ou un manuscrit historique, tandis qu'une source secondaire est une analyse ou un travail postérieur sur cette entité. Les « sources tertiaires », expression moins fréquente, consistent en des compilations, comme des catalogues ou des index. L'usage de ces trois termes varie grandement au sein des sciences humaines et des pratiques documentaires et archivistiques (University of Maryland University Libraries, 2013). Comme nous le verrons au chapitre 7, les sources primaires peuvent également être des représentations d'originaux ayant disparu ou encore des compilations qui rendent les originaux plus lisibles.

Dans les sciences humaines, une grande partie de la pratique historiographique se consacre au repérage de liens entre des livres, des traités et d'autres documents séculaires au fil de leurs copies, de leurs interprétations, de leurs traductions et de leurs transferts entre cultures et contextes. Il arrive que les sources primaires aient été perdues, détruites ou détériorées depuis longtemps. Les sources secondaires se ramifient en un nombre inconnu de variantes et peuvent avoir été divisées et recomposées de nombreuses manières dans de nombreux buts. Ce qui relève du primaire sera choisi en fonction du contexte et du point de départ du chercheur ou de la chercheuse. La source secondaire de l'un peut être la source primaire d'un autre.

Une particularité importante dans le rapport des sciences humaines aux données est la façon dont elles gèrent l'incertitude dans les représentations des connaissances (Kouw *et al.*, 2013). Cette incertitude peut prendre de nombreuses formes : épistémique, statistique, méthodologique, socioculturelle (Petersen, 2012). Par exemple, l'ambiguïté et l'hétérogénéité de documents historiques peuvent engendrer de l'incertitude. Lorsque des spécialistes des sciences humaines utilisent des technologies conçues pour d'autres formes de recherche, comme des outils statistiques ou des systèmes d'information géographique, elles et ils sont confrontés à un

dilemme : adapter leurs méthodes aux outils ou adapter ces outils à leurs méthodes. De nouveaux outils conduisent à de nouvelles représentations et interprétations. Chaque discipline et chaque scientifique évaluent, collectivement et individuellement, le degré d'incertitude acceptable et déterminent ce qui constitue la « vérité » d'une recherche. Dans les méthodes de recherche et la représentation des données, l'implicite réside dans les choix faits pour diminuer l'incertitude.

Conclusion

Malgré ses cinq siècles d'existence, le terme *data* ou « donnée » n'a toujours pas trouvé une définition consensuelle. Il ne s'agit pas d'un pur concept ; ce n'est pas non plus un objet naturel possédant une essence propre. La synthèse la plus exhaustive consiste à dire que les données sont des représentations d'observations, d'objets ou d'autres entités qui servent à mettre en évidence des phénomènes à des fins de recherche. Selon l'*Oxford English Dictionary*, une *entity* (entité) est « quelque chose qui possède une existence réelle par opposition à un simple attribut, fonction, relation, etc ». Ces entités peuvent avoir une existence matérielle, par exemple des textes sur papier ou papyrus, ou être numériques, comme les signaux émis par des capteurs ou les questionnaires remplis d'une enquête en ligne. Une entité ne devient une donnée que lorsqu'elle est utilisée pour mettre en évidence un phénomène et une même entité peut être la manifestation de plusieurs phénomènes. Par exemple, des photographies trouvées dans un vieux album familial ou un annuaire d'école peuvent devenir des données si un chercheur ou une chercheuse les utilise pour démontrer les modes vestimentaires et capillaires d'une époque. D'autres scientifiques peuvent les exploiter pour attester un regroupement familial ou une identité sociale. Les indications météorologiques dans d'anciens journaux de bord, collectionnés à des fins de navigation et de commerce, servent aujourd'hui de données pour étudier le changement climatique. Des brevets peuvent être une indication de l'époque et du lieu de fabrication d'un objet découvert.

Une liste d'entités pouvant être considérées comme des données ne constitue pas une définition satisfaisante. Pourtant, de telles « définitions » abondent dans la littérature scientifique et les documents d'orientation. L'incapacité à fixer ce concept de manière à clarifier ce que les données sont et ne sont pas dans une situation donnée contribue grandement à la confusion autour des plans de gestion des données, des politiques d'ouverture et de la conservation. C'est dans des contextes opérationnels que l'on trouve, le plus souvent, des définitions concrètes et bornées, comme celle de l'OAIS que nous avons citée plus haut : « une représentation réinterprétable formalisée de l'information, adaptée à la communication, à l'interprétation ou au traitement ».

Les définitions catégorielles, comme celles fondées sur l'origine des données ou les types d'ensembles, sont également utiles dans les contextes pour lesquels elles ont été imaginées. Il est néanmoins nécessaire de s'accorder sur une définition générale afin de peser dans les problématiques locales et mondiales associées aux données de la recherche.

Nous nous efforçons d'utiliser le terme « donnée » ou *data* de façon aussi cohérente que possible. Lorsque nous faisons référence aux données en tant qu'entités, nous utilisons le terme au pluriel, suivant l'usage éditorial standard en anglais (Bryson, 2008) dans la littérature sur la communication savante. Nous utilisons la forme au singulier lorsque nous faisons référence au concept : par exemple « “donnée” n'est pas un mot nouveau » ou « le *big data* est le pétrole d'aujourd'hui ». Cependant, l'usage du mot « donnée » varie en fonction des contextes et des locuteurs, souvent de manière subtile, mais lourde de sens. Nous nous conformons, notamment pour les études de cas, aux usages en vigueur dans la discipline étudiée. À moins que nous l'employions comme concept ou que nous signalions une convention particulière, le terme « données » désigne « des entités utilisées pour mettre en évidence des phénomènes à des fins de recherche ».

3. Le travail scientifique des données

Introduction

« Travail scientifique des données » (*data scholarship*)¹ est un terme que j'ai créé pour exprimer l'agencement complexe des relations entre données et recherche. Les données sont aujourd'hui animées d'une vie propre, indépendamment du contexte scientifique où elles servent à mettre en évidence des phénomènes ; c'est du moins ce qui ressort de la presse grand public. Scientifiques, universités et cabinets de conseil sont tous d'accord pour dire que disposer de suffisamment de données et des techniques adéquates pour les exploiter permet de poser des questions inédites et d'obtenir des indicateurs d'un nouveau genre. Dans certains cas, on peut réaliser des choses extrêmement utiles. Cependant, il peut s'avérer très difficile de déterminer la valeur d'un jeu de données quelconque ou d'anticiper à quoi il pourrait servir.

La notion de travail scientifique des données a d'abord été formulée comme une « recherche à forte concentration de données » dans les initiatives stratégiques du début des années 2000, telles que l'*eScience*, l'*eSocial Science*, les *eHumanities*, l'*eInfrastructure* et la *cyberinfrastructure* (Atkins *et al.*, 2003 ; Edwards *et al.*, 2007 ; Hey et Trefethen, 2005 ; Unsworth *et al.*, 2006). Les trois premiers termes se sont finalement fondus en celui d'*eResearch*. Le programme britannique Digital Social Research a mis en commun des investissements antérieurs dans l'*eSocial Science* : il englobait ainsi la recherche à forte concentration de données dans les sciences sociales et l'étude de l'*eResearch* (Digital Social Research, 2013). Le nom *eScience* désigne souvent le travail des données dans l'ensemble des champs de recherche, y compris les sciences humaines (Crane *et al.*, 2007). *Cyberinfrastructure* reste un terme résolument américain qui recouvre à la fois le travail des données et le cadre technique qui rend ces activités possibles.

Un individu ou une communauté scientifique peut savoir exploiter ses données pour ses besoins propres tout en ignorant quelles données et méthodes, utilisées par des communautés contiguës, pourraient lui servir et vice versa. Passer à de très grands volumes de données produit des différences qualitatives dans les méthodes et les problématiques. Les anciennes approches ne sont plus viables ; pourtant, il faut bien souvent combiner de vieilles données avec des nouvelles.

1. Pour la traduction des concepts de *scholarship* et *data scholarship*, voir la Note de la traductrice p. 7.

L'expertise se transmet difficilement d'un domaine à l'autre, d'une méthode à l'autre. Les personnels de recherche réagissent de façons très diverses à ces difficultés. Certains veulent diffuser largement leurs données, quelle que soit leur forme, pour qu'elles puissent être exploitées à loisir. D'autres préfèrent les conserver par devers eux indéfiniment plutôt que de laisser des personnes extérieures en extraire des trésors qui leur auraient échappé. Beaucoup sont paralysés par le risque de mésusage, d'interprétation erronée, de charge à gérer, de manque d'expertise, de manque d'outils et de ressources, de manque de financement, de perte de contrôle et de pollution de banques de données communes, ainsi que par ce redoutable défi qu'est la pérennisation. Ce n'est qu'aujourd'hui que ces difficultés dans la gestion des données apparaissent au grand jour. Nous comprenons suffisamment certaines d'entre elles pour nous diriger vers des solutions. D'autres semblent insurmontables.

Pour les chercheurs et chercheuses, la gestion des données peut apparaître comme une responsabilité individuelle supplémentaire dans un contexte de réduction des ressources consacrées à la recherche et à l'éducation ; pourtant, le travail scientifique des données est profondément ancré dans les infrastructures de la connaissance. Comme nous l'avons évoqué au chapitre 1, les données suscitent des tensions en raison de l'intérêt que représentent leur possession, leur contrôle et leur accès ; de la difficulté à les transmettre pour des contextes et périodes futurs ; et du besoin de pérennité des données et d'autres contenus scientifiques. Les enjeux sont considérables pour les scientifiques, la population étudiante et les sociétés humaines où la recherche est menée. Les infrastructures de la connaissance fournissent un cadre pour évaluer les interactions sociales et techniques, les implications de la recherche ouverte et la convergence des formes de communication savante.

Les infrastructures de la connaissance

Le terme « infrastructure de la connaissance » s'inspire de travaux antérieurs sur l'information, l'infrastructure et l'Internet : la recherche sur l'infrastructure en soi s'est considérablement développée, l'Internet s'est enraciné dans la vie académique et l'information circule en continu. L'expression « infrastructure de l'information », qui évoquait une vision globale d'un monde, a pris le sens plus restreint d'architectures techniques de communication. De même, en anglais, *national information infrastructure* (infrastructure nationale de l'information) et *global information infrastructure* (infrastructure mondiale de l'information) désignent désormais des politiques états-unien des années 1990 (Borgman, 2000, 2007 ; Busch, 2013 ; Kahin et Foray, 2006 ; Wouters *et al.*, 2012).

Une infrastructure n'est pas un processus conçu pour être cohérent. Elle se rapproche plutôt de l'écosystème, d'un système adaptatif complexe. Elle est constituée de nombreuses composantes qui interagissent, avec plus ou moins de succès, à travers des processus sociaux et techniques. Paul Edwards (2010, p. 17) définit les « infrastructures de la connaissance » (*knowledge infrastructures*) comme des « réseaux robustes de personnes, d'artefacts et d'institutions qui créent, partagent et entretiennent des connaissances spécifiques sur les mondes humains et naturels ». Ces réseaux intègrent la technologie, les activités intellectuelles, l'apprentissage, la collaboration et l'accès décentralisé à l'expertise humaine et à une information étayée. La recherche disciplinaire sur ces idées s'est alors penchée sur trois thèmes : comment les infrastructures de la connaissance évoluent-elles ? Comment renforcent-elles ou redistribuent-elles autorité, influence et pouvoir ? Comment pouvons-nous mieux étudier, connaître et imaginer les infrastructures de la connaissance d'aujourd'hui (et de demain) (Edwards *et al.*, 2013) ?

La donnée est un mode d'information qui semble toujours en mouvement, difficile à fixer en une forme statique. En outre, les limites du savoir tacite et des terrains d'entente sont également mouvantes, puisque de nombreuses parties négocient l'appréhension des données dans de multiples disciplines, domaines et périodes. Les normes de la connaissance n'ont certes jamais été stables, mais elles sont plus complexes encore à établir à l'ère du *big data*. Que signifie « savoir » quelque chose si le résultat du calcul ne peut être parfaitement expliqué ? Que peut-on connaître des origines des données, que faut-il en savoir pour les transmettre dans d'autres contextes ? La « structure de confiance » implicite dans le partage d'informations entre collaborateurs et collaboratrices est ardue à recréer dans des échanges avec des inconnus, en particulier avec d'autres communautés et sur de longues périodes. Certains transferts peuvent se réaliser par l'entremise de la technologie, mais beaucoup dépendent de l'expertise de médiateurs humains, qu'il s'agisse de scientifiques spécialisés, de bibliothécaires, d'archivistes ou d'autres acteurs émergents dans les personnels de recherche. De plus, cet espace est également investi par les intérêts commerciaux.

La notion d'infrastructures de la connaissance recoupe celle de « biens communs de la connaissance », comme nous le développerons au chapitre 4. En effet, un commun est aussi un écosystème complexe, défini simplement comme « une ressource partagée par un groupe confronté à des dilemmes sociaux » (Hess et Ostrom, 2007a, p. 3). Le présent ouvrage est parsemé d'exemples de ces systèmes qui renforcent ou redistribuent autorité, influence et pouvoir. Les personnes possédant des compétences en analyse de mégadonnées sont aujourd'hui recherchées. Les scientifiques capables d'exploiter ces nouvelles formes de données tirent également leur épingle

du jeu. De nouvelles formes de connaissance, comme le *data mining* (exploration de données) et le *crowdsourcing* (production participative) participent à reconfigurer et redessiner le territoire intellectuel.

Des investissements massifs dans l'infrastructure profitent à tous, au sens où la marée montante soulève tous les navires. Mais cette comparaison est contrecarrée par d'autres tendances économiques et sociales, comme la théorie du *knowledge gap* (écart des connaissances) (Ettema et Kline, 1977 ; Tichenor *et al.*, 1970), les compétences médiatiques (Jenkins *et al.*, 2009) et l'effet Matthieu (Merton, 1968, 1988, 1995). De manière générale, les riches s'enrichissent. Ceux qui sont plus à même de tirer profit des technologies et des informations nouvelles bénéficient d'avantages différentiels. L'effet Matthieu existe dans la recherche ; il a d'ailleurs été formulé pour la première fois dans une étude sur les lauréats du prix Nobel. En effet, les individus et les centres d'enseignement supérieur jouissant d'un grand prestige ont tendance à accaparer la reconnaissance et les ressources de manière disproportionnée. Les découvertes réalisées par des institutions d'excellence recevront plus d'attention que les centres moins connus. À l'inverse, les scientifiques travaillant dans des établissements de faible niveau et des pays peu développés possèdent souvent des compétences inférieures et disposent de moins de ressources locales pour exploiter les innovations technologiques.

L'invisibilité pose problème dans la conception et la maintenance des infrastructures d'au moins deux façons. La première est que les infrastructures, par définition, ne sont visibles que lorsqu'elles ne marchent plus (Star et Ruhleder, 1996). Le public n'a souvent pas conscience de sa dépendance vis-à-vis de l'une d'elles, que ce soit le réseau électrique ou l'interopérabilité entre deux instruments, avant que celle-ci ne cesse de fonctionner. La deuxième est la quantité de travail invisible nécessaire à l'entretien des infrastructures, qu'il s'agisse de réseaux électriques, de réseaux d'instruments scientifiques ou de référentiels décentralisés de données de recherche. Les personnes qui jouissent de l'usage de ces infrastructures ignorent souvent le labeur effectué en arrière-plan pour faire fonctionner harmonieusement tous ces éléments. Le travail invisible est une caractéristique saillante des infrastructures de la connaissance parce que l'effort de documentation, d'organisation et de gestion de l'information scientifique est essentiel pour que d'autres puissent la découvrir et l'utiliser. Il constitue à la fois un ciment et une source de frictions dans les collaborations, le développement d'outils, le partage et la réutilisation de données et beaucoup d'autres composantes de l'infrastructure (Bietz et Lee, 2009, 2012 ; Birnholtz et Bietz, 2003 ; Borgman, 2003 ; Edwards *et al.*, 2011 ; Ehrlich et Cash, 1999 ; Lee *et al.*, 2006 ; Paisley, 1980 ; Ribes et Jackson, 2013 ; Star et Strauss, 1999).

« Infrastructures de la connaissance » est un terme nouveau, mais l'idée est ancienne. Dès les premiers instants du questionnement intellectuel, les savants et savantes ont appris à nager – ou à concevoir de meilleurs bateaux – pour ne pas se noyer dans la vague, le déluge, le tsunami de données prêt à s'abattre sur eux. Ils se plaignent d'être bombardés d'information depuis au moins le premier siècle de notre ère, comme le raconte Ann Blair (2010)² dans *Tant de choses à savoir*. La crainte d'une surabondance de livres est apparue bien avant l'époque moderne. Au milieu du ^{xiii}^e siècle, on élabore des solutions au problème de l'information : pages de titre, concordanciers et florilèges. Ces derniers, dont le nom signifie « fleurs choisies », étaient des compilations des bonnes feuilles d'un livre sur un sujet donné. Au ^{xvi}^e siècle, les index étaient devenus courants. Blair étudie comment les premiers savants et savantes ont géré, chacun à leur manière, leurs lectures et leurs interprétations grâce à des mécanismes sophistiqués de prise de note et d'organisation. Comme le rapporte Price (1975) dans *Science Since Babylon*, l'augmentation de l'offre de services d'indexation et de synthèse est allée de pair avec la croissance des revues depuis le milieu du ^{xix}^e siècle. Depuis, les relations entre les métiers d'auteur, d'éditeur, d'indexeur, de bibliothécaire, de libraires et autres n'ont cessé d'évoluer.

Le social et la technique

Le travail scientifique des données est parcouru de tensions entre l'aspect social et la dimension technique. On ne peut que rarement séparer ces facteurs, car ils sont réflexifs et s'influencent mutuellement. L'outil rend possible la création de données, mais c'est la capacité à imaginer quelles données pourraient être recueillies qui engendre l'outil. Plutôt que de tenter de résoudre ce débat ancien, nous prenons comme prémisses de nos provocations l'idée que les aspects sociaux et techniques de la recherche sont indissociables. Les données ne peuvent être comprises sans l'outil et vice versa ; ces deux éléments trouvent leur sens dans la relation qui les unit (Bijker *et al.*, 1987 ; Bijker, 1995 ; Hughes, 1989, 2004 ; Latour et Woolgar, 1979 ; Latour, 1987, 1988, 1993 ; Meyer et Schroeder, 2014 ; Schroeder, 2007).

Bruno Latour (1987) a introduit le terme « technoscience » pour expliquer le rapport de la pratique scientifique aux technologies. En bon philosophe, Latour tend à utiliser le mot « science » pour évoquer pratiquement toutes les formes de recherche savante. Dans l'anglais d'Amérique du Nord, au contraire, les *sciences* désignent en fait les sciences exactes par opposition aux sciences humaines et sociales. On peut aller

2. Référence originale : Blair, Ann M., 2010. *Too Much to Know. Managing Scholarly Information before the Modern Age*. New Haven : Yale University Press.

plus loin dans les distinctions, par exemple en séparant les champs disciplinaires des carrières dans l'ingénierie, la médecine, le droit et la recherche³.

L'attention croissante portée aux données pose des questions plus larges pour l'histoire et la philosophie des sciences. La recherche est un investissement public coûteux. Depuis la seconde guerre mondiale et surtout la fin de la guerre froide, le public exige que la communauté scientifique rende des comptes, il demande à peser dans les orientations de la recherche et à avoir davantage accès aux découvertes (Kwa, 2011 ; Latour, 2004). Dans ce contexte de reconfiguration des relations entre l'entreprise scientifique et la population, les spécialistes des sciences sociales se sont intéressés de plus près au travail de recherche comme objet d'étude. De leur côté, les scientifiques et autres savants et savantes ont accepté d'être étudiés, à la fois pour faire entendre leur voix et pour bénéficier d'un regard externe sur leur activité (Salk, 1986). Depuis le début des années 1950, une littérature de plus en plus foisonnante traite de l'histoire, de la philosophie et de la sociologie des sciences et des technologies (Hackett *et al.*, 2007 ; Latour et Woolgar, 1986 ; Lievrouw et Livingstone, 2002 ; Van Raan, 1988 ; Woolgar, 1988). Elle a, à son tour, donné naissance à des enquêtes sur les pratiques scientifiques dans les sciences humaines et sociales (Borgman, 2009 ; Case, 2006 ; Flor *et al.*, 2010 ; Friedlander, 2009 ; Jirotko *et al.*, 2006 ; Wouters *et al.*, 2012).

Communautés et collaboration

Les politiques, les pratiques et les infrastructures en matière de données se rapportent généralement aux communautés de recherche associées à celles-ci. Les plans de gestion des données en sont un excellent exemple : « La communauté considérée déterminera ce qui constitue une donnée par l'évaluation par les pairs et la gestion de programme » (National Science Foundation, 2010a). De même, les politiques d'archivage numérique évoquent la « communauté désignée » (Consultative Committee for Space Data Systems, 2012). Par ailleurs, les données représentent souvent des « objets-frontières » existant de manière précaire à la limite de plusieurs domaines (Star et Griesemer, 1989). L'étude des rôles des données dans les collaborations fait apparaître les frontières, les champs d'application, les accords et les désaccords des communautés.

La collecte, la création, l'analyse, l'interprétation et la gestion de données supposent une expertise dans le domaine de recherche concerné. De nombreuses compétences

3. NdT : Dans la traduction française, nous adoptons l'usage de Latour et utilisons les termes « science » et « scientifiques » aussi bien pour les sciences humaines et sociales que pour les sciences exactes.

sont nécessaires, certaines théoriques et pratiques, d'autres sociales et techniques. Certaines d'entre elles peuvent être enseignées ou apprises grâce à des livres, des revues ou de la documentation, mais la plus grande partie se rapporte à un savoir profondément ancré et difficile à formuler. Connue sous le nom de « savoir implicite » – une notion elle-même complexe –, cette expertise des plus critiques est souvent la moins transmissible entre communautés et contextes (Agre, 1994 ; Collins et Evans, 2007 ; Darch *et al.*, 2010 ; Duguid, 2005 ; Polanyi, 1966 ; Schmidt, 2012).

La communauté est une construction théorique bien connue des sciences sociales. Les communautés de pratiques et les cultures épistémiques représentent des notions clés de la sociologie des sciences. La « communauté de pratique » est un concept créé par Lave et Wenger pour décrire la manière dont des groupes acquièrent et partagent des connaissances (Lave et Wenger, 1991 ; Wenger, 1998), qui a été largement étudié et approfondi depuis (Osterlund et Carlile, 2005). En revanche, les « cultures épistémiques » ne sont ni des disciplines ni des communautés (Knorr-Cetina, 1999). Il s'agit plutôt d'une série « d'arrangements et de mécanismes » associés à l'élaboration du savoir et qui englobe les individus, les groupes, les artefacts et les technologies (Van House, 2004). Les communautés de pratiques et les cultures épistémiques ont en commun l'idée que le savoir est situé et local. Nancy Van House (2004, p. 40) résume cette perspective de manière succincte : « Il n'existe pas de "vue de nulle part" : le savoir est toujours situé en un lieu, une époque, des conditions, des pratiques et des *a priori*. Il n'existe pas un savoir, mais des savoirs ».

Connaissances et représentation

En dépit des tentatives de marchandisation, les données ne sont des « choses qui brillent » qu'au sens où c'est un thème en vogue qui attire et détourne l'attention (Schrier, 2011 ; Starke, 2013). Signaux, enregistrements, notes, observations, spécimens et autres entités ont été considérés comme des données au terme d'un long processus culturel au sein des domaines de recherche, des disciplines et des spécialités. On appelle « inscriptions » les enregistrements des pratiques scientifiques (Latour et Woolgar, 1979, 1986 ; Latour, 1988 ; Lynch et Woolgar, 1988a, 1988b). Chaque champ de recherche élabore ses propres inscriptions pour consigner, décrire et représenter ce qu'il considère comme des données. Les méthodes courantes de représentation – métadonnées, langages de balisages, formats, étiquetage, espaces de noms, thésaurus, ontologies, etc. – facilitent l'échange de données au sein d'un domaine. Ainsi, une forme de représentation commune peut délimiter une communauté scientifique. Ces mêmes limites peuvent devenir des barrières empêchant d'apporter des données à des champs ayant recours à des formes de représentation concurrentes. Certains médicaments, maladies, animaux, plantes et phénomènes

peuvent être connus sous beaucoup de noms différents. La capacité à combiner des données issues de plusieurs sources dépend donc de ces inscriptions.

Les données, les critères de preuve, les formes de représentation et les pratiques de recherche sont étroitement liés. Les différences entre communautés de recherche n'apparaissent souvent que lorsqu'on tente d'utiliser ou de combiner des sources externes, de collaborer ou d'imposer les pratiques d'une communauté à une autre. Transmettre des connaissances au fil du temps et dans plusieurs contextes est difficile, comme nous l'avons formulé dans la deuxième provocation. Or, les données ne sont pas plus simples à transférer que les autres formes de connaissance. C'est même souvent une tâche particulièrement délicate, car leur sens dépend du dispositif qui les entoure : logiciel, matériel, documentation, publications et ainsi de suite.

Les articles de revue, les communications de conférences, les ouvrages et les autres types de publications sont des lots d'information conçus pour être interprétables de manière indépendante, du moins par le lecteur averti. Ils constituent des représentations du savoir scientifique et comprennent souvent des représentations de données sous des formes adaptées à la diffusion, à la découvrabilité et à l'échange. Les formes adoptées par les publications savantes ont évolué au cours des siècles. Pages de titre, mentions d'auteur, tables des matières, index et autres éléments qui sont aujourd'hui considérés comme des parties intégrantes des ouvrages scientifiques sont apparus graduellement. Certains de ces éléments, comme les mentions de responsabilité, sont passés des ouvrages aux articles avec la parution des premières revues scientifiques, *Le Journal des sçavans* à Paris et *Transactions of the Royal Society* à Londres. Depuis, des infrastructures de la connaissance en évolution se sont multipliées autour de l'édition scientifique. Les maisons d'édition, l'évaluation par les pairs, le renvoi bibliographique, les services d'indexation et de synthèse, les systèmes de recherche d'informations et les mesures d'évaluation comme le facteur d'impact des revues sont autant de composantes de cette infrastructure de la connaissance, ainsi que nous le verrons au chapitre 9.

Théorie, pratique et politique

Le concept de travail scientifique des données transcende la théorie, la pratique et la politique. À petite échelle, les politiques en matière de données correspondent aux choix effectués par les chercheurs et chercheuses quant à ce qu'ils considèrent comme une donnée, ce qu'ils sauvegardent et ce qu'ils conservent ; ce qu'ils partagent, quand et avec qui ; et ce qu'ils déposent, quand et pour combien de temps. À grande échelle, elles consistent dans les décisions des pouvoirs publics et des organismes de financement sur des questions telles que ce qui est considéré comme une

donnée ; ce que les laboratoires doivent préserver ; ce qu'ils doivent diffuser, quand, comment et à qui ; quelles données doivent être conservées, par qui et pour combien de temps ; et comment ces exigences doivent se manifester dans les demandes de financement, les récompenses et les conditions d'utilisation des référentiels de données. Enfin, à l'échelle intermédiaire, elles correspondent aux choix effectués par les institutions de recherche, les universités, les maisons d'édition, les bibliothèques, les référentiels et les autres parties prenantes quant à ce qu'ils considèrent être une donnée et leur rôle dans sa conservation et sa diffusion. Ces acteurs s'appuient eux-mêmes sur des politiques plus larges en matière de financement de la recherche, de propriété intellectuelle, d'innovation, d'économie, de gouvernance et de vie privée.

Les politiques destinées à fluidifier la communication savante, établies par les pouvoirs publics, les organismes de financement, les revues et les institutions, ont souvent une vision simpliste de la marchandisation et de l'échange d'informations. Celles qui ne respectent pas les importantes différences théoriques, pratiques et culturelles entre domaines, bien qu'elles visent généralement l'équité entre communautés et disciplines, risquent d'être mal appliquées, de s'avérer contre-productives ou d'être ignorées par leurs parties prenantes. Par exemple, une communauté de recherche peut avoir sa propre économie morale en matière de collecte, de gestion et de partage des données (Kelty, 2012 ; McCray, 2000). Comme nous l'observerons dans les études de cas et au chapitre 8, les politiques qui régissent actuellement les plans de gestion et le partage des données ont tendance à insister sur la diffusion au détriment de la réutilisation ou de la pérennisation, qui sont des dimensions complexes et coûteuses des infrastructures de la connaissance.

La recherche ouverte

Open access, open source, open data, normes ouvertes, référentiels ouverts, réseaux ouverts, bibliographies ouvertes, annotations ouvertes... la liste ne cesse de s'allonger. Le mouvement de l'*open access* (accès ouvert ou libre accès) existe depuis les années 1970, comme évoqué au chapitre 1. Les infrastructures de la connaissance d'aujourd'hui ont été façonnées par la recherche en accès ouvert ; les améliorations de l'interopérabilité des systèmes, des outils et des services ; les avancées en matière de technologies et de réseaux informatiques distribués ; et la quasi-omniprésence de l'accès à Internet. Elles continuent de les modeler aujourd'hui.

Il n'est pas plus aisé de définir la recherche ouverte que le travail scientifique des données. Elle se rapproche du concept de science ouverte. Pour les besoins du présent ouvrage, la recherche ouverte désigne les politiques et les pratiques

associées à l'édition en accès ouvert, à l'ouverture des données, à leur diffusion et à leur partage. La recherche ouverte est censée accélérer les études, impulser de nouvelles questions et de nouvelles formes d'investigation, diminuer la fraude et les fautes professionnelles, favoriser la croissance d'une main-d'œuvre éduquée au point de vue technique et scientifique et exploiter les investissements publics dans la recherche et l'éducation (David *et al.*, 2010 ; Esanu et Uhler, 2004 ; Nielsen, 2011 ; Boulton *et al.*, 2012 ; Uhler et Schröder, 2007).

Employer un terme unique comme « recherche ouverte » comporte cependant un risque : celui de masquer des différences substantielles entre les nombreuses formes qu'adopte l'*open access*. Les publications et les données jouent chacune un rôle distinct dans la recherche, comme esquissé dans la troisième provocation et comme nous allons l'expliquer ci-dessous. L'accès ouvert aux publications et les données ouvertes visent tous deux à fluidifier la circulation de l'information, à réduire les restrictions dans l'usage de ressources intellectuelles et à rendre les pratiques de recherche plus transparentes. Néanmoins, elles diffèrent de par leurs valeurs respectives pour la recherche, les acteurs qu'elles mobilisent et leur mobilité au fil du temps et dans plusieurs contextes.

L'accès ouvert aux résultats de la recherche

Avec l'apparition des premières revues en 1605, la recherche est sortie du domaine de la correspondance et des réunions privées pour passer à la diffusion ouverte. Les lecteurs et lectrices avaient désormais accès à des livres, des périodiques et d'autres publications grâce aux bibliothèques, aux librairies et aux abonnements. L'échange privé de lettres, de brouillons, de manuscrits et de prépublications s'est poursuivi en parallèle.

L'accès ouvert aux résultats de la recherche a effectué un formidable bond en avant en 1991 avec le lancement d'arXiv, d'abord connu sous son adresse xxx.lanl.gov, avant l'apparition du World Wide Web (Ginsparg, 1994, 2001). Dans les décennies qui ont suivi, arXiv s'est étendu à de nouveaux domaines scientifiques, a quitté les laboratoires nationaux de Los Alamos pour rejoindre l'université Cornell et a obtenu un large soutien de la part d'institutions partenaires. Son usage augmente toujours de manière exponentielle. Près de huit mille articles sont déposés chaque mois sur arXiv et plus de soixante millions ont été téléchargés rien qu'en 2012 (ArXiv.org, 2013).

Nous pouvons tirer du lancement d'arXiv plusieurs leçons importantes pour l'accès ouvert aux données aujourd'hui. La première est que ce système représente une excroissance de la culture d'échanges de manuscrits qui imprégnait les milieux de la

physique des particules. Il s'est constitué sur une infrastructure de la connaissance existante qui permettait la circulation de l'information au sein de réseaux de collègues proches, que l'on appelle collègues invisibles (Crane, 1972).

La deuxième leçon est qu'arXiv a bouleversé cette infrastructure de la connaissance en transformant les relations entre les parties prenantes de la communication savante en physique – auteurs et autrices, maisons d'édition, bibliothèques et lectorat. Les personnels de recherche comme la population étudiante, de pays riches ou pauvres, ont tous pu accéder à des articles bien avant leur date officielle de publication. Étant donné la rapidité de son adoption, les éditeurs et éditrices de revues n'ont eu d'autre choix que d'accepter l'existence d'arXiv. Auparavant, beaucoup refusaient d'examiner des articles publiés en ligne, au motif qu'il s'agissait d'une « publication antérieure ». Des politiques semblables sont encore en vigueur dans de nombreux domaines actuels.

La troisième leçon est que la réussite d'arXiv n'a pas rapidement fait d'émules dans les autres champs de recherche. Les serveurs de prépublication gagnent certes en taille et en popularité dans d'autres disciplines, mais aucun ne s'est implanté aussi profondément que l'original dans les pratiques scientifiques. Néanmoins, même arXiv ne s'est pas intégré dans l'ensemble de la physique, des mathématiques, de l'astronomie ou des autres domaines qu'il couvre. Dans certaines spécialités, il est omniprésent, tandis que dans d'autres, il n'est que marginalement utilisé.

L'accès ouvert aux publications s'est bâti sur ces enseignements. Le concept est simple, mais souvent mal compris, en grande partie à cause des intérêts divergents des nombreuses parties prenantes. La définition de Peter Suber (2012a, p. 4⁴) est la plus succincte : « La littérature en accès ouvert est numérique, en ligne, gratuite et exempte de la plupart des restrictions en termes de droit d'auteur et de licence d'utilisation »⁵. Suber s'empresse de préciser que le libre accès à la littérature scientifique se fait dans un contexte très différent de celui d'autres contenus. La littérature en accès ouvert repose sur un premier postulat : les auteurs et autrices détiennent les droits de leur travail, à moins qu'ils ne les transfèrent à un tiers, par exemple une maison d'édition. Le deuxième postulat est que les scientifiques sont rarement payés pour rédiger des articles de recherche. Ils et elles peuvent diffuser largement leur travail sans perdre de rentrées d'argent, contrairement à la plupart des autres auteurs, autrices, artistes, créateurs et créatrices. En effet, les auteurs et autrices académiques écrivent des articles pour augmenter leur impact ou leur influence et

4. Référence originale : Suber, Peter, 2012a, *Open Access*, Cambridge, MA : MIT Press.

5. Citation traduite par Marie Lebert in Suber, Peter. 2016. *Qu'est-ce que l'accès ouvert ?* [en ligne], Marseille : OpenEdition Press.

non leurs revenus. Toucher un lectorat aussi large que possible est dans leur intérêt. Par ailleurs, les principales sources de financement de la recherche sont les salaires académiques et les subventions. Il est également dans l'intérêt des institutions qui emploient et financent des scientifiques que leurs publications aient le plus grand impact possible.

On peut rendre possible l'accès ouvert à la littérature de nombreuses manières, sous divers modèles de gouvernance et sous de nombreux noms (par exemple, « vert », « doré », « gratuit », « libre », etc.). Ces modèles ont en commun les deux postulats que nous avons évoqués. En général, les auteurs et autrices conservent leur droit d'auteur ou une licence de leur travail diffusé en *open access*. On leur en reconnaît également la paternité. En ce qui concerne l'accès ouvert aux ouvrages scientifiques, aux manuels et aux autres travaux desquels ils tirent généralement des revenus directs, des considérations nuancées s'appliquent (Budapest Open Access Initiative, 2002 ; Directory of Open Access Journals, 2013 ; Howard, 2013a ; Jacobs, 2006 ; Laakso et Björk, 2013 ; Leptin, 2012 ; Pinter, 2012 ; Research Councils UK, 2013 ; Suber, 2012a ; Van Noorden, 2013b ; Wickham et Vincent, 2013 ; Wilbanks, 2006 ; Willinsky, 2006).

Depuis le début des années 2000, un nombre croissant d'établissements de recherche dans le monde adoptent des politiques d'accès ouvert pour les revues de leur faculté. Aux États-Unis, c'est par exemple le cas de Harvard, du Massachusetts Institute of Technology, du California Institute of Technology et de l'université de Californie (Harvard University, 2010 ; MIT Libraries, 2009 ; Caltech, 2013a ; Office of Scholarly Communication, 2013). Ces politiques octroient généralement aux universités une licence non exclusive pour diffuser des travaux, le plus souvent par le truchement d'un dépôt public. En 2012 et 2013, l'accès ouvert aux publications a été bouleversé. En effet, en 2012 au Royaume-Uni, les Research Councils (RCUK) ont annoncé que tous les articles de revues et les communications de conférence découlant entièrement ou partiellement de leur financement seraient soumis à des revues en accès ouvert à partir d'avril 2013. La définition de « revue en accès ouvert » dans cette politique était particulièrement controversée et a depuis été modifiée et réinterprétée à plusieurs reprises. Elle implique en effet des périodes d'embargo, une série de modèles économiques et des subventions intermédiaires (Research Councils UK, 2012a, 2012b). En 2013, l'exécutif états-unien a annoncé une politique similaire d'accès ouvert aux publications issues d'un financement fédéral qui suivait, dans l'ensemble, les périodes d'embargo et les politiques établies par les National Institutes of Health et PubMed Central (Office of Science and Technology Policy, 2013 ; Holdren, 2013b). À l'heure où nous écrivons, l'Union européenne, l'Australie et d'autres pays envisagent des politiques semblables.

Ces diverses politiques, modèles économiques et publications d'un nouveau genre ont abouti à un élargissement de l'accès à la littérature scientifique périodique. En tenant compte de la période d'embargo, environ la moitié des articles de revue sont librement disponibles en ligne dans l'année qui suit leur publication et l'on s'attend à ce que cette proportion augmente (Laakso *et al.*, 2011 ; Van Noorden, 2013a). Cependant, les tensions entre parties prenantes n'ont pas pour autant disparu. Auteurs et autrices continuent de publier en ligne des articles, des mémoires et d'autres travaux auxquels ces politiques d'accès ouvert ne s'appliquent pas et certaines maisons d'édition contrôlent de plus en plus âprement l'accès aux publications dont les droits leur sont réservés (Howard, 2013b ; SHERPA/RoMEO, 2014).

L'accès ouvert aux données

De nombreuses politiques d'accès aux données mises en place par les organismes de financement sont liées à des politiques d'accès ouvert aux publications. Celles du Royaume-Uni sont les plus explicites en la matière (Research Councils UK, 2012b, p. 1) : « Le Gouvernement, conformément à son engagement majeur en faveur de la transparence et de l'ouverture des données, veille à ce que les résultats de recherche publiés soient libres d'accès ». La politique de publication dans des revues en accès ouvert des RCUK exige des auteurs et autrices qu'ils spécifient comment retrouver les données associées à leur publication, mais reconnaît que c'est une tâche complexe (Research Councils UK, 2012b, p. 4) : « veiller à ce que les personnels de recherche envisagent les problématiques d'accès aux données... Cependant, cette politique n'exige pas l'ouverture des données. Si l'on considère qu'il existe des raisons convaincantes de protéger l'accès à celles-ci, par exemple la confidentialité commerciale ou le caractère sensible de données issues de participants humains potentiellement identifiables, ces raisons doivent figurer dans la note ».

Les National Institutes of Health, aux États-Unis, demandent que les publications financées par leurs subventions soient déposées dans PubMed Central. Elles requièrent également que les demandes de financement comportent des plans de gestion des données (Basken, 2012 ; National Institutes of Health, 2013 ; PubMed Central, 2009 ; Zerhouni, 2006). La National Science Foundation a des exigences en matière de plans de gestion des données, mais pas en matière de publication en accès ouvert. Néanmoins, la politique fédérale états-unienne de 2013 s'applique à la NSF, aux NIH et aux autres organismes fédéraux consacrant plus de 100 millions de dollars par an à la recherche et au développement. Elle enjoint à chacun d'eux d'élaborer un plan d'ouverture de l'accès aux publications et aux données numériques scientifiques (Holdren, 2013b ; Burwell *et al.*, 2013).

Cependant, les données ouvertes diffèrent de l'accès ouvert aux articles de revue d'après les deux postulats de Suber (2012a). Alors que les auteurs et autrices détiennent les droits de leurs écrits, au moins au départ, il en va rarement de même pour les données. Comme nous le verrons plus avant au chapitre 9, déterminer qui peut être auteur est une question âprement débattue au sein des disciplines et entre elles. Une fois celle-ci réglée, des droits et des devoirs incombent aux auteurs et aux autrices de travaux. En revanche, déterminer « l'auteur » de données est un sujet qui n'est qu'effleuré dans la plupart des collaborations (Wallis, 2012). Même quand des individus et des groupes attribuent une autorité à des données, les droits et les devoirs afférents peuvent manquer de clarté. De nombreuses formes de données sont créées et contrôlées par des chercheurs et chercheuses, mais leur propriété est une question différente. Certaines formes de données sont considérées comme des faits qui ne relèvent pas du droit d'auteur. Les scientifiques utilisent aussi des données détenues par des tiers ou obtenues dans des ressources communes. Certains types de données, comme des dossiers confidentiels sur des sujets humains, sont contrôlés par les personnels de recherche, mais ne peuvent être diffusés. Les politiques en matière de droits varient selon l'institution, l'organisme de financement, le contrat, le droit applicable et d'autres facteurs.

Le deuxième postulat de Suber est que les auteurs et autrices académiques rédigent des articles de revue et des publications sous de nombreuses formes pour augmenter leur influence, et non leurs revenus. Les scientifiques, leurs employeurs et leurs financeurs ont intérêt à diffuser leurs publications aussi largement que possible. Or, pour la plupart des données, ni l'un ni l'autre de ces principes ne sont vrais. Les articles de revue sont conçus pour être diffusés à un lectorat, tandis que les données doivent être extraites avec difficulté du processus de recherche. Communiquer ces dernières suppose souvent un investissement considérable en plus de la recherche elle-même et de la rédaction. Les données constituent des atouts précieux que l'on accumule tout au long d'une carrière pour ne les diffuser qu'avec prudence, voire pas du tout.

Les modèles de gouvernance pour l'accès ouvert aux données sont, au mieux, balbutiants, comme nous le développerons aux chapitres 4 et 10. La formule « libre d'accès » (*freely accessible*) utilisée dans la politique des RCUK que nous avons citée semble se rapporter davantage à la liberté d'expression qu'à la bière gratuite, pour reprendre l'aphorisme de Stallman mentionné au chapitre 1. Pour tous les communs, les questions essentielles à traiter concernent l'équité, l'efficacité et la pérennité (Hess et Ostrom, 2007a). Quelques domaines seulement sont parvenus à prendre des mesures en faveur de l'équité et de l'efficacité en mettant en place des référentiels pour absorber les données, les conserver et y donner accès. Pour certains

champs de recherche, les échanges privés sont suffisants. D'autres se tournent vers les bibliothèques de recherche. Dans tous les cas, la pérennité est un problème. Certains référentiels bénéficient d'un financement à long terme, mais d'autres seulement à court terme. Certains fournissent leurs données à tous gratuitement, d'autres ne les proposent qu'aux membres des consortiums qui les ont financés. Les données partagées dans des échanges privés ou publiées sur le site web d'un laboratoire peuvent n'être accessibles qu'à court terme (voir chapitre 8).

La notion de données ouvertes est donc tout à fait différente de celle d'accès ouvert à la littérature scientifique. Il n'y a guère de consensus sur ce que signifie « ouvert » pour une donnée. La première explication de ce qu'est l'*open data*, par Peter Murray-Rust et Henry Rzepa (2004), recouvre la plupart des idées qui lui succéderont. Les deux chimistes s'inquiétaient du libre accès et de leur capacité à explorer des données structurées. Lorsque des entités, telles que des molécules, sont représentées de manière à ce qu'un algorithme puisse identifier leur structure, elles deviennent des données que l'on peut avantageusement explorer, extraire et manier. En revanche, si ces mêmes molécules ne sont représentées que sous forme d'images dans un fichier texte, il faut disposer d'yeux humains pour identifier leur structure. Les données ouvertes, prises en tant que « *datument* »⁶ dans la formulation de Murray-Rust et Rzepa, sont des données structurées pour être lisibles par une machine et qui sont libres d'accès.

Murray-Rust et d'autres ont plus tard développé, sous l'égide de l'Open Knowledge Foundation, une définition juridique succincte des données ouvertes : « Une donnée ou un contenu est ouvert si quiconque est libre de l'utiliser, de le réutiliser et de le redistribuer et qui n'est conditionné, au plus, que par l'attribution et le partage à l'identique » (Open Data Commons, 2013). Dans des contextes commerciaux, les définitions sont plus ambiguës : « Donnée ouverte : information lisible par une machine, en particulier des données publiques, qui est mise à disposition d'autrui » (Manyika *et al.*, 2013). Les *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics* (Organisation de coopération et de développement économiques, 2007) formulent l'accès ouvert aux données au moyen de treize principes, comme nous l'expliquons au chapitre 8. Le rapport de l'UK Royal Society intitulé *Science as an Open Enterprise* (Boulton *et al.*, 2012, p. 14) définit la donnée ouverte comme « une donnée qui répond aux critères de l'ouverture intelligente. Les données doivent être accessibles, exploitables, évaluable et intelligibles ». Pour les données biomédicales, cela implique des compromis entre coûts et bénéfices, un déclencheur et un calendrier de diffusion, des moyens de s'assurer de leur qualité, une délimitation des données à inclure, la

6. NdT : mot-valise composé de *data* et *document*.

confidentialité, la protection de la vie privée, la sécurité, la propriété intellectuelle et le droit applicable (Boulton *et al.*, 2011).

L'ouverture peut favoriser la création de données. Par exemple, l'accès ouvert aux textes permet de traiter les entités qu'ils contiennent comme des données. On peut, grâce au recours aux techniques de fouille de texte (*text mining*), localiser tous les articles ou ouvrages qui décrivent une molécule, un objet céleste, une personne, un lieu, un événement ou une autre entité. Pour l'exploration de données, les bases de publications, les archives de données et les collections de livres numérisés commencent à se ressembler (Bourne *et al.*, 2011 ; Bourne, 2005 ; Crane, 2006 ; Murray-Rust *et al.*, 2004). Elles sont cependant très différentes les unes des autres au point de vue de la propriété intellectuelle, comme nous l'étudierons au chapitre 4.

L'ouverture des données peut aussi comprendre la capacité à traiter des représentations d'objets de recherche comme des données, que les objets décrits soient ouverts ou non. Un exemple en est la création de tags ou d'annotations ouvertes sur des publications, des jeux de données et d'autres contenus. Annotations et bibliographies apportent une valeur ajoutée aux éléments décrits en les rendant plus découvrables. Le partage d'annotations a rapidement suscité l'intérêt dans la recherche sur les bibliothèques numériques, ce qui a conduit à des méthodes concurrentes et à des efforts de normalisation pour permettre l'interopérabilité des systèmes d'annotation (Foster et Moreau, 2006 ; Hunter, 2009 ; Phelps et Wilensky, 1997, 2000). Les premiers outils de gestion de bibliographies personnelles, tels que ProCite, BiblioLink, RefWorks et EndNote, supposaient un auteur ou une autrice unique et des fichiers stockés en local. Au début des années 2000, des individus se sont mis à « taguer » et annoter des sites web, des images, des publications et des données et à partager ces tags sur des réseaux sociaux comme Delicious ou Flickr. À la fin des années 2000, bibliographies personnelles et annotations ouvertes convergeaient. Zotero, Mendeley, LibraryThing et d'autres outils permettent de partager bibliographies, tags et notes. Le mouvement de la bibliographie ouverte a reçu un coup de fouet lorsque les bibliothèques nationales se sont mises à diffuser leurs catalogues pour qu'ils soient utilisés librement (Open Bibliography and Open Bibliographic Data, 2013). Le nombre de notices bibliographiques en accès ouvert a augmenté et, avec lui, les données à explorer. Les outils d'annotations sont compatibles avec de plus en plus de types de données (Agosti et Ferro, 2007 ; Das *et al.*, 2009 ; Kurtz et Bollen, 2010 ; Renear et Palmer, 2009 ; Rodriguez *et al.*, 2007).

Les technologies ouvertes

La recherche ouverte s'inscrit dans l'ouverture, sur plus de quarante ans, des réseaux et des technologies. Les récits sur l'origine et la trajectoire d'Internet racontent cette transition (Abbate, 1999 ; Goldsmith et Wu, 2006 ; Kraut *et al.*, 2002 ; Levien *et al.*, 2005 ; MacLean, 2004 ; O'Brien, 2004 ; Odlyzko, 2000 ; National Research Council, 2001 ; Zittrai, 2005). On s'accorde généralement sur le fait que les réseaux informatiques se sont développés grâce à des fonds publics à des fins scientifiques et militaires. Des premières interconnexions internationales à la fin des années 1960 aux changements de politique du début des années 1990, Internet n'a été accessible qu'aux communautés de recherche, aux universités et à l'armée par le truchement de marchés publics. Ces réseaux sont devenus les réseaux nationaux de la recherche et de l'enseignement, ou NREN (pour National Research and Education Networks). D'abord conçu pour répartir les coûteux cycles informatiques sur l'ensemble du réseau, Internet a vu ses capacités s'étendre jusqu'à comprendre le courrier électronique, le transfert de fichiers et d'autres fonctionnalités de même type. D'autres réseaux à commutation de paquets existaient dans le commerce, comme Telenet et Tymnet, et fournissaient des moyens de communication à des entreprises privées pour faciliter leur activité et les nouveaux services d'information, telles les bases de données bibliographiques (Borgman *et al.*, 1984 ; Borgman, 2000).

Les changements de politique intervenus en 1993-1994, sous les titres de *national information infrastructure* et *global information infrastructure*, ont permis l'interconnexion des réseaux publics et commerciaux. Ce furent les débuts de l'Internet commercial et de la conversion de réseaux de communication, jusqu'alors propriété de l'État ou protégés, vers les activités lucratives. Internet fut déclaré « ouvert » à l'interconnexion et à la fourniture de services publics comme privés. L'ouverture du réseau a coïncidé avec la première démonstration du World Wide Web sur les interfaces initiales d'explorateurs. Dans les deux décennies qui ont suivi, les technologies, les capacités et les communautés d'internautes sont allées au-delà des rêves les plus fous de ses concepteurs et conceptrices. Néanmoins, de nouveaux modèles économiques, des fluctuations dans l'équilibre des parties prenantes et des problèmes imprévus en matière de sécurité et de confidentialité redessinent aujourd'hui l'infrastructure (Borgman, 2000 ; Estrin *et al.*, 2010).

Acquérir des données sur des réseaux ouverts ne revient pas à pouvoir les utiliser. Les données et les représentations numériques ne sont lisibles qu'à l'aide de technologies adéquates. Pour interpréter un jeu de données, il faut disposer de nombreuses connaissances sur le matériel utilisé pour les générer, qu'il s'agisse de réseaux de capteurs ou de machines en laboratoire ; sur les logiciels employés pour les encoder

ou les analyser, qu'il s'agisse de traitement d'image ou d'outils statistiques ; et sur les protocoles et le savoir nécessaires pour les combiner. Les technologies évoluent très rapidement, surtout dans les milieux de la recherche. De nombreux instruments produisent des données qui ne peuvent être lues que dans des logiciels propriétaires : utiliser ou réutiliser ces données suppose donc d'avoir accès à la bonne version du logiciel et, parfois, à l'instrument lui-même. Beaucoup d'outils d'analyse sont des logiciels propriétaires et génèrent donc des jeux de données dans des formats exclusifs, quel que soit le degré d'ouverture des données au moment de l'analyse. Les scientifiques conçoivent souvent leurs propres outils et écrivent leur propre code pour résoudre leurs problèmes. Bien que cette pratique soit très efficace à court terme, du code et des outils locaux peuvent s'avérer bien difficiles à maintenir à long terme, car ils se conforment rarement aux normes sectorielles du génie logiciel. Les outils locaux sont souples et adaptables, souvent au détriment de la portabilité entre sites et situations (Easterbrook et Johns, 2009 ; Edwards *et al.*, 2013 ; Segal, 2005, 2009).

Le degré d'ouverture des données, des normes et des technologies influe sur la capacité à échanger des données entre outils, laboratoires, partenaires et au fil du temps. Les normes peuvent améliorer la circulation de l'information au sein de communautés, mais aussi ériger des barrières entre ces dernières. En effet, elles peuvent apparaître prématurées ou inadaptées, créant des obstacles et empêchant l'innovation. L'interopérabilité technique des systèmes et des services a longtemps été le Graal des bibliothèques numériques et du génie logiciel. Elle permet à certaines données et parties prenantes d'accéder au système et empêche d'autres d'y entrer. Les politiques, les pratiques, les normes, les modèles économiques et les intérêts catégoriels sont souvent plus déterminants pour l'interopérabilité que la technologie en elle-même (Brown et Marsden, 2013 ; Busch, 2013 ; DeNardis, 2011 ; Lampland et Star, 2009 ; Libicki, 1995 ; Palfrey et Gasser, 2012).

La convergence des communications

Les communications formelles et informelles sont en train de converger, dans les affaires et l'administration comme dans la recherche. Une entreprise ne peut plus se contenter d'avoir pignon sur rue et d'apparaître dans les quotidiens : pour rester dans la course, il lui faut aussi être présente sur le Web, les réseaux sociaux, les blogs et microblogs et les chaînes vidéo. Les administrations doivent être accessibles à la population dans les grandes villes comme dans les petits quartiers ; il leur faut aussi, avec le développement de l'administration numérique, être présentes en ligne et fournir un service public 24 heures sur 24, 7 jours sur 7. De même, les scientifiques n'exercent plus seulement leur influence à travers la littérature de leurs domaines,

mais aussi sur des pages web, des serveurs de manuscrits, des archives de données, des dépôts institutionnels, des archives de présentations et de schémas, des blogs et microblogs, des réseaux sociaux et d'autres médias au fur et à mesure de leur invention. Les nouvelles technologies favorisent de nouveaux moyens de communication, mais déstabilisent les modèles existants. Quand les nouveaux modèles sont calqués sur les anciens, les métaphores touchent à leurs limites.

Les métaphores de la donnée

Les rôles respectifs des publications et des données dans la communication savante se confondent dans des expressions telles que « publication de donnée » et « publier des données ». Les présupposés simplistes de ces métaphores font courir des risques aux nouveaux modèles de communication savante, comme nous l'avons dit dans la troisième provocation.

Au sens strict, publier signifie « rendre public » ; de nombreux actes peuvent donc être considérés comme une « publication ». Cependant, dans le contexte de la recherche, la publication poursuit trois buts principaux : 1) la légitimation, 2) la diffusion et 3) l'accessibilité, la sauvegarde et la conservation (Borgman, 2007). La première fonction est généralement remplie par l'évaluation par les pairs (*peer review*). Le point fixe de la publication, qui est habituellement le document de référence, matérialise le processus de légitimation, conférant un sceau de qualité qui marque la confiance de la communauté scientifique. Parce qu'ils sont les références légitimes de la recherche, c'est vers ces éléments publiés que pointent les renvois bibliographiques. La deuxième fonction, la diffusion, est essentielle, car la recherche n'existe que si elle est communiquée à autrui (Meadows, 1974, 1998). Les maisons d'édition diffusent la recherche sous forme de revues, de livres, d'actes de colloque et d'autres genres. Les auteurs et autrices diffusent leur propre travail en distribuant leurs publications à leurs collègues, en les mettant en ligne et en les mentionnant dans des présentations, des blogs, sur des réseaux sociaux, etc. La troisième fonction consiste à rendre les travaux disponibles et découvrables, à veiller à ce que des exemplaires soient préservés et, généralement, à s'assurer qu'une copie est conservée pour un usage à long terme. Les auteurs et autrices, les maisons d'édition et les bibliothèques tendent à assumer conjointement la responsabilité de cette dernière fonction. Les scientifiques sont motivés pour publier leur travail, puisque la publication est la forme de reconnaissance primordiale qui permet de bénéficier d'une embauche, d'une promotion ou d'autres récompenses.

La métaphore de la publication des données n'est juste que dans un sens particulier, proche de celui de la publication de revues et d'ouvrages. Par exemple,

l'Organisation de coopération et de développement économiques (OCDE) publie une vaste série de statistiques nationales et internationales sur le produit intérieur brut, l'emploi, le revenu, la démographie, le travail, l'éducation, le commerce, la finance, les prix et ainsi de suite. Divers organismes publics publient des données de recensement et d'autres statistiques de même type. Hors de la sphère scientifique, la publication de données peut faire référence à la dissémination de documents constitués de listes, de faits ou d'annonces plutôt que de récits. Ainsi, une société états-unienne nommée Data Publishing produit depuis 1986 des annuaires téléphoniques (Data Publishing, 2013).

Dès que l'on sort de cette acception restreinte, la métaphore de la publication de données fait long feu. Le plus souvent, l'expression désigne la divulgation de jeux de données associés à un article de revue individuel. Les données peuvent être jointes aux articles, mais elles bénéficient rarement d'un examen indépendant par les pairs, car elles sont difficiles à évaluer. Dans cette acception de la publication de données, ces dernières sont souvent déposées dans une archive et liées à l'article et non publiées en tant que telles. Les jeux de données peuvent être trouvés et conservés, mais ils ne sont pas diffusés comme une entité autonome et ne sont pas autodescriptifs comme le sont les articles de revue. La publication de données peut aussi désigner le téléversement sur le site web de l'auteur ou de l'autrice, auquel cas aucune des trois fonctions de la publication n'est vraiment remplie. Dans quelques cas, l'expression fait référence à une archive qui recueille des données et y donne accès (PANGAEA : « Data Publisher for Earth & Environmental » *Science*, 2013). Les données peuvent alors être trouvées et conservées ; en revanche, l'évaluation par les pairs et la diffusion ne constituent pas, pour la plupart des archives de données, des activités essentielles.

Un argument en faveur de cette métaphore est son côté familier : les scientifiques savent comment publier et citer des articles. L'idée sous-jacente est que cette familiarité encouragera la divulgation de données (Crosas *et al.*, 2013 ; Klump *et al.*, 2006 ; Lawrence *et al.*, 2011 ; Murphy, 2013 ; Parsons et Fox, 2013). Bien que cela soit souvent présenté comme un fait, il n'existe guère d'éléments permettant d'affirmer que la citation de données constitue une incitation à la diffusion. Par ailleurs, la métaphore de la publication promeut les intérêts des maisons d'édition, qui étendraient leurs modèles économiques en faisant des données des unités à diffuser.

De nombreux arguments s'élèvent en défaveur de la métaphore ; pourtant, elle persiste. Les appels à « publier vos données » avec chaque article de revue sont dangereux parce qu'ils consolident un lien binaire entre articles et jeux de données. La mise en correspondance peut servir les intérêts de la communauté dans les domaines où

cette relation binaire existe déjà et où les articles se doivent d'être reproductibles à partir des jeux de données. Cependant, comme nous le verrons au chapitre 9, ces domaines sont rares. Pour qu'une telle correspondance des publications et des données fonctionne, elle doit s'inscrire dans une infrastructure de la connaissance plus vaste qui comprenne l'évaluation par les pairs des jeux de données, des référentiels disponibles, des politiques et des technologies de la part des revues pour faciliter l'établissement des liens et un accès au matériel, aux logiciels et autres dispositifs nécessaires à la reproductibilité.

Une mise en correspondance d'un article de revue avec un jeu de données n'est qu'une des nombreuses relations possibles entre publications et données. Les correspondances sont souvent multivoques. En effet, il peut s'avérer impossible de spécifier de manière complète le champ d'application des données et informations associées à une publication donnée. Les liens directs sont utiles dans la mesure où ils contribuent à la découvrabilité et à la reproductibilité, mais une architecture qui suppose des liens univoques réduit la capacité à découvrir ou à réutiliser des jeux de données à d'autres fins. Le mouvement de l'*open data* repose sur la possibilité de rassembler et de comparer des données issues de nombreuses sources, ce qui nécessite des technologies ouvertes.

La publication n'est que la première des cinq métaphores identifiées par Parsons et Fox (2013) pour désigner l'intendance de données, métaphores qu'ils jugent toutes problématiques et incomplètes. La deuxième est celle du « *big data* » ou « *big iron* »⁷, en référence à la production industrielle et à la culture technique associées aux données en astronomie, en climatologie, en physique des particules et dans des domaines similaires. Le *big iron* s'intéresse avant tout à l'assurance qualité, à la réduction des données, à la gestion de versions, aux standards de données et métadonnées et au débit. « L'infrastructure scientifique » (*science support*) constitue la troisième métaphore, qui se rapporte à des domaines tels que l'écologie de terrain, où séparer la science des données ou les collecteurs des conservateurs peut s'avérer difficile. La « cartographie », la quatrième métaphore de Parsons et Fox, fait référence aux données géospatiales qui sont essentielles aux modèles climatiques, à l'affectation des terres, aux enquêtes et à bien d'autres usages. Ces données sont intégrées en couches et publiées sous forme de cartes plutôt que d'articles ou de textes.

Les « données liées » constituent la dernière métaphore. S'il s'agit de relier des jeux de données et des publications, c'est aussi une méthode participant d'un mouvement plus large pour rassembler des données, des publications et des documentations

7. NdT : un terme argotique en anglais pour désigner les « gros calibres » que sont les superordinateurs.

qui présentent un lien. La notion de données liées est fondamentale pour le Web sémantique. Pour fonctionner, celles-ci reposent sur des modèles graphiques d'organisation et supposent un accord sur les ontologies et les normes. Les données ouvertes occupent une place centrale dans cette vision ; en revanche, la sauvegarde, la conservation et l'assurance qualité n'y sont pas des priorités.

Les unités de données

Les métaphores autour des données illustrent une vision simpliste de l'élément qu'il conviendrait de diffuser, citer, utiliser ou conserver. Les données peuvent être représentées sous forme d'unités de toute taille : pixels, photons, caractères, frappes, lettres, mots, cellules de tableur, jeux ou archives de données. Le terme « jeu de données » même est problématique en ce qu'il fait couramment référence à au moins quatre thèmes – regroupement, contenu, parenté et destination – qui comportent chacun plusieurs catégories (Renear *et al.*, 2010). Les jeux de données peuvent être traités comme des objets indépendants, qu'ils représentent quelques bits ou plusieurs centaines de téraoctets. L'unité adéquate dépend de l'usage qu'on souhaite en faire. Il est parfois utile de rassembler de nombreux éléments pour les comparer et les explorer ; dans d'autres cas, on cherchera plutôt à extraire des portions de ressources plus vastes.

Les ouvrages et les articles de revue, unités de communication commodées qui n'existaient autrefois que sous forme papier, peuvent aujourd'hui être divisés en éléments bien plus petits. Les moteurs de recherche retrouvent les articles sous forme d'entités indépendantes et non comme des parties d'une revue assemblées avec soin par un comité éditorial. Au sein de ces articles, chaque tableau, figure et jeu de données peut avoir ses propres identifiants de manière à pouvoir être trouvé indépendamment, séparé du contexte de la méthode de recherche, de la théorie et des conclusions. Livres, articles et autres formes de texte peuvent être traités comme des unités ou comme des agrégats de données, interrogeables par mot, par expression et par chaîne de caractères.

La frontière entre communication formelle – les publications – et informelle – les présentations et les discussions – continue de se brouiller à mesure que le contenu scientifique s'atomise et est traité comme des données. Articles de revue, prépublications, manuscrits, billets de blogs, *slides*, tableaux, figures, présentations vidéo, tweets, publications sur Facebook et LinkedIn et d'innombrables autres entités peuvent être partagés indépendamment. Les dépôts de *slides* et de figures, tout commerciaux qu'ils soient, rencontrent un vif succès grâce à leur facilité d'utilisation et d'accès et à leur acceptation d'entités utiles difficilement publiables ailleurs. La

souplesse des objets numériques en matière de division en unités et d'établissement de liens promeut de nouvelles formes de communication. Par exemple, quand des revues facturent un prix à la page substantiel pour publier un schéma en couleur – même pour une publication numérique –, les auteurs et autrices peuvent choisir de publier une version en noir et blanc dans l'article et déposer ou publier ailleurs la figure en couleur. Parce que les couleurs sont nécessaires pour interpréter les résultats découverts, ces auteurs optent pour une manière peu coûteuse de toucher leur lectorat, au moins à court terme. Une partie de ces contenus sera pérennisée, mais beaucoup disparaîtront : en effet, les liens se brisent souvent et vite.

Pour répondre à ce problème de désagrégation, la réponse habituelle consiste à réagréger les contenus, à la fois pour reconstituer les relations d'origine entre éléments et pour créer de nouveaux ensembles. Les méthodes des données liées peuvent servir à reconstruire la chaîne de valeur de la recherche, à relier des articles, des données, de la documentation, des protocoles, des prépublications, des présentations et d'autres unités. Cette stratégie convient bien aux éléments que l'on peut aisément mettre en réseau, mais ne constitue pas pour autant une solution générique pour relier des ressources dans différents systèmes et services (Bechhofer *et al.*, 2013 ; Goble *et al.*, 2013 ; Pepe *et al.*, 2010 ; Van de Sompel *et al.*, 2012). De même, l'exploration de données (*data mining*) dans la littérature ouverte permet d'identifier des données dans des textes, des tableaux et des figures, mais pas dans des ressources complémentaires ou dans des archives. Il faut appliquer des stratégies multiples pour régler les problèmes de désagrégation, de réagrégation, de citation et de choix d'unités à publier (Bourne *et al.*, 2011 ; Parsons et Fox, 2013 ; Uhlig, 2012).

Les documents de référence

Alors que les communications savantes se fragmentent, se dispersent et se réagèrent de façon nouvelle, maintenir la publication de référence comme telle devient plus difficile. Les objets cités disparaissent, les liens se brisent et les algorithmes de recherche évoluent souterrainement, protégés par le secret industriel. Les données sont plus instables encore, car elles servent des causes multiples dans des contextes divers. Les publications sont des éléments fixés dans le temps et l'espace au terme d'une longue élaboration. Au moment de la parution, on assigne aux travaux des identifiants uniques et permanents : un *digital object identifier* (DOI, « identifiant numérique d'objet »), des numéros de volume et de page ou d'autres immatriculations. Pour recevoir un identifiant permanent, les unités de données doivent être figées dans le temps, dans une version et sous une forme uniques. Reste toujours à déterminer lesquelles sont probantes.

La prolifération des versions d'un même document n'est pas un problème nouveau. De grands quotidiens, tel le *New York Times*, publient plusieurs éditions dans la journée. Dans la presse, la « dernière édition » du jour était considérée comme le document de référence ; mais dans le monde du numérique, le *New York Times* existe sous de multiples versions imprimées et numériques adaptées à des lectorats locaux, nationaux et internationaux. Chaque article peut évoluer en continu.

Dans l'univers de l'imprimé, il n'y a pas d'ambiguïté sur ce qui constitue le document de référence : le numéro de la revue, les actes de la conférence, le livre publié. Une fois que les documents ont paru, ils le restent. C'est à la version publiée que les renvois bibliographiques font référence. Les lecteurs et lectrices peuvent retracer le flux des idées et des informations en suivant ces références bibliographiques, sachant pertinemment qu'ils peuvent trouver une copie stable de l'œuvre auprès d'une bibliothèque, d'un collègue ou d'une librairie, même des années plus tard. En revanche, aujourd'hui, les articles de revue paraissent en ligne des semaines, voire des mois avant leur mise sous presse, s'ils sont seulement imprimés. Jusqu'à une époque récente, la version en ligne était considérée comme une prévisualisation et la version imprimée constituait le document de référence. Désormais, la version en ligne est devenue la référence et le DOI est en passe de remplacer le référencement par volume, numéro et page.

Les données sont diffusables d'une manière qui rend compte de leur caractère dynamique, bien mieux que ne peuvent le faire les formats statiques des articles de revue. Le *Journal of Visualized Experiments*, par exemple, publie des rapports d'expériences évalués par les pairs qui s'accompagnent de vidéos démonstratives (*JoVE : Peer Reviewed Scientific Video Journal*, 2013). Les auteurs et autrices enrichissent leurs publications de visualisations interactives, comme des PDF 3D qui permettent au lectorat de manipuler les modèles de données astronomiques contenus dans l'article (Goodman *et al.*, 2009). De nombreuses expérimentations de ce type sont en cours (Flor *et al.*, 2010 ; Dutton et Jeffreys, 2010 ; Wouters *et al.*, 2012).

Les méthodes et les résultats de recherche peuvent aussi être divulgués par d'autres moyens que les publications formelles dans les revues et les ouvrages. Par exemple, les vidéos brèves constituent une manière courante de montrer des techniques de laboratoire. D'autres concepts peuvent être illustrés par des dessins et des commentaires audio. Si l'on cherche sur Internet des vidéos concernant la protéomique, les équations différentielles, l'économétrie ou l'intertextualité, on trouvera des dizaines d'enregistrements gratuits, certains visionnés un demi-million de fois. Les médias sociaux sont utilisés par les auteurs et autrices pour promouvoir leurs travaux et par les lecteurs et lectrices pour y découvrir de nouvelles idées, trouvailles et compétences.

Les tweets et les billets de blogs sont devenus un système de notification important pour les contenus scientifiques. Le magazine *Science* organise chaque année *Dance Your PhD*, un concours très en vue de clips musicaux illustrant les thèses de doctorantes et doctorants (Bohannon, 2013a). Ces ressources se trouvent certes aux marges de l'édition scientifique, mais elles occupent une place de plus en plus importante dans les infrastructures de la connaissance des chercheurs et chercheuses.

Conclusion

Les données et la recherche sont indissociables. Elles existent ensemble au sein d'un écosystème complexe, qui doit être entendu comme une infrastructure de la connaissance. Les infrastructures de la connaissance sont prises dans un mouvement perpétuel à mesure des avancées de la recherche et des technologies, des évolutions des politiques et des rééquilibrages des intérêts des parties prenantes. Elles s'adaptent et se transforment sur de longs laps de temps. De même, il est impossible de dissocier les influences sociales et techniques sur ces infrastructures. Les scientifiques appartiennent à de multiples communautés qui se chevauchent, chacune possédant ses propres méthodes, cultures, pratiques et moyens pour représenter ses connaissances. Les représentations communes facilitent le partage des données et d'autres ressources, mais elles créent aussi des frontières entre communautés.

À bien des égards, la recherche est de plus en plus ouverte ; mais chaque forme d'ouverture révèle de nouveaux obstacles à la communication et à l'échange de ressources telles que les données. L'absence de consensus sur la signification du terme « ouvert » est due aux nombreuses caractéristiques qui peuvent faire considérer une entité, une politique, une technologie ou un service comme « ouvert », chacune avec ses circonstances atténuantes. Les communications savantes formelles et informelles convergent par certains aspects et divergent par d'autres. Les métaphores de la publication scientifique font long feu quand on les applique aux données. L'atomisation des unités de communication crée des perspectives nouvelles pour l'agrégation et la désagrégation des connaissances, mais met à mal la conception traditionnelle de ce qui doit constituer la référence du travail de recherche à valoriser, croire et conserver, et de ce qui doit être écarté.

4. La diversité des données

Introduction

Étant donné la diversité des théories, des méthodes, des cultures et des questions dans la recherche, il n'est pas surprenant que les données soient tout aussi variées. Pour être efficaces, les infrastructures de la connaissance doivent fournir un terrain d'entente aux parties prenantes, soutenir autant que possible des méthodes et des normes spécifiques et faciliter, le cas échéant, le transfert de données, de méthodes, d'outils, de technologies et de théories entre communautés, entre contextes et au fil du temps. Dans le meilleur des mondes possibles, les scientifiques pourraient s'appuyer sur des infrastructures de la connaissance de domaines très divers.

Les bonnes pratiques en physique ne conviendront pas forcément à la philologie et vice versa. Les champs de recherche diffèrent par de nombreux aspects, y compris par leur volume, leur variété, leur vitesse (Laney, 2001) et leurs types de données ; leurs habitudes en matière de partage et de réutilisation ; et leurs critères d'attribution de la paternité et de la responsabilité des publications et des données. Une analyse poussée des pratiques en matière de données – comment elles sont recueillies, gérées, utilisées, interprétées, divulguées, réutilisées, déposées, conservées, etc. – permet d'identifier le degré de variation, les méthodologies transversales, les méthodologies nécessitant une importante adaptation aux contraintes locales et les possibilités de transfert. Souvent, la recherche la plus transformatrice est celle qui emprunte des pratiques à d'autres domaines. En quête d'expertise et d'idées, certaines personnes se tournent vers leur communauté et d'autres regardent à l'extérieur. Une approche peut rester relativement inchangée à l'échelle d'une carrière ou s'adapter en permanence à des sources de données, des outils, des questions, des possibilités et des contraintes nouvelles. Des foyers d'innovation peuvent surgir dans des recoins inattendus de l'univers scientifique, mais il en va de même des zones de stagnation. Les provocations du chapitre 1 varient en fonction des domaines et d'autres facteurs. Le présent chapitre développe un modèle qui s'applique aux études de cas en sciences exactes, humaines et sociales des trois chapitres suivants, établissant les fondements des analyses comparatives des politiques et des pratiques en matière de données de la partie III.

Disciplines et données

Les pratiques en matière de données sont tout aussi variables que les autres aspects du travail scientifique. Mener des recherches s'apprend tout au long d'une carrière et requiert une expertise et une expérience profondes. La théorie, les questions, les méthodes et les ressources influent sur le choix des données pour résoudre un problème. À l'inverse, s'apercevoir que quelque chose pourrait servir de donnée influence la manière de poser les questions et d'appliquer les méthodes. Ces décisions reposent souvent sur un savoir tacite et sur des suppositions qu'il est difficile d'explicitier.

Les disciplines sont un moyen commode, bien que très imparfait, de classer les différentes pratiques en matière de données. Cependant, les frontières entre sciences exactes, sciences sociales et sciences humaines sont poreuses et artificielles. Toute tentative de catégorisation générale de l'activité humaine se heurte à des difficultés considérables ; néanmoins, le regroupement reste nécessaire à la formulation d'un argumentaire. Il existe de nombreuses catégorisations des distinctions disciplinaires, dont aucune ne fait réellement consensus. Parmi les dichotomies courantes dans les méthodologies, on citera : recherche fondamentale ou appliquée, axée sur l'hypothèse ou sur la découverte, déductive ou inductive et théorique ou empirique. En dépit de la fréquence de leur usage dans les politiques scientifiques, ces termes sont litigieux depuis le ^{xvi}^e siècle et Francis Bacon.

Donald Stokes (1997) s'est penché sur les origines des pratiques de recherche en Occident. Il a révélé les nombreuses manières dont ces catégories sont traitées, ainsi que la variation des distinctions – comme recherche fondamentale ou appliquée – en fonction de leurs rôles, au fil du temps et des circonstances. En examinant un travail de recherche, Stokes demande si celui-ci est une quête de savoir fondamental et si les chercheurs et chercheuses envisagent les applications de leurs découvertes. Si l'on répond oui à ces deux questions, le travail peut être catégorisé en tant que « recherche fondamentale motivée par l'application », que Stokes appelle le « quadrant de Pasteur ». En effet, Louis Pasteur souhaitait comprendre le fonctionnement de la fermentation, mais il voulait aussi appliquer ces connaissances pour empêcher le lait de tourner. La recherche fondamentale pure, dont Niels Borh est l'illustration par excellence, consiste en la quête d'un savoir fondamental, sans s'inquiéter de s'il trouvera une application. Thomas Edison incarne la recherche appliquée pure, profondément préoccupée par l'utilisation et très peu par le savoir fondamental. Le quatrième quadrant, où la réponse aux deux questions est non, n'est pas vide. Stokes note en effet (p. 74-75) que des scientifiques peuvent simplement se montrer curieux, sans être motivés par des applications ou par la théorie fondamentale. Le

quatrième quadrant est particulièrement intéressant si l'on envisage la dimension temporelle des données. Par exemple, des guides d'observation des oiseaux et des papillons, qui ne recelaient ni applications ni théorie lorsqu'ils ont été conçus, sont plus tard devenus de précieuses traces du changement climatique.

Kwa (2011) remonte l'histoire du raisonnement scientifique jusqu'à ses sources gréco-latines, bien avant les traditions occidentales. S'appuyant sur les travaux d'Alistair Crombie (1994), il examine six styles distincts : déductif, expérimental, hypothético-analogique, taxonomique, statistique et évolutionniste. Plus tard, Meyer et Schroeder (2014) comparent les méthodologies de l'*eResearch* aux styles de savoir de Kwa (et Crombie). Ils appliquent également le modèle d'organisation sociale des sciences de Whitley (2000) à l'*eResearch* de façon transversale. Les dimensions de ce modèle sont la dépendance mutuelle et l'incertitude de la tâche. Certains domaines présentent une dépendance mutuelle élevée et requièrent donc l'agrégation des ressources et la formalisation des structures de gouvernance, tandis que ceux présentant une faible dépendance mutuelle ont moins besoin de ces structures. La dépendance mutuelle interagit avec le degré d'incertitude de la tâche ; ainsi, les domaines où ce degré est faible (c'est-à-dire où il y a davantage de certitude), comme la physique des particules, développent des administrations intégrées.

Thomas Kuhn (1962, 1970) avançait un argument semblable à propos du degré de consensus entre communautés sur les questions dignes d'être étudiées. Tony Becher (1989, 1994) fait partie de ceux qui distinguent entre les méthodologies « dures » et « molles » dans les sciences. Il s'appuie sur les travaux de Biglan (1973) et Kolb (1981) pour y ajouter une dichotomie entre sciences pures et sciences appliquées, créant ainsi un modèle à quadrants analogue à ceux de Stokes et Whitley. Le quadrant dur et pur représente les méthodologies abstraites et réflexives que l'on trouve dans les sciences naturelles ; le quadrant mou et pur correspond au mode concret et réflexif des sciences humaines et sociales ; le dur et appliqué équivaut aux caractéristiques abstraites et actives des métiers à teneur scientifique ; et le mou et appliqué est constitué des modes concrets et actifs pratiqués dans les professions sociales. Bien que tous ces modèles soient fondés sur une étude empirique des domaines, chacun d'eux s'efforce de réduire le travail de communautés diverses et complexes à quelques éléments de base (Furner, 2010).

Les styles de recherche des sciences humaines et sociales n'ont pas bénéficié de la même attention que ceux des sciences exactes. Néanmoins, les quadrants désignant des méthodologies scientifiques s'appliquent aussi aux sciences sociales, notamment aux domaines quantitatifs qui se situent à la limite des sciences exactes. Dans les sciences humaines et les domaines des sciences sociales se rapprochant de

celles-ci, on emploiera surtout les termes de théorie et d'interprétation. L'incertitude de la tâche y est plus élevée et les ressources partagées, moindres. Les chercheurs et chercheuses en sciences humaines travaillent plus souvent seuls et sont moins sujets à la dépendance mutuelle. Toutes ces pratiques évoluent au fil des avancées technologiques en matière de répartition : les équipes s'agrandissent, les ressources communes se développent et les tâches sont davantage systématisées (Beaulieu *et al.*, 2012 ; Burdick *et al.*, 2012 ; Friedlander, 2009 ; Meyer et Schroeder, 2014 ; Unsworth *et al.*, 2006 ; Wouters *et al.*, 2012).

La taille compte

Le *big data* est une question d'échelle et non de taille absolue, comme nous l'avons expliqué au chapitre 1. Les différences d'échelles de données influent à leur tour sur les entités qui deviennent des données, les méthodes appliquées, les formes d'investigation et les fins et objectifs de la recherche. La *big science* (mégascience, science lourde), pour reprendre la terminologie de Price (1963), consiste en des sciences mûres caractérisées par des équipes nombreuses et décentralisées, des collaborations internationales, de vastes recueils de données et de grandes installations instrumentales. Les *little sciences* ou *small sciences* (sciences légères) représentent tout le reste.

Pourtant, ce qu'il y a au milieu est peut-être le plus intéressant. Comme le montreront les études de cas, aucun champ de recherche ne s'inscrit strictement dans les limites de la *big science* ou de la *little science*. Les domaines qui disposent de vastes ensembles de données, d'instrumentation partagée, de partenaires internationaux et de structures de gouvernance complexes présentent aussi des poches d'activité où des chercheurs et chercheuses constituent individuellement des collections locales de données qui ne rejoindront jamais les ressources partagées. De même, un scientifique isolé peut fonder son travail sur d'immenses ensembles recueillis par d'autres.

Les dimensions qui font varier la taille, l'échelle ou le périmètre des données sont influencées par les entités qui deviennent des données, les pratiques qui leur sont associées, leur représentation et leurs modes de diffusion ou de réutilisation. Les facteurs institutionnels, comme les normes et symboles, les intermédiaires, les procédures, les standards et les objets matériels, s'avèrent également importants (Mayernik, 2015). Les dimensions de taille sont indépendantes du domaine ou de la discipline, mais interagissent avec de nombreux autres facteurs, tels que les objectifs du projet, les méthodes de collecte des données et les techniques d'analyse.

Ces dimensions peuvent indiquer si les transferts sont judicieux ou non et s'il est nécessaire d'investir dans les infrastructures de la connaissance.

Objectifs du projet

Les objectifs d'un projet de recherche influent sur le périmètre des preuves à avancer. Par exemple, pour une recherche exploratoire, le ou la scientifique souhaitera recueillir de petits ensembles de données qui décrivent des événements particuliers ou phénomènes, souvent de manière souple. En revanche, si le but est de modéliser des systèmes entiers, il est nécessaire de collecter systématiquement de larges volumes de données. Quand ils cherchent à formuler une problématique, chercheurs et chercheuses se montrent ouverts à de nombreuses sources potentielles. À mesure qu'ils cernent le problème, ils resserrent le périmètre de leur collecte.

Les données météorologiques et climatologiques illustrent combien le périmètre d'une collecte de données peut varier en fonction de l'objectif de recherche. Les données météorologiques sont des mesures locales recueillies afin d'émettre des prévisions à court terme de la température, des précipitations, de la neige, des nuages, des vents, des orages et d'autres phénomènes. Ces prévisions sont rarement fiables à plus de dix jours. Les mesures météorologiques locales peuvent être cohérentes au fil du temps si elles sont prises au même endroit, avec les mêmes instruments, à intervalles réguliers et en suivant des protocoles établis. Cependant, les données recueillies à un endroit peuvent ne pas correspondre à celles venant d'ailleurs, parce que les sites concernés peuvent utiliser des instruments et des protocoles différents pour la collecte.

À la différence des prévisions météorologiques, la climatologie nécessite de rassembler des observations sur des sites dispersés et sur de longues périodes. Les données climatiques peuvent être collectées de manière systématique grâce à des observations satellites ou être dérivées de séries de données météorologiques. On peut alors les saisir dans des modèles de processus physiques pour étudier le climat. Une même observation peut d'ailleurs être entrée dans des modèles multiples, développés par des équipes diverses sur la base de théories et de paramètres différents. Paul Edwards (2010) nous offre un récit exhaustif du siècle de prouesses scientifiques, techniques, sociales et politiques nécessaires à la mise en place d'accords internationaux, lesquels ont permis la collecte cohérente de données indispensable à la climatologie. Aujourd'hui, ces données servent à prédire les tempêtes, les inondations, les sécheresses et d'autres phénomènes climatiques qui transcendent les frontières nationales.

Malgré leurs imperfections, les modèles climatiques représentent une belle réussite au sens où les intérêts locaux et mondiaux ont fini par converger (Edwards, 2013). Il semblerait que cela ne soit pas la norme – les échecs étant consignés moins précisément que les réussites. Davis Ribes et Thomas Finholt (2009) ont ainsi identifié des tensions dans le développement à long terme de quatre grandes infrastructures scientifiques. Au départ, toutes ces infrastructures semblent partager des objectifs communs, du moins pour l'œil extérieur. Mais au fur et à mesure de leur collaboration, des différences dans les buts, les motivations, les communautés d'utilisateurs et utilisatrices, les technologies et d'autres facteurs se sont fait jour. Rapprocher les données hydrographiques – une question essentielle pour le changement climatique – est particulièrement compliqué en raison des divergences méthodologiques fondamentales entre des domaines comme le génie de l'environnement, l'hydrologie, la politique publique, l'architecture et l'urbanisme. Les différences d'objectifs peuvent être attribuées aux dimensions de Stoke – à savoir l'application *versus* le savoir fondamental –, à des facteurs temporels, aux standards, à l'interopérabilité et au contexte local (Jackson et Buyuktur, 2014 ; Jackson *et al.*, 2011 ; Jackson, 2006 ; Lampland et Star, 2009 ; Ribes et Finholt, 2009 ; Ribes et Jackson, 2013).

Collecte des données

Le fait que des observations, des traces ou d'autres entités deviennent des données dépend en partie de qui les collecte et comment. Les mégadonnées (*big data*) sont généralement recueillies par des machines : c'est par exemple le cas des gigantesques volumes de photons captés par des télescopes ou de l'échantillonnage intermittent d'indicateurs de la qualité de l'eau par des réseaux de capteurs. Le développement de la technologie nécessaire peut réclamer des années de travail, mais une fois mise en place, elle produit des données caractérisées par une vitesse et un volume élevés. Ce sont ces données qu'évoquent Parsons et Fox (2013) dans la métaphore du « *big iron* »¹ ; elles s'inscrivent dans une production industrielle et une culture technique. Les individus et les petits laboratoires peuvent également recueillir des mégadonnées, tirées aussi bien de réseaux sociaux que de réseaux de capteurs. À l'extrême opposé se trouve la collecte artisanale de données, qui nécessite un savoir-faire considérable. Il s'agit par exemple d'échantillons d'eau et de sol qui sont prélevés à la main, traités par dilution et centrifugeuse, mis en culture et comptés (Wallis *et al.*, 2013). Les données artisanales locales sont souvent conservées sur place par les laboratoires qui les ont recueillies. La collecte, l'utilisation et l'interprétation de données – *big* ou *little* – dépendent fortement de l'expertise dans le domaine considéré (Beaulieu *et al.*, 2012 ; Collins et Evans, 2007).

1. NdT : « Gros calibre », terme argotique désignant à l'origine des armes à feu avant d'être appliqué aux macroordinateurs.

Les méthodes de collecte peuvent s'avérer déterminantes pour la motivation et la capacité à échanger des données. Hilgartner et Brandt-Rauf (1994) établissent une distinction entre la disponibilité d'une technique et l'avantage concurrentiel qu'elle procure. Le continuum va des « doigts de fée » (*magic hands*), où des techniques nouvelles ne peuvent être appliquées que par des spécialistes, aux « kits », où celles-ci sont suffisamment banalisées pour être vendues comme du matériel standardisé. Les « protocoles normalisés » se situent au milieu de ce continuum. Une technique unique au sein d'un seul laboratoire peut passer du statut de « doigts de fée » à celui de kit. Pritchard, Carver et Anand (2004) ont découvert que ce sont les pratiques de collecte, plus que les domaines, qui fédèrent les groupes. Les données produites par des techniques industrielles, standardisées ou en kit sont mieux documentées que celles provenant de méthodes artisanales et de doigts de fée.

Les unités de données, que nous avons évoquées au chapitre 3, et leurs liens avec les pratiques de publication influent aussi sur ce qui devient des données et les volumes collectés. Dans certains cas, une journée passée sur le terrain fournira suffisamment de matériaux pour rédiger plusieurs textes. Dans d'autres, un article de revue représentera des années de collecte. Un seul gros jeu de données peut être analysé pour répondre à plusieurs questions de recherche, soit seul, soit combiné à d'autres ressources. Des notes prises dans des archives fourniront quelques données ou bien des années de ressources à explorer. Les scientifiques peuvent accumuler des données sur un site, une espèce, une communauté ou un manuscrit précieux tout au long d'une carrière. Il est rare qu'à une excursion – qu'il s'agisse d'un séjour sur le terrain ou d'une visite dans des archives – corresponde une publication. Quand les données sont précieuses, elles sont d'autant plus susceptibles d'être intégrées à de multiples articles, appelés « plus petites unités publiables ». Cette pratique est aussi parfois désignée sous le terme de « saucissonnage » (Owen, 2004).

Le degré de contrôle sur la collecte de données dont dispose la ou le scientifique ou le laboratoire influe aussi sur les pratiques, les motivations et les contraintes de transfert et de réutilisation. Lorsqu'ils récupèrent des données dans un référentiel ou une autre ressource, les chercheurs et chercheuses doivent se fier à la documentation pour savoir comment elles ont été recueillies. La collecte peut être externalisée, par exemple en engageant des sondeurs et sondeuses ou par du *crowdsourcing* (ou production participative). Dans ce dernier cas, les scientifiques trouvent un moyen de former des volontaires au repérage d'oiseaux, d'espèces invasives ou d'autres phénomènes. On leur fournit des outils, comme des sites web et des applications mobiles, et la collecte de données est supervisée et vérifiée par des personnels experts (eBird, 2013 ; What's Invasive!, 2010).

Le *crowdsourcing* permet, sans être généralisable à toutes les collectes, de recueillir de grandes quantités de données en mobilisant un public sensibilisé.

Analyse des données

Les investissements réalisés dans les données au stade de l'analyse, ainsi qu'au cours de la collecte et de l'interprétation, sont également déterminants pour une utilisation, une réutilisation, un échange et une conservation aisés. Certaines données n'ont besoin que d'une intervention minime pour être utiles ; d'autres nécessitent pour cela un nettoyage substantiel, un étalonnage et d'autres opérations. Les décisions de départ sur ce qui sera collecté, comment et pourquoi ont des répercussions sur l'ensemble des processus. Parfois, les mesures seront directes, comme le mesurage de feuilles d'arbres et l'attribution de paramètres géospatiaux et temporels (par exemple, le lieu et l'heure exacte de la cueillette) à celles-ci. D'autres mesures seront indirectes et nécessiteront des modèles d'interprétation, comme la tension électrique d'un capteur qui représente un type de nitrate. Pour interpréter une donnée, il peut être nécessaire de disposer d'informations sur le type d'appareil, son calibrage, la date, l'heure et le lieu de collecte et la méthode de captation de l'échantillon. Certaines seront analysées et documentées immédiatement, donnant lieu à la création d'une trace permanente. D'autres seront accumulées pour analyse ultérieure. Certains échantillons sont fragiles, comme les prélèvements d'eau contenant des protozoaires qui meurent en quelques heures. D'autres, tels les parchemins, restent relativement stables des siècles durant.

Les pratiques d'analyse de données dépendent des objectifs du projet, de la préoccupation pour les applications ou la théorie, de la disponibilité d'outils, d'expertise et d'infrastructures et de nombreux autres facteurs. De même que certaines personnes rangent méticuleusement leur domicile alors que d'autres s'épanouissent dans le chaos, les scientifiques ont des habitudes extrêmement diverses en matière de collecte et d'analyse. Certains organisent leurs données rapidement, distinctement et de manière complète le plus tôt possible. D'autres les laissent s'accumuler des jours, des mois ou des années durant avant de s'atteler à la tâche. Certains rédigent des mémorandums à chaque étape de la collecte et de l'analyse, d'autres les évaluent après coup. Certains mettent en place des procédures et des protocoles clairs pour leur laboratoire ou leur équipe et promeuvent ainsi des pratiques cohérentes qui facilitent les analyses comparatives et longitudinales. D'autres privilégient la flexibilité et exploitent l'ensemble des compétences de chaque membre de l'équipe. Si les pratiques en matière de données tendent à se regrouper par communauté, les détails de la gestion sont aussi divers que les données elles-mêmes (Bowker, 2005 ; Edwards *et al.*, 2011 ; Jackson *et al.*, 2011 ; Mayernik, 2011, 2015 ; Wallis, 2012).

Bien que ce modèle ne soit pas généralisable, certains types d'analyse de données peuvent bénéficier du *crowdsourcing* (ou production participative), comme le montre Zooniverse. Cette plateforme soutient des projets où la classification ou l'interprétation des données n'est pas réalisable par des méthodes informatiques, ne nécessite que peu ou pas de formation et de connaissance du domaine et où les objets peuvent être classés en fonction d'un petit nombre de paramètres. Les projets en cours proposent aux participantes et participants de catégoriser le temps qu'il fait sur Mars, de transcrire des relevés météorologiques issus de vieux journaux de bord, d'écouter des baleines communiquer et d'analyser des données liées au cancer. L'interface ludique incite les gens à donner de leur temps pour ces projets de recherche. Début 2014, plus d'un million de personnes étaient membres de Zooniverse. Les données sont présentées à de multiples bénévoles afin d'être classées. L'exactitude de chaque classification est évaluée, de même que la rapidité et la précision de chaque volontaire (Zooniverse, 2014).

Quand est-ce une donnée ?

Il est rare qu'on puisse définir un instant magique où une chose devient une donnée. En général, cela implique un processus par lequel un ou une scientifique reconnaît qu'une observation, un objet, une trace ou une autre entité pourrait servir à mettre en évidence un phénomène, puis recueille, acquiert, représente, analyse et interprète ces entités en tant que données. J'utiliserai un exemple simple pour illustrer les nombreux instants où les choses deviennent données. Une chercheuse détermine que les jardins publics sont un lieu idéal pour étudier le comportement des enfants. Elle se rend dans l'un d'eux et observe un petit courir sur le terrain de jeu. En fonction de la question de recherche, du phénomène étudié, du cadre théorique, de la méthode et des objectifs de l'étude, beaucoup d'aspects différents de cette observation pourraient devenir des données. Peut-être la chercheuse examinera-t-elle la démarche de l'enfant afin d'évaluer son âge, son état de santé et le type d'activité en cours. Les points de départ et d'arrivée de la course peuvent l'intéresser : le petit donne-t-il l'impression de courir au hasard, entre deux parents ou gardiens ou bien avec d'autres enfants ? Joue-t-il dans son coin ou en interaction avec d'autres ? La chercheuse peut effectuer ses observations sur l'enfant seul, le couple parent-enfant ou un groupe d'enfants comme unité d'analyse.

Le parc peut aussi constituer l'unité d'analyse, pour peu que la chercheuse compare l'activité de différents jardins publics aux démographies différentes ; ou alors elle conduira une étude ethnographique à long terme sur un seul site. Elle pourra rassembler des informations contextuelles comme la date et l'heure, des données

socioéconomiques sur le quartier, les moyens de transport utilisables pour se rendre au parc, des données météorologiques ou encore les niveaux de pollens ou de pollution de l'air.

Les observations des activités de l'enfant peuvent être enregistrées de nombreuses façons, en fonction de ce que la chercheuse connaît de ses sujets de recherche et du degré de consentement qu'elle a obtenu pour les observer. Peut-être s'assiéra-t-elle sur un banc dans un coin discret, prenant en silence des notes dans un carnet. En revanche, si la chercheuse a obtenu le consentement éclairé des participantes et participants, elle pourra recueillir davantage d'informations. De multiples observateurs pourront être présents. Peut-être des caméras, des microphones et d'autres capteurs seront-ils installés sur le terrain de jeu. Les participants pourront même être équipés de capteurs pour enregistrer leur activité, leurs mouvements, leur rythme cardiaque, etc. La chercheuse pourra s'entretenir avec les enfants et leurs gardiens avant ou après le jeu.

Quels aspects des observations de notre chercheuse sur l'enfant constituent des données ? Comme nous l'avons vu au chapitre 3, une observation doit être représentée d'une façon ou d'une autre pour pouvoir être traitée en donnée. L'observatrice humaine saisit mentalement les activités de l'enfant, puis les retranscrit sous forme d'entités sur papier ou sur ordinateur. Celles-ci peuvent prendre la forme de notes narratives, de croquis, d'un tableau d'entrées ou d'autres représentations. Elles peuvent être saisies directement dans un outil d'analyse ou transcrites pour une saisie ultérieure. Si les participantes et participants sont interrogés, on combinera leurs réponses avec les notes. On intégrera des informations contextuelles sur le parc et les conditions au moment de l'observation dans un tableur ou un progiciel statistique. On peut enregistrer l'activité de l'enfant sous forme audio ou vidéo et ajouter automatiquement à ces représentations numériques le lieu, la date, l'heure, le niveau de luminosité et d'autres paramètres. On peut aussi envisager une segmentation manuelle des enregistrements pour les apparier à des données d'autres participants. Ces représentations peuvent être analysées de nombreuses façons, en fonction de l'objectif de la recherche. Par exemple, on peut compiler des modèles détaillés des activités de chaque enfant, de chaque couple ou de chaque parc ; ou bien orienter l'analyse sur des regroupements d'activités et des phénomènes à plus grande échelle. Dans tous les cas, les observations sont sélectives : elles enregistrent les détails présentant un intérêt pour l'étude considérée et ignorent les autres.

Le contexte de la recherche influence donc par beaucoup d'aspects ce qui devient données et les modes d'enregistrement et de représentation de celles-ci. Les liens entre connaissances et données peuvent être catégorisés de nombreuses façons

(Furner, 2010). Aux fins de la description des cas des chapitres suivants, nous classons les facteurs qui influent sur le moment et la manière dont les choses deviennent données dans deux catégories générales. La première regroupe les considérations de distance entre le ou la scientifique et les origines de ses données ; elle inclut les sources et ressources, les métadonnées et la provenance. La deuxième comprend les influences extérieures sur ce qui devient données, avec pour sous-catégories l'économie et la valeur, les droits de propriété et l'éthique.

La distance compte

On peut mesurer la distance entre les origines d'une donnée et son utilisation de nombreuses manières. L'époque, le contexte, la méthode, la théorie, la langue et l'expertise ont leur importance. Par ailleurs, la distance géographique entre collaborateurs et collaboratrices peut également influencer significativement leur capacité à partager, interpréter et utiliser les données (Cummings *et al.*, 2008 ; Hollan et Stornetta, 1992 ; Olson et Olson, 2000). Le terme « origines » est tout aussi problématique ; dans le présent ouvrage, il désigne la première fois où quelqu'un traite quelque chose comme une donnée. Les origines et la distance deviennent des facteurs significatifs de représentation et de réutilisation.

Plus les travaux d'un chercheur ou d'une chercheuse sont proches des origines d'une donnée, quelle que soit la dimension considérée, moins son interprétation dépendra de représentations formelles de connaissances. À une extrémité du spectre, on trouve les scientifiques qui ont une relation personnelle, voire intime, à leurs données, comme les ethnographes, qui consacrent leur carrière à étudier une communauté ou un phénomène. Une chercheuse seule analysant le comportement d'un enfant au parc pourra écrire un article de revue uniquement grâce à sa prise de notes, sans jamais présenter ces dernières d'une manière qui permettrait à d'autres de les interpréter à d'autres fins.

À l'autre extrémité du spectre se trouvent les scientifiques qui ne travaillent qu'avec des ensembles de données rassemblés par d'autres. Ceux-ci sont souvent très vastes, compilés à partir de sources nombreuses sur de longues périodes. Les démographes qui cartographient l'activité des jardins publics d'une ville, d'un département ou d'un pays sont contraints de se reposer sur des représentations de données réalisées par un grand nombre de personnes et institutions, parfois dans des buts très divers. Il est rare que les démographes ou d'autres analystes de mégadonnées aient une connaissance intime des modalités, de la temporalité ou des raisons de la collecte de leurs données. Les utilisateurs et utilisatrices ultérieurs dépendent de représentations ou d'inscriptions de ces données, de

la disponibilité d'un logiciel pour lire les formats numériques et d'autres formes de documentation : autant d'occasions d'interprétations nouvelles, mais aussi d'incompréhensions.

Sources et ressources

Parfois, les scientifiques collectent leurs propres données ; dans d'autres cas, elles et ils en acquièrent des existantes sur un phénomène ; le plus souvent, ils combinent des données nouvelles et anciennes. La distance entre ces deux catégories – ici considérées respectivement comme des sources et des ressources – sert la réflexion sur comment gérer les données, comment les transférer d'un contexte à l'autre et d'une période à l'autre et comment déterminer lesquelles valent la peine d'être préservées. Les « sources » sont les données créées par les chercheurs et chercheuses d'un projet quelconque. Les « ressources » sont des données existantes réutilisées dans un projet. La distinction source-ressource suit l'usage proposé par l'*Oxford English Dictionary*, nous l'avons développée ailleurs (Borgman, 2007).

Chercheurs et chercheuses ont davantage de contrôle sur les sources qu'ils génèrent eux-mêmes que sur les données qu'ils tirent d'autres ressources. En général, les sources sont gérées par les individus ou les laboratoires qui les ont créées, du moins jusqu'au moment du dépôt, de la destruction ou de l'abandon des données. Lorsque les chercheurs ont recours à des ressources qu'ils ne contrôlent pas, qu'il s'agisse de jeux de données personnels créés par d'autres, d'archives, de manuscrits, de flux de médias sociaux ou de documents officiels, ils sont dépendants d'autrui pour accéder à ces ressources et pour obtenir le droit de divulguer, publier ou réutiliser les données générées.

Le fait de constituer une source ou une ressource pour un projet de recherche quelconque dépend de l'utilisation des données. Lorsque des astronomes étudient le ciel, elles et ils se servent de télescopes pour effectuer des observations d'une région déterminée, dans un laps de temps déterminé et à l'aide d'un instrument déterminé. Chaque série d'observations est une source unique et originale de données. En revanche, lorsque d'autres les récupèrent dans des archives astronomiques, ils traitent ces données comme des ressources.

Les traces qui n'ont pas été créées à des fins de recherche peuvent aussi être considérées comme des ressources. Par exemple, les informations démographiques sur le quartier, les bulletins météorologiques et les relevés de la qualité de l'air constituent des ressources pour les personnes qui étudient l'activité d'un parc. De même, des chercheurs et chercheuses qui travaillent sur l'évolution de la langue utilisent des manuscrits historiques comme ressources. Ces documents ont été rédigés pour

communiquer des idées à leurs contemporains : ce sont les chercheurs d'aujourd'hui qui les traitent en ressources pour la recherche textuelle. Cependant, ces derniers peuvent aussi créer de nouvelles sources de données grâce au recueil d'informations jusque-là insaisissables, par exemple en encodant une écriture manuscrite ou en extrayant le texte de documents historiques à l'aide d'algorithmes de reconnaissance optique de caractères.

Métadonnées

« Classer est humain », pour reprendre la formulation de Geoffrey Bowker et Susan Leigh Star (1999, p. 1) quant au problème de « la classification et ses conséquences ». Les êtres humains classifient leurs mondes avec des noms de fichier, des dossiers, des gommettes, des étagères, des placards, des piles de papier et d'autres médias et d'innombrables autres moyens. La première étape de la classification consiste généralement à nommer les choses. Une fois nommées, ces catégories peuvent être organisées en hiérarchies, réseaux, graphes ou autres relations présentant différents degrés de formalité. Dans le cas des données et des relations qui les unissent, ces dénominations peuvent avoir des répercussions profondes sur la capacité à découvrir, échanger et conserver.

Les métadonnées, que l'on peut définir simplement comme « des données sur les données », sont un moyen de nommer les choses et de représenter les données et leurs relations.

Définitions et découvrabilité

La multiplicité des définitions des métadonnées est la manifestation de leurs nombreux usages. Le milieu de la normalisation propose la définition générale suivante : « une métadonnée est une information structurée qui décrit, explique, localise une ressource informationnelle ou la rend facile à trouver, utiliser ou gérer ». (National Information Standards Organization, 2004, p. 1) La NISO distingue ainsi trois principaux types de métadonnées : descriptives, structurelles et administratives. Cette dernière catégorie comprend deux sous-types : métadonnées de gestion des droits et métadonnées de préservation. Le milieu archivistique a recours à une typologie semblable : métadonnées administratives, descriptives, de préservation, techniques et d'utilisation. Les métadonnées techniques comprennent la documentation matérielle et logicielle, les données d'authentification et d'autres informations spécifiques à un système. Les métadonnées d'utilisation recouvrent le suivi d'utilisation et d'utilisateur, la réutilisation de contenu et l'information sur les versions (Gilliland, 2008 ; Gilliland-Swetland, 1998). Les méthodes d'organisation propres aux bibliothèques, comme les règles de catalogage, les index et les thésaurus, constituent des métadonnées descriptives au sens indiqué plus haut, bien que la pratique documentaire

distingue en outre les métadonnées descriptives du catalogage des matières (Svenonius, 2000). Parmi les nombreuses façons d'évaluer les métadonnées scientifiques, on trouve les exigences suivantes : abstraction du schéma ; extensibilité, flexibilité et modularité du schéma ; exhaustivité et suffisance ; simplicité ; échange de données ; recherche de données ; archivage de données et enfin publication de données (Willis *et al.*, 2012).

Ce qui fait d'une unité d'information considérée une donnée ou une métadonnée dépend souvent de ce qui est décrit et à quelles fins. Ainsi, les notices bibliographiques constituent des métadonnées sur les ouvrages, les articles de revue, les actes de colloques et les autres documents qu'elles décrivent. Ces notices peuvent à leur tour être traitées en données pour la recherche sur les réseaux sociaux en montrant qui cite qui ; elles peuvent aussi servir de données à l'étude de la communication savante pour examiner le transfert d'idées entre disciplines au fil du temps (Borgman et Furner, 2002 ; Borgman, 1990). Les données des uns sont les métadonnées des autres et vice versa.

Les distinctions entre données et métadonnées ont constitué parmi les révélations les plus fracassantes de la fuite en 2013 de dossiers classés de la NSA (US National Security Agency, Agence nationale de sécurité des États-Unis) [The NSA Files, 2013]. Si la NSA ne peut légalement recueillir d'informations sur la teneur d'appels téléphoniques ou obtenir sans mandat les noms et adresses de clientes ou clients, elle a le droit – et elle en use – de collecter ce qu'elle considère comme des métadonnées : « le numéro de téléphone de chaque appelant et correspondant ; le numéro de série unique des téléphones impliqués ; l'heure et la durée de chaque appel téléphonique ; et potentiellement, la localisation de chaque participant au moment de l'appel » (Ball, 2013). Les juristes du gouvernement états-unien assimilent ces éléments aux informations figurant sur l'enveloppe d'un courrier. Ces unités d'information, bien qu'elles puissent légalement être considérées comme des métadonnées, peuvent servir de données pour modéliser qui appelle qui, quand, d'où, où et pour combien de temps. L'opinion publique estime qu'il s'agit là de données personnelles : pour ceux qui sont placés sous surveillance, la distinction juridique entre données et métadonnées est peu convaincante (Gallagher, 2013).

Les formes de métadonnées et les méthodes pour les créer correspondent plus ou moins bien aux moyens de collecte de données. Lorsque les données sont recueillies ou générées par des technologies informatiques, cela peut engendrer des métadonnées automatiques. On citera par exemple les historiques d'appels téléphoniques, les transactions de cartes bancaires et les données de sortie d'instruments scientifiques. Les métadonnées automatiques sont intégrées à des fichiers audio, vidéo, texte et

image afin de faciliter la recherche, le traitement et le suivi de licence (Embedded Metadata Initiative, 2013). Les sites de partage de photos et de vidéos, par exemple, les utilisent pour taguer et géolocaliser les fichiers téléversés. Les logiciels de gestion de médias, comme iPhoto, se servent de ces métadonnées pour indiquer l'heure de prise de vue, la date, le lieu, les visages, l'exposition et d'autres informations sur des objets. Les outils d'analyse des données de la recherche n'exploitent pas encore aussi bien les métadonnées automatiques que les produits grand public et les médias sociaux.

Dans le cas où les données sont recueillies manuellement, la création de métadonnées se fait généralement à la main : les scientifiques notent les dates, les heures, les procédures et les résultats de la collecte et du test d'échantillons physiques d'eau ou de sol sur le terrain. Néanmoins, même lorsque les données sont collectées grâce à un processus automatisé, que ce soit par des télescopes ou des enquêtes en ligne, une certaine production de métadonnées manuelles peut rester nécessaire. Les personnels de recherche en service se chargent d'enregistrer les informations contextuelles nécessaires à l'interprétation des données, telles que les conditions météorologiques, les artefacts des instruments et les erreurs logicielles. Les chercheurs et chercheuses attribuent des noms de variables aux réponses d'une enquête et des titres aux lignes et aux colonnes d'un tableur. Pour pouvoir utiliser efficacement des textes numérisés, les lectrices et lecteurs humains doivent souvent étiqueter les chapitres, les sections, les paragraphes, les noms, les lieux et d'autres éléments. Il arrive que des informations importantes sur les objets de recherche, qu'ils soient matériels ou numériques, ne fassent pas partie de l'objet lui-même. Par exemple, selon son ancienneté et son état, un livre imprimé peut ne pas indiquer son auteur, sa date de publication ou son éditeur. L'identité des propriétaires du livre, de l'imprimerie, de l'entreprise de reliure et des annotateurs et annotatrices peut aussi s'avérer nécessaire pour déterminer ses origines. Les catalographes de bouquins effectuent des recherches poussées pour créer des métadonnées qui feront autorité sur de tels objets.

Qu'elles soient générées automatiquement ou à la main, ce sont les métadonnées qui permettent aux autres de découvrir les objets qu'elles représentent. Elles constituent aussi les moyens par lesquels les données sont appariées, comparées et distinguées. Elles peuvent être cumulatives : des représentations sont ajoutées aux données à mesure de leur utilisation et de leur réutilisation dans de nouveaux contextes, migrées vers d'autres formats et technologies et conservées sur de longues périodes. Les métadonnées peuvent devenir plus volumineuses que les données qu'elles représentent.

Communautés et normes

Les métadonnées peuvent combler la distance qui sépare les sources des ressources en formalisant et en normalisant la dénomination et la description des données. Lorsque des communautés développent un thésaurus, une taxonomie, une ontologie ou une autre structure de ce type pour organiser leurs connaissances, elles se mettent d'accord sur les catégories d'entités existantes et sur les relations entre ces catégories. Pour une utilisation locale au sein d'un laboratoire, une simple ontologie de quelques dizaines de termes et relations peut suffire à analyser un jeu de données et à communiquer les résultats. En revanche, pour un usage généralisé dans de grandes communautés qui échangent des données, des ontologies bien plus complexes sont nécessaires. Par exemple, dans les sciences de la vie, les taxonomies d'organismes constituent des formes de métadonnées essentielles. La base de données taxonomiques du National Center for Biotechnology Information de la National Library of Medicine, aux États-Unis, contient la description de plus de 230 000 espèces ayant reçu un nom officiel ; pourtant, on estime que le contenu de cette base ne représente que 10 % des espèces vivantes décrites de notre planète (Bard, 2013 ; Taxonomy Database, 2013). Certains champs des sciences de la vie ont élaboré leurs propres classifications, qui sont de plus en plus souvent des ressources ouvertes et interopérables (National Center for Biotechnology Information, 2013). Néanmoins, elles ne sont pas nécessairement compatibles entre elles pour autant. L'existence d'un si grand nombre d'ontologies en concurrence révèle les divisions profondes au sein des domaines de recherche et entre eux.

Les mécanismes de classification tels que les ontologies, les taxonomies et les thésaurus, ainsi que les structures de métadonnées comme les fichiers FITS en astronomie, la DDI dans les sciences sociales et la TEI dans les sciences humaines (que nous aborderons tous dans les études de cas), facilitent l'interopérabilité entre communautés. En s'accordant sur la dénomination d'objets et sur les relations entre catégories, les chercheurs et chercheuses peuvent découvrir, explorer et combiner des données issues de sources multiples. Le revers de la médaille est que ces mécanismes créent des frictions entre les domaines dont les schémas de métadonnées sont inconciliables. Ainsi, la drosophile ou mouche des fruits est étudiée à tant de fins différentes que plusieurs communautés de recherche ont développé des classifications incompatibles. Certains de ces mécanismes de classement remontent aux échanges de données sur la drosophile au travers des bulletins d'information professionnels au début du ^{xx}e siècle (Kelty, 2012). De nouvelles couches de métadonnées ont émergé pour réunir ces schémas. OpenFlyData, par exemple, est un projet de démonstration de faisabilité qui vise à intégrer dans des systèmes multiples des données sur l'expression des gènes chez la drosophile.

Bien que de nombreux domaines scientifiques disposent de schémas de métadonnées à maturité, d'autres n'en ont pas créé ou n'ont pas appliqué largement ceux qui existent déjà (Bowker, 2005 ; Mayernik, 2011 ; Wallis *et al.*, 2013). Les ensembles de données ressources, de recherche et de référence se distinguent justement par leur degré de déploiement des standards de métadonnées (National Science Board, 2005). Le Digital Curation Center (DCC) tient à jour un important catalogue de spécifications de métadonnées élaborées par des champs individuels des sciences exactes, humaines et sociales. Cette ressource est en outre organisée par spécifications de métadonnées, par profils, par utilisations et par logiciels associés (Digital Curation Center, 2013).

Les investissements réalisés par les scientifiques et leurs communautés dans les métadonnées constituent une indication quant aux données qu'ils jugent précieuses. Cependant, l'inverse n'est pas nécessairement vrai : de nombreuses données de valeur ne se voient guère adjoindre de métadonnées, ce qui limite leur diffusion et leur utilisation (Mayernik, 2011). En effet, la création de métadonnées peut représenter un investissement coûteux, soit pour les individus qui décrivent leurs propres jeux de données, soit pour les communautés qui développent et tiennent à jour un schéma de classification. Ce ne sont pas des ressources statiques : elles doivent être révisées et adaptées en continu à mesure que la langue évolue et que le front de recherche avance. Entretenir un schéma de métadonnées est souvent un travail invisible, essentiel à la réussite de la recherche, mais qui n'est guère reconnu à l'extérieur (Borgman, 2003 ; Star et Strauss, 1999).

Il y a pléthore de cas où les normes en général et les normes informationnelles en particulier créent autant de problèmes qu'elles en résolvent. L'ouvrage *Sorting Things Out* de Bowker et Star (1999) regorge d'exemples de classifications ayant pris un tour politique. Les taxonomies de maladies relèvent moins du savoir médical que des relations de pouvoir entre médecins et personnel infirmier et des frais facturables pour les interventions. La cause de décès officielle peut résulter de la décision de la ou du médecin sur ce qui lui semble socialement acceptable, étant donné que les facteurs sont souvent multiples. Ainsi, les suicides et les morts dues à la syphilis peuvent facilement être attribués à une attaque, à une crise cardiaque, voire à une pneumonie. Dans certains pays, il est plus acceptable de mourir d'un problème digestif que cardiovasculaire (Bowker et Star, 1999). Les actes de naissance, créés à la demande pour une tâche immédiate « faisable », sont tout aussi problématiques (Pine *et al.*, 2014). Derrière chaque mécanisme d'organisation des connaissances se cachent des présupposés tacites et non consignés sur les motifs de création et d'utilisation des données et métadonnées. Lorsque ces dernières sont utilisées à d'autres fins, dans d'autres contextes et à d'autres époques, la documentation sur ces présupposés peut ne pas exister ou s'avérer introuvable. Les statistiques sur les

naissances, les décès, les mariages et le nombre de poissons dans l'océan reposent donc sur des bases chancelantes ; pourtant, la politique publique et la recherche en dépendent. Il en va de même pour toutes les formes de données et de métadonnées. Les métadonnées sont nécessaires à la découverte, à l'interprétation et à l'utilisation des données, mais il faut les manipuler avec précaution. « Les données mordent », comme David Ribes et Steven Jackson (2013) l'ont si bien dit.

Provenance

Les métadonnées peuvent représenter de nombreuses dimensions des données. Pour les infrastructures de la connaissance, il est particulièrement important de pouvoir représenter où les données sont apparues et par quel chemin elles sont parvenues jusqu'à nous. En anglais, le terme *provenance* est couramment utilisé dans le milieu des musées et des archives pour désigner l'historique d'un objet. L'une des méthodes employées par les musées pour vérifier l'authenticité d'une pièce est de consulter les titres de propriété ou de possession. Si une documentation continue permet de remonter aux origines de l'objet, celui-ci a plus de chances d'être authentique. On peut s'intéresser à différents types de provenance : l'historique de conservation, l'historique de création et d'altération d'un objet ou encore l'historique du contexte culturel où il était utilisé (Contadini, 2010).

Le terme « provenance » possède des acceptions à la fois plus restreintes et plus larges que celui de « métadonnée ». Entré dans la langue anglaise au XVIII^e siècle, le mot *provenance* peut désigner simplement, selon l'*Oxford English Dictionary*, l'origine ou l'historique de quelque chose et la documentation se rapportant à cet historique. Dans un sens plus restreint, la provenance peut se référer – en français comme en anglais – à un type de métadonnées qui décrit l'origine des données. Lorsque les objets numériques sont devenus des traces juridiques et scientifiques, la notion de provenance s'est étendue, en théorie comme en pratique. Sur le Web, le concept recouvre différents aspects, tels que l'attribution d'un objet, l'instance qui en est responsable, son origine, les traitements qu'il a subis et le contrôle de version (Moreau, 2010 ; Gil *et al.*, 2010). Connaître ou non la provenance d'un jeu de données peut déterminer, aux yeux des scientifiques, si un résultat est fiable, reproductible et admissible comme preuve ou identifier qui doit être crédité (Buneman *et al.*, 2001 ; Moreau *et al.*, 2008). La citation de données permettrait, entre autres choses, de mieux retracer la provenance (Uhlir, 2012). De même, certains arguments en faveur de la sauvegarde du processus de recherche se fondent sur le besoin de meilleurs registres de provenance (Gamble et Goble, 2011).

Tout comme les musées ont recours aux informations de provenance pour vérifier l'authenticité d'une sculpture, les scientifiques peuvent utiliser la provenance pour

évaluer l'authenticité, la fiabilité et l'utilité d'un jeu de données. Les registres de provenance permettent de connaître les protocoles de collecte des données, qu'il s'agisse de l'observation d'un enfant dans un parc ou de l'expression d'un gène de *Drosophila melanogaster*, des transformations subies par un jeu, des critères de nettoyage et de réduction, des routines logicielles appliquées à chaque étape et de la documentation ou d'autres informations nécessaires à la réplication ou à l'interprétation des résultats. Plus le chercheur ou la chercheuse se trouve loin de la source originelle des données, plus cette réutilisation est dépendante des informations de provenance. Par ailleurs, la traçabilité joue un rôle déterminant lorsqu'on combine des données issues de sources multiples : l'origine de chaque jeu contribuera à décider de son usage, de son analyse et du degré de confiance qui lui est accordé. Enfin, la licence et les autres informations sur les droits peuvent être nécessaires pour déterminer ce qu'il est possible d'utiliser, quand, comment, et avec quelle attribution.

Les influences extérieures

Les scientifiques décident de ce qu'elles ou ils utiliseront comme données dans le cadre des possibilités et des contraintes des infrastructures de la connaissance où ils travaillent. Comme le docteur Pangloss, ils peuvent rêver d'exister dans le meilleur des mondes possibles, mais, comme Candide, ils doivent transiger avec le monde réel et ses influences extérieures (Voltaire, 1759). Ces dernières peuvent avoir une puissante incidence sur ce qui devient données et sur quelles données seront transmises au fil du temps dans différents contextes. À mesure que les technologies, les politiques et les acteurs interagissent, le pouvoir de chaque partie prenante évolue. Les normes sociales, les marchés, le droit, les architectures techniques et le code informatique sont constamment rééquilibrés (Lessig, 2004).

Économie et valeurs

Les données revêtent des valeurs de nombreuses sortes, dont la valeur monétaire n'est qu'un exemple. Une analyse exhaustive de l'économie des données de la recherche serait éminemment utile, mais dépasse largement le cadre du présent ouvrage. Pour nos besoins, les données de la recherche seront ici considérées comme une ressource des communs, notion que nous avons abordée au chapitre 3. En effet, les données répondent clairement à la définition des communs, « une ressource partagée par un groupe confronté à des dilemmes sociaux » (Hess et Ostrom, 2007b, p. 3). Un commun est une ressource partagée qu'il faut régir et non un espace dénué de coûts ou de propriété. Les modèles de production fondés sur les communs coexistent avec les modèles mercantiles ; la question est généralement celle de l'équilibre entre les deux. De même, les façons de faire fondées sur les communs peuvent donner de bons résultats comme des mauvais (Bollier, 2007 ; Disco et Kranakis, 2013 ; Hess

et Ostrom, 2007b). Suivant l'usage du terme « communs » (*commons*) dans un sens extensif par Hess et Ostrom, notre analyse envisagera cette notion à la fois comme un système de ressources partagées et comme un régime de droits de propriété. La concurrence et les « passagers clandestins » sont des dilemmes sociaux récurrents parce que les intérêts ne sont pas tous partagés. Les communs de la connaissance sont confrontés à la marchandisation, à la dégradation et au manque de pérennité. La connaissance, dans le contexte des communs de la connaissance, se définit comme « toutes les idées, informations et données intelligibles, quelle que soit la forme sous laquelle elles sont exprimées ou obtenues » (Hess et Ostrom, 2007b, p. 7).

La marchandisation de la littérature scientifique et la flambée des prix facturés aux bibliothèques ont été des moteurs du mouvement pour l'accès ouvert. Au milieu des années 1990, les débuts de la large diffusion des ressources numériques et la sévérité croissante des régimes de propriété intellectuelle ont stimulé l'étude des communs. À la fin des années 1980, les ressources scientifiques partagées coûtaient très cher et le manque d'accès a conduit à leur sous-utilisation potentielle. Le mouvement de l'*open data* semble traversé par des préoccupations similaires. Les données numériques sont fragiles et risquent d'être perdues ; les rendre pérennes est difficile. Des interrogations s'élèvent : peut-être les données, du génome humain aux plans de lieux publics, sont-elles des ressources partagées trop précieuses pour être traitées en marchandises.

La métaphore de la « tragédie des communs » de Garrett Hardin (1968) est bien plus connue que les nombreux travaux ayant remis en cause ses présupposés et ses conclusions. Les communs peuvent être régis par des acteurs multiples et en concurrence à condition de respecter des principes fondamentaux. Néanmoins, ces derniers ne doivent pas se traduire par des séries de règles pointilleuses. La gouvernance est un processus continu de négociation entre parties prenantes (Hess et Ostrom, 2007b). Nous commençons seulement à puiser des données dans les communs de la connaissance. Le système du World Data Center a d'abord constitué un moyen de formaliser des échanges de données, qui faisaient jusque-là l'objet de troc. Ces modèles de gouvernance continuent de mûrir (Hugo, 2013 ; Shapley et Hart, 1982). Les communautés des sciences de la vie et d'autres disciplines apprennent à administrer les ressources des communs, telles que les grands ensembles de données de référence.

Pour étudier la valeur des données de la recherche, il est important de connaître quelques principes de base de l'économie de l'information. Les biens peuvent être classés en fonction de deux dimensions, dans une simple matrice binaire. La première dimension est le degré auquel des individus peuvent se voir interdire l'usage du

bien considéré. Elle correspond à la distinction classique entre les biens publics purs, comme les couchers de soleil, pour lesquels l'exclusion est difficile, et les biens privés purs, tels que les objets physiques que l'on peut posséder, pour lesquels l'exclusion est aisée. La deuxième dimension est la rivalité ou soustraitabilité (*subtractability*). Si la rivalité est forte (le bien est rival), cela signifie que la consommation par une personne en prive les autres, comme lorsqu'on possède un ordinateur ou une voiture. Si la rivalité est faible (le bien est non rival), l'utilisation par une personne ne gêne pas l'utilisation par d'autres, comme dans le cas de l'abonnement à une revue (Hess et Ostrom, 2007b, p. 8-9). Les caractéristiques économiques de biens informationnels peuvent dépendre autant de leur « emballage » que de leur contenu. Par exemple, un même logiciel peut être distribué en *open source* ou vendu en tant que produit commercial ; la même information peut être diffusée dans un fichier numérique ou dans un ouvrage imprimé. Le fichier est certes non rival, mais une seule personne peut détenir ou lire le livre physique à un moment donné (Disco et Kranakis, 2013 ; Kelty, 2008).

La recherche peut être classée dans n'importe laquelle des quatre catégories de cette matrice binaire, comme le montre le tableau 4.1. Les biens publics, dont l'exclusion est difficile et qui présentent une faible rivalité, sont ceux qui ont reçu le plus d'attention dans les discussions sur la valeur des données. Lorsque les données sont diffusées avec peu de restrictions d'utilisation, elles peuvent être traitées comme des biens publics (Berman *et al.*, 2010 ; David *et al.*, 2010 ; David, 2004a ; Guibault, 2013 ; Murray-Rust *et al.*, 2010 ; Pearson, 2012 ; Wilbanks, 2009, 2011).

Ostrom et d'autres placent les bibliothèques dans la catégorie des réservoirs communs de ressources, qui présentent une rivalité forte et dont l'exclusion est difficile (Hess et Ostrom, 2007b ; Ostrom et Ostrom, 1977). La plupart des référentiels de données se trouvent également dans cette catégorie : ils constituent des réservoirs communs de ressources parce que les communautés investissent dans ces contenus partagés et qu'une gouvernance y est nécessaire. Avec la hausse des coûts des livres et des revues, les communs de la connaissance ont risqué l'enclosure, qui transforme les biens publics en biens privés. Les bibliothèques ne pouvaient plus acquérir autant de ressources et leur capacité à y donner accès est devenue plus restreinte. La réussite des communs de la connaissance dépend de la capacité à limiter les enclosures, à rendre l'exclusion difficile et à pérenniser des modèles de gouvernance efficaces. Bibliothèques, archives, référentiels de données et autres ressources en informations partagées sont confrontés en permanence à la menace des « passagers clandestins », des enclosures et de l'absence de pérennisation (David, 2003, 2004b ; Disco et Kranakis, 2013 ; Kranich, 2004, 2007 ; Lessig, 2001, 2004).

Hess et Ostrom situent les abonnements à des revues dans la catégorie des biens de club, dont l'exclusion est facile et qui présentent une rivalité faible. Les données qui ne sont accessibles qu'en supplément de revues payantes appartiennent aussi à cette catégorie, tout comme les services qui proposent d'accéder à des données moyennant un abonnement. Les référentiels de données qui ne sont accessibles qu'aux membres cotisants d'un consortium peuvent aussi être considérés comme des biens de club. On ajoutera à ces exemples les données commerciales, telles que les flux des médias sociaux. Le dernier quadrant, les biens privés, dont l'exclusion est facile et pour qui la rivalité est forte, consiste en des biens physiques ou consommables dans la plupart des analyses économiques (Hess et Ostrom, 2007b ; Ostrom et Ostrom, 1977). De nombreuses données peuvent être considérées comme des biens privés parce qu'elles ne sont pas prêtes à l'emploi comme un logiciel *plug and play*. Ainsi, l'expertise, les instruments, les logiciels et le matériel peuvent créer des rivalités autour des données.

Tableau 4.1. Types de données dans un commun de la connaissance

Source : Adapté à partir de la matrice des types de biens (Hess et Ostrom, 2007b, p. 9)

		Rivalité	
		Faible	Forte
Exclusion	Difficile	Biens publics	Réservoirs communs de ressources
		Données ouvertes	Référentiels de données
	Facile	Biens de club	Bien privés
		Données sur abonnement	Données concurrentielles, données « brutes »

Dans le meilleur des mondes possibles, les données numériques seraient des biens publics. Paul David (2003, p. 4) estime, dans une analyse des données, de l'information et de la connaissance en tant que biens économiques, que les idées sont l'équivalent du feu plutôt que du charbon. Il cite Thomas Jefferson, qui déclarait en 1813 : « Celui qui reçoit une idée de moi reçoit un enseignement qui ne diminue pas le mien, tout comme celui qui allume sa chandelle à la mienne reçoit de la lumière sans me

plonger dans l'obscurité »². Les données ne sont pas des idées, certes, néanmoins on peut réaliser de multiples copies parfaites d'objets numériques sans dégrader les originaux. Cependant, elles peuvent devenir des biens privés lorsque l'exclusion est aisée et la rivalité forte. En effet, la mainmise exclusive sur des données peut offrir un avantage concurrentiel à des personnels de recherche, des scientifiques, des institutions, des gouvernements ou des sociétés privées. Par exemple, contrôler l'accès à des données sur la découverte de médicaments et des essais cliniques peut déterminer quelle société sera la première sur le marché. Les données qui revêtent une telle valeur scientifique et commerciale sont la préoccupation première de la gouvernance des communs de la connaissance. Lorsque les universités, les organismes publics de financement et les sociétés pharmaceutiques s'allient dans la recherche de médicaments, ils cherchent à diffuser les données dans l'intérêt du public tout en protégeant le commerce privé (Boulton, 2012 ; Edwards *et al.*, 2009 ; Edwards, 2008a ; Goldacre, 2012 ; Thomas, 2013).

Les données de la recherche présentent souvent une rivalité forte en raison du savoir-faire humain et du dispositif technique qui leur sont associés. Même au sein des communautés possédant l'une et l'autre, elles peuvent être traitées comme des marchandises dans une culture du don et du contre-don. Elles peuvent être marchandées et troquées contre d'autres données, un financement ou du capital social (Hilgartner et Brandt-Rauf, 1994 ; Lyman, 1996 ; Wallis *et al.*, 2013). Dans d'autres cas, les données constituent davantage une matière première qu'un bien pouvant être échangé sur un marché. En effet, extirper des données du savoir-faire et du dispositif qui leur sont associés peut nécessiter un investissement considérable dans les représentations, les métadonnées et la provenance. Il faut parfois divulguer en même temps le logiciel, le code, les scripts et d'autres outils. Ici, la distinction entre données et représentations est essentielle : ce sont en effet ces dernières qui sont échangées. Or de nouvelles représentations peuvent être nécessaires pour rendre ces données utiles à d'autres parties.

Ainsi, les quatre catégories dépendent de l'utilisation des données et non de leurs caractéristiques inhérentes. Par exemple, la structure d'une protéine peut constituer un bien privé lorsqu'elle est détenue par une société pharmaceutique, un bien de club quand elle est jointe à un article de revue accessible uniquement sur abonnement (payant), faire partie d'un réservoir commun de ressources quand elle est déposée dans un référentiel comme la Protein Data Bank et constituer un bien public si elle est diffusée comme donnée ouverte avec des restrictions de licence minimales.

2. Traduction de Marie Lebert in Suber, Peter. 2016. *Qu'est-ce que l'accès ouvert ?* [en ligne], Marseille : OpenEdition Press.

Droits de propriété

Les données sont-elles des biens ? Les juristes répondront : « ça dépend ». Le droit, les politiques et les pratiques relatives à la détention, au contrôle et à la diffusion de données varient en fonction des organismes de financement, des gouvernements et des juridictions. Le droit des biens, le droit de la propriété intellectuelle, le droit des bases de données, les contrats et les réglementations s'entremêlent au point d'embrouiller les scientifiques, les juristes, les bibliothécaires, le législatif et le grand public. Dans les politiques en matière de recherche, les droits de propriété et l'économie des données sont étroitement liés (Dalrymple, 2003 ; David et Spence, 2003 ; David, 2003 ; Esanu et Uhler, 2004 ; Scotchmer, 2003 ; Uhler, 2007). Chacun de ces facteurs est susceptible d'avoir des répercussions sur le moment et la manière dont un chercheur ou une chercheuse traite quelque chose en donnée.

Que les données relèvent ou non de la propriété intellectuelle, elles sont souvent traitées comme telles. Même celles dites « ouvertes » sont mises sous licence par les personnels de recherche, les universités et d'autres parties (Ball, 2012 ; Hirtle, 2011 ; Korn et Oppenheim, 2011 ; Wilbanks, 2013). La confusion qui règne autour de la propriété, des droits, des devoirs et des conditions de diffusion des données constitue un obstacle significatif pour l'accès à celles-ci (David et Spence, 2003 ; Erdos, 2013a, 2013b). Illustrons les problèmes de propriété qui peuvent se poser lors de la collecte ou de la création des données par quelques exemples. Imaginons qu'une chercheuse souhaite acquérir des informations à partir de bases de données d'éditeurs ou de réseaux sociaux (par exemple Twitter ou Facebook). Elle pourra peut-être le faire selon les clauses contractées avec les sociétés qui les gèrent. Si elle acquiert des données grâce à une interface de programmation d'application (API, pour *application programming interface*), en copiant des sites web ou par d'autres moyens automatisés, le contenu n'en reste pas moins soumis au droit des biens, bien que les droits de propriété puissent être plus difficiles à déterminer ou à faire respecter. Des clauses semblables peuvent s'appliquer à des données issues d'organismes publics, comme les notes d'élèves rassemblées par les académies. Les chercheurs et chercheuses obtiennent ces données dans le cadre de contrats qui régissent leur utilisation, les personnes qui y recourent, la période, les modalités de publication et les conditions de dépôt ou de partage.

Pour reproduire un contenu dans une publication, il faut souvent obtenir l'autorisation écrite de son ayant droit, ce qui peut s'avérer compliqué. Si cette personne peut être identifiée, contactée et qu'elle accorde sa permission, l'auteur ou l'autrice suivra les consignes de la maison d'édition et les lois en vigueur. Les règles peuvent différer selon qu'il s'agit de textes, de photographies, d'images, d'enregistrements audio et vidéo ou d'autres genres de contenu. Disposer de principes directeurs pour différents

types de contenu est utile, mais il y a peu de chances que des règles définitives soient instaurées, puisque la technologie évolue plus rapidement que le droit (Hirtle *et al.*, 2009 ; Visual Resources Association, 2012).

Travailler avec des contenus suffisamment anciens pour relever du domaine public ne dégage pas les scientifiques des contraintes liées à la propriété. En effet, une représentation d'objet peut être soumise au droit d'auteur indépendamment du statut de l'objet lui-même. Ainsi, les archives et les musées peuvent vendre ou autoriser l'usage de versions numériques de manuscrits ou d'autres objets de collection qu'ils possèdent. Un ou une artiste peut céder un tableau à un musée tout en conservant le droit d'en faire des reproductions. Pour publier une photographie d'un objet, un auteur ou une autrice peut avoir besoin de l'autorisation de la ou du propriétaire de celui-ci, de la ou du photographe et parfois d'autres parties encore. Les musées restreignent l'accès aux images des œuvres de leurs collections, mais commencent à partager davantage de contenus en accès ouvert (Getty Research Institute, 2013 ; Whalen, 2009).

Le problème de propriété le plus compliqué pour les scientifiques concerne les « œuvres orphelines », c'est-à-dire des œuvres dont on ne peut déterminer le statut au regard de la propriété intellectuelle ou, si on y parvient, dont l'ayant droit ne peut être identifié ou contacté, ou dont on n'obtient aucune réponse (Hirtle, 2012 ; Society of American Archivists, 2009 ; United States Copyright Office, 2006). Plus de 75 % de tous les livres connus seraient des œuvres orphelines (Grafton, 2007) et jusqu'aux deux tiers du corpus de Google Books (Courant, 2009 ; Lewis *et al.*, 2010 ; Samuelson, 2009a, 2010).

Les œuvres orphelines sont un exemple des anticommons, c'est-à-dire de cas où des ressources informationnelles rares sont sous-employées en raison d'un excès de droits de propriété intellectuelle ou de brevets (Hess et Ostrom, 2007a). Lorsqu'elles et ils ne peuvent reproduire, réutiliser ou citer des contenus qui les intéressent, les scientifiques peuvent être contraints de sélectionner des contenus moins adaptés à leur piste de recherche, voire d'abandonner complètement leur investigation (J. E. Cohen, 2012 ; Lessig, 2001, 2004 ; Nisbet, 2005). Les données régies à la manière de réservoirs communs de ressources, comme les données météorologiques ou les génomes de drosophiles, peuvent se retrouver sous-utilisées si leurs membres s'efforcent d'en restreindre l'accès ou de violer les règles de gouvernance (Edwards, 2013 ; Kelty, 2012).

Une solution partielle au problème des œuvres orphelines consiste à intégrer des informations de licence sous forme de métadonnées dans les documents

numériques. Ainsi, l'être humain ou le moteur de recherche peut retrouver des objets numériques qui répondent à ses besoins en matière d'usage, dans la mesure où les droits de propriété et les informations de licence peuvent être spécifiés. En théorie, les scientifiques pourraient distinguer les objets utilisables sans restriction de ceux qui ne le sont pas et de ceux qui peuvent être utilisés sous certaines conditions de licence. Cependant, ces solutions techniques posent problème. En effet, encoder des informations de licence s'avère plus difficile que ce que l'on pensait. Les licences Creative Commons – la forme de licence en accès ouvert la plus utilisée au monde – sont conçues pour être lisibles par des machines, par des êtres humains et par des juristes (Creative Commons, 2013). On peut identifier des catégories générales d'utilisation, mais la réutilisation d'objets peut être soumise à des conditions supplémentaires. L'accès ouvert aux articles de revue est difficile à encadrer parce que ceux-ci peuvent être « ouverts » à différents égards. Certains sont accessibles gratuitement, parfois sous forme de contenu structuré et interrogeable. La licence de certains articles, livres et autres travaux peut autoriser une utilisation de l'œuvre entière, d'autres permettent la réutilisation de sous-ensembles, et toutes peuvent avoir des conditions d'attribution différentes. L'atomisation de licences de propriété intellectuelle, parfois appelée « cumul de licences », a créé des barrières juridiques et techniques à la réutilisation des données (Van Houweling, 2009). Pour compliquer plus encore la tâche, il existe une distinction juridique entre la lecture de textes pour leur contenu et « l'utilisation sans consommation », telle que l'indexation ou les analyses statistiques qui « n'exploitent pas la finalité créative et expressive de l'œuvre » (Australian Law Reform Commission, 2014). Les exceptions au droit de la propriété intellectuelle à des fins d'utilisation sans consommation ont constitué un point saillant de l'affaire Google Books (Samuelson, 2009b).

Les scientifiques rencontrent ces problèmes de droits de propriété dans le cadre du contrôle et de la diffusion de leurs propres données, ainsi que dans le cadre de l'accès aux contenus qu'elles et ils veulent utiliser comme données. Il ne devrait pas être nécessaire de recourir à des juristes spécialisés en propriété intellectuelle pour déterminer quelles données on peut ou doit divulguer, quand et comment. Les scientifiques peuvent se montrer réticents à exploiter des données dont on ne peut pas déterminer les droits, qui sont difficiles à obtenir ou pour lesquelles ils risquent de porter indûment la responsabilité. Les droits et les devoirs associés aux données se trouvent au premier plan des préoccupations des politiques de l'information.

Éthique

La perception de ce qui est bien et de ce qui est mal dans une situation donnée varie de façon incommensurable selon les époques, les cultures et les circonstances. Le choix de données de recherche n'est pas épargné par ces considérations éthiques.

Ainsi, des méthodes de collecte jugées novatrices il y a quelques décennies, comme l'étude de Tuskegee sur la syphilis (Jones, 1981) ou les expériences de Milgram (1947) en psychologie sociale, ont plus tard été regardées comme abominables. S'il s'agit là d'exemples extrêmes dont le caractère inacceptable apparaît clairement au vu des normes d'aujourd'hui, de nombreux domaines se trouvent en zone grise, tandis que d'autres sont soumis à une réglementation et à la législation.

La déontologie quant à ce qui peut être traité en donnée, comment, quand et pourquoi évolue rapidement maintenant que les enregistrements numériques deviennent la norme et que le *data mining* gagne en sophistication. Les données personnelles sont les plus litigieuses. Les notions de confidentialité, d'anonymat et de vie privée évoluent de manière imprévue. Des données concernant des sujets humains, autrefois considérées comme parfaitement anonymisées, peuvent maintenant être identifiées en les combinant à d'autres sources, ce qui permet aux *data miners* de constituer des profils très complets sur les individus. Dans certains domaines, la divulgation des données est soumise à des restrictions sévères, tandis que dans d'autres, les sujets de recherche sont nommés et identifiés, soit pour leur attribuer le mérite de leurs contributions, soit pour faciliter la combinaison des données. Certains chercheurs et chercheuses analysent les données des réseaux sociaux recueillies par des sociétés privées, alors que d'autres remettent en cause la moralité de ces sociétés autorisées à collecter ces données personnelles (Bruckman *et al.*, à paraître ; Ohm, 2010 ; Wilbanks, 2009 ; Zimmer, 2010).

La recherche sur des personnes vivantes – appelées sujets humains – est réglementée afin que les personnes soient traitées de manière éthique et que les informations recueillies sur elles restent confidentielles. Les personnels de recherche doivent respecter la vie privée au cours de la collecte, de la communication et de la diffusion des données. Il est généralement obligatoire de recueillir le consentement éclairé des sujets. Des comités locaux de protection des personnes ou leurs équivalents (*human subjects review committees*) ont souvent le dernier mot sur les données pouvant être divulguées et celles devant être détruites à la fin du projet de recherche, ce qui conduit à des applications inégales des règles déontologiques (Association of Internet Researchers, 2012 ; National Research Council, 2013 ; US Department of Health and Human Services, 1979).

Le recours d'une étude à des sujets humains dépend d'abord de la définition des termes « recherche » et « sujets humains ». Ces termes sont clairement établis dans les réglementations de tout organisme de financement, instance décisionnelle et comité local de protection des personnes. Les spécialistes des sciences sociales assimilent ces règles et procédures au cours de leur formation universitaire.

Elles et ils y apprennent les règles en vigueur dans leurs disciplines, organismes et universités. S'ils partent dans un autre établissement ou un autre pays, ou s'associent avec des scientifiques venus d'ailleurs, ils apprennent souvent de nouvelles règles ou des variantes de celles-ci.

Les spécialistes des sciences humaines, des sciences exactes et des technologies sont généralement moins au fait de ces réglementations, car elles ne sont pas ancrées dans leur formation universitaire ou leurs pratiques de recherches. Les données que l'on peut recueillir dans le respect de la déontologie ne seront pas les mêmes d'une discipline à l'autre. Les chercheurs et chercheuses peuvent ignorer que leurs activités de recherche sont soumises à une réglementation particulière. Si une équipe d'ingénieurs installe des caméras dans un vestibule sur un campus afin de perfectionner des algorithmes de reconnaissance faciale, est-ce de la recherche sur des sujets humains ? La localisation du vestibule dans le bâtiment du département d'ingénierie ou dans un dortoir a-t-elle une importance ? Le fait que les données soient détruites à la fin du projet change-t-il la donne ? Et qui a accès à ces données ? Si un étudiant ou une étudiante en philosophie observe des enfants faire du sport dans un jardin public pour étudier leur gestion du conflit, s'agit-il de recherche sur des sujets humains ? Le fait d'avoir obtenu l'autorisation du parc, de l'équipe ou des parents a-t-il une importance ? Faut-il prévoir un formulaire de consentement pour chaque enfant et parent ? La réponse à la plupart de ces questions sera un juridiquement correct « ça dépend ».

La diffusion de données associées à la recherche sur des animaux, des espèces menacées, des déchets toxiques, des maladies infectieuses et d'autres sujets sensibles fait l'objet de réglementations et de préoccupations éthiques (J. Cohen, 2012 ; Enserink et Cohen, 2012 ; Enserink, 2006, 2012b ; Fouchier *et al.*, 2012). Les données climatiques sont devenues sensibles ces dernières années en raison de la dimension politique du réchauffement de la planète. En effet, des cas où des données ont été sorties de leur contexte ou délibérément faussées ont incité à la prudence quant à la diffusion de celles-ci (Kintisch, 2010).

Les réservoirs communs de ressources, au sens économique, peuvent être très litigieux. Des sources précieuses de données scientifiques, telles que les sites archéologiques, les manuscrits anciens et les espèces rares, peuvent être considérées soit comme appartenant au patrimoine mondial, soit comme la propriété culturelle d'un pays ou d'une population, soit comme des biens privés. Les avis divergent fortement sur la moralité du fait de contrôler ces sources pour obtenir un avantage concurrentiel. Dans certains domaines et dans certains pays, protéger indéfiniment des contenus pour les exploiter est une pratique acceptée, quand d'autres la

jugent contraire à l'éthique. Les pratiques varient selon les époques et les lieux et se heurtent dans le cadre des collaborations internationales et interdisciplinaires. À mesure que les échanges de données s'accélèrent, que les technologies gagnent en sophistication et se diffusent plus largement et que les frontières nationales et disciplinaires se brouillent, ces tensions éthiques vont vraisemblablement s'accroître.

Conclusion

La diversité des données n'est pas seulement liée à la variété des méthodes de recherche dans l'ensemble des domaines, mais aussi aux nombreuses manières dont des observations, des objets, des traces, des textes, des spécimens, des personnes, des animaux ou d'autres entités peuvent être représentées en tant que données. Les disciplines se distinguent par ces facteurs et par d'autres encore, comme les objectifs de leurs projets de recherche, leurs méthodes de collecte et d'analyse de données et leurs choix en faveur de sources nouvelles ou de ressources existantes. Les scientifiques peuvent travailler au plus près des origines des idées ou des phénomènes étudiés, ou bien à grande distance, que celle-ci soit temporelle, spatiale, théorique, méthodologique, linguistique ou autre. La même entité peut être représentée de nombreuses manières, par de nombreux moyens et à de nombreuses fins, sur de longues périodes. Dans le processus de représentation, ces entités peuvent acquérir beaucoup de noms différents, qui peuvent être agencés de nombreuses manières. Les métadonnées, c'est-à-dire les informations structurées utilisées pour gérer les données et d'autres ressources informationnelles, comprennent les noms, les étiquettes et les relations employés pour représenter les données. Les mécanismes de classification que sont les taxonomies, les thésaurus et les ontologies sont ensuite utilisés pour organiser ces métadonnées. Puisqu'une même entité peut être représentée de nombreuses manières pour mettre en évidence différents phénomènes, c'est le contexte qui détermine quand elle est une donnée et quand elle est une métadonnée.

Les métadonnées, la provenance et les mécanismes de classification facilitent l'échange de données au sein des communautés. Elles peuvent aussi être source de friction dans les collaborations et marquer les frontières entre communautés de recherche. Des facteurs externes peuvent déterminer si les données sont traitées en biens économiques, comment et pourquoi. Les mêmes données peuvent constituer des biens publics purs, des biens privés purs, des réservoirs communs de ressources ou des biens de club à différentes étapes ou même simultanément. Les droits de propriété et la déontologie associés aux données dépendent eux aussi largement du contexte. Reconnaître la diversité des données, des représentations et

des points de vue divergents des parties prenantes sur les questions de valeur, de droit et d'éthique est essentiel à la conception d'infrastructures de la connaissance efficaces. Les sources de tensions sont nombreuses, comme nous le développerons dans les études de cas à suivre.

II. Études de cas en travail scientifique des données

La partie II jette les bases des analyses de la partie III, « Politiques et pratiques en matière de données ». Bien que la littérature regorge d'études de l'activité de recherche au sein de plusieurs disciplines, notamment dans les sciences exactes, les différences de méthodes et de buts de chacune rendent leur comparaison délicate. Les pratiques de recherche des sciences humaines et sociales sont moins bien étudiées, en particulier en ce qui concerne leur utilisation des données. La partie II compare le travail scientifique des données dans les sciences exactes, les sciences sociales et les sciences humaines en fonction des facteurs identifiés au chapitre 4.

Chacun des trois chapitres à suivre comporte deux études de cas. La première étude de chaque chapitre montre un exemple de travail scientifique des données nécessitant de grands volumes d'informations, tandis que le second illustre un travail ayant recours à des types de données plus modestes et localisés. Environ la moitié de ces études de cas sont tirées de travaux empiriques que nous avons menés avec d'autres collaborateurs et collaboratrices : nous citons les publications rapportant nos résultats au fur et à mesure. L'autre moitié a été élaborée spécialement pour le présent ouvrage. Les études de cas du chapitre 5 comparent l'astronomie avec la science et la technologie des réseaux de capteurs. Au chapitre 6, les enquêtes et les médias sociaux s'opposent aux études sociotechniques de la science et de la technologie des réseaux de capteurs. Le chapitre 7, sur le travail scientifique des données dans les sciences humaines, développe des cas tirés de l'histoire de l'art et de l'archéologie antiques par opposition à une analyse approfondie de la philologie bouddhiste chinoise. Certaines de ces études de cas comportent plusieurs exemples. L'aide reçue pour étudier ces cas est expliquée plus longuement dans la section des remerciements à la fin du livre.

Par nature, les études de cas sont plus propices à l'approfondissement de l'analyse qu'à la généralisation. Présenter chaque exemple dans une structure parallèle rend possibles la comparaison et la formation de perspectives plus générales. Nous devons cependant émettre plusieurs mises en garde. Chaque cas représente l'activité d'un ou plusieurs laboratoires sous le prisme du travail scientifique de leurs données. Aucun de ces exemples ne saurait être considéré comme un archétype de la conduite de la recherche dans l'intégralité d'une discipline. Les six études de cas se veulent assez restreintes pour caractériser les pratiques d'un chercheur, d'une chercheuse ou d'un laboratoire, mais suffisamment larges pour évoquer l'ensemble des problèmes rencontrés par les scientifiques dans tout le domaine.

Chaque cas est présenté sous l'angle du travail scientifique des données, en adoptant la perspective théorique des sciences de l'information. Les scientifiques ne se décriraient vraisemblablement pas de cette manière ; il s'agit plutôt de réflexions sur leur travail. Nous utilisons ces cas pour expliciter les provocations présentées au chapitre 1, mettant ainsi en évidence les infrastructures de la connaissance, les tensions entre parties prenantes et l'éventail des motivations et des pratiques de collecte, de partage et de réutilisation des données.

5. Le travail scientifique des données dans les sciences exactes

Introduction

Les sciences exactes recouvrent la recherche sur le monde naturel, notamment l'étude de la faune et de la flore, des phénomènes biologiques et physiques et de leurs interactions. Pour les besoins du présent ouvrage, le terme générique « sciences exactes » englobe les disciplines que l'on désigne en Amérique par l'acronyme STEM (Science, Technology, Engineering, and Mathematics). Qu'elles soient lourdes ou légères, les sciences exactes emploient aujourd'hui des logiciels, des méthodes computationnelles et des statistiques pour leur recherche. La numérisation des recueils, des outils et des technologies y est la norme, de même que les objets matériels de toutes sortes : spécimens, lames, substances chimiques... Nos provocations se fondent sur la diversité des données et de la recherche entre les domaines et en leur sein. Dans certains d'entre eux, les infrastructures de la connaissance fonctionnent, du moins pour l'heure. Dans d'autres, des questions fondamentales sur la propriété des données, leur contrôle, leur partage, leur échange, leur pérennité, les risques et les responsabilités qu'elles suscitent sont âprement débattues. Enfin, un petit nombre de domaines commence tout juste à voir des tensions apparaître.

Les méthodes de recherche et les pratiques en matière de données

Les sciences exactes sont tout aussi cosmopolites que d'autres domaines d'érudition. Les références à la « méthode scientifique » ou aux « normes scientifiques » sont trompeuses, car il n'existe pas une, mais de nombreuses méthodes. La panoplie classique des normes scientifiques avancée par Robert Merton, il y a plusieurs générations de ça (Merton, 1963a, 1970, 1973 ; Storer, 1973), a laissé place à une vision bien plus nuancée du caractère émergent et varié des pratiques savantes. Des notions telles que les communautés de pratique (Lave et Wenger, 1991) et les cultures épistémiques (Knorr-Cetina, 1999) fournissent des cadres mieux aptes à cerner les pratiques en matière de données d'aujourd'hui.

La dichotomie entre science lourde (ou mégascience) et science légère proposée par Derek de Solla Price est une autre manière d'envisager la diversité des méthodes de recherche appliquées aux sciences exactes. Certaines disciplines, comme l'astronomie

ou la physique, œuvrent en vastes collaborations décentralisées ; celles-ci partagent souvent des infrastructures instrumentales colossales bâties sur des années. D'autres, comme l'écologie de terrain ou la biologie du comportement des oiseaux, travaillent en petits groupes et utilisent des technologies adaptées aux contraintes locales. Des aspects tels que la quête de connaissance théorique ou d'applications concrètes, comme nous l'avons décrit au chapitre 4, exercent également une influence profonde sur les pratiques en matière de données et les infrastructures de la connaissance nécessaires.

Les ensembles de données numériques représentent, dans certaines disciplines, des ressources communautaires essentielles, alors qu'elles sont inexistantes dans d'autres. Les méthodes élaborées dans un champ peuvent rester cantonnées à l'échelle locale ou être adaptées à d'autres domaines. Par exemple, BLAST (acronyme de Basic Local Alignment Search Tool, « outil basique local de recherche d'alignement ») est un outil développé pour apparier des séquences ADN (Altschul *et al.*, 1990). Algorithme permettant d'identifier rapidement des chaînes de caractères, BLAST a depuis été utilisé comme méthode d'identification pour du texte, du son et des images. Les tableurs, à l'origine créés pour des applications commerciales, sont largement utilisés dans les sciences exactes. Dans certains cas, c'est la principale méthode de gestion des données. Dans d'autres, ils constituent le plus petit dénominateur commun des solutions de partage de données entre des groupes utilisant des outils incompatibles. Les modalités, la temporalité et les raisons du transfert de méthodes et d'outils entre domaines sont encore mal connues.

Études de cas en sciences exactes

Parce qu'elle est l'un des plus anciens domaines de recherche et qu'elle est familière aux scientifiques comme au grand public, l'astronomie est un cas idéal pour entamer notre exploration des pratiques en matière de données dans différentes disciplines. Les connaissances astronomiques étaient indispensables à l'agriculture, à la navigation et à la religion ; ces applications se poursuivent aujourd'hui. Galilée fut le premier scientifique à se rendre compte de l'intérêt des télescopes pour l'observation du ciel et améliora ces technologies pour ses propres travaux. Ses notes sur le mouvement des satellites de Jupiter en 1610 sont des documents canoniques de la science moderne. La recherche en astronomie nous donne un aperçu du *big data*, des sciences physiques et des infrastructures de la connaissance. C'est une science du temps long, qui recourt largement aux technologies de l'information pour recueillir, analyser et visualiser des données.

La science et la technologie des réseaux de capteurs, la deuxième étude de cas de ce chapitre, présentent de nombreux contrastes avec l'astronomie. La télédétection par caméras satellites a révolutionné les sciences de l'environnement en fournissant de vastes vues de la Terre depuis le ciel. Les réseaux de capteurs intégrés, à l'inverse, fournissent des clichés microscopiques de l'environnement depuis le sol. Ils peuvent être déployés aux fins d'une collecte continue de données sur de longues périodes ou pour une brève étude de phénomènes spécifiques.

La recherche en réseaux de capteurs nous offre un aperçu de ce que sont la science légère, les sciences de la vie, la recherche technologique et les collaborations interdisciplinaires.

Ces deux études de cas, qui s'appuient sur des exemples tirés de projets dans les champs concernés, permettent de comparer la sélection, la création, l'utilisation, le partage, la réutilisation, la gestion et la conservation des données dans les sciences exactes, ainsi que l'évolution de la recherche dans un monde interconnecté. Deux cas ne sauraient néanmoins suffire à représenter toute la diversité du travail scientifique des données. Ceux-ci, et ceux des deux chapitres suivants, ont été choisis pour les contrastes qu'ils présentent, pour les données probantes dont ils disposent et pour leurs résonances avec les provocations.

Les études des cas des sciences exactes illustrent le caractère complexe et brouillon du processus de recherche, la difficulté à séparer les données d'observation des instruments utilisés pour les recueillir et des méthodes employées pour les analyser, ainsi que les nombreuses étapes de la manipulation des données pour qu'elles deviennent interprétables et puissent mettre en évidence des phénomènes. Chaque étude commence par un examen des sources de données et se termine par une explication de la conduite de la recherche grâce à ces données.

L'astronomie

L'histoire de l'astronomie est longue, riche et bien attestée. D'après l'*Oxford English Dictionary*, l'astronomie, dans son acception contemporaine, « étudie l'univers au-delà de l'atmosphère terrestre, y compris les objets célestes et les phénomènes extra-terrestres, ainsi que la nature et l'histoire de l'univers ». Ses frontières, comme pour tout domaine de recherche, sont poreuses. Les astronomes collaborent fréquemment avec des spécialistes de la physique, de l'ingénierie et de l'informatique. Certains comptent également des partenaires dans le domaine de la statistique, de la médecine, de la biologie, de l'art et du *design* et d'autres encore.

L'astronomie présente de nombreuses caractéristiques qui nous aideront à éclairer les grandes thématiques de ce livre. C'est un domaine qui relève du *big data* et dispose de décennies d'expérience en matière de données numériques et de calcul. Au cours de cette période, les astronomes ont été confrontés au bouleversement des pratiques en matière de données, passant d'un travail solitaire à de vastes collaborations et de la maîtrise locale des instruments à des ressources internationales partagées. L'astronomie est devenue une discipline tout aussi diverse que le reste des sciences, constituée de poches de *big data*, de *little data* et parfois, de *zéro data*. Certaines et certains astronomes sont des pionniers du partage et de la réutilisation de données ; d'autres sont connus pour conserver jalousement leurs ressources. Ces scientifiques ont bâti une infrastructure de la connaissance sophistiquée pour coordonner des ressources informationnelles à l'échelle mondiale ; pourtant, ils dépendent de l'expertise humaine pour identifier les relations entre objets. Des lacunes subsistent dans leur capacité à trouver et exploiter des ressources informationnelles.

Notre étude de cas se concentre sur l'astronomie en tant que science observationnelle ; les exemples présentés ici ne sont pas représentatifs de l'ensemble de la discipline. Le périmètre de l'étude est suffisamment large pour explorer les sources, les utilisations et le traitement final des données, mais suffisamment restreint pour proposer des explications cohérentes des choix de données dans un domaine et jeter les bases d'une analyse du partage, de la réutilisation et de la responsabilité dans les chapitres suivants. Le contenu est tiré de la littérature publiée et de recherches en cours sur les pratiques en matière de données dans la discipline. Notre exemple de conduite de recherche en astronomie résulte d'entretiens, d'observations et d'autres analyses du travail d'un laboratoire basé au Harvard-Smithsonian Center for Astrophysics (CfA) (Borgman, 2013 ; Borgman *et al.*, 2014 ; Sands *et al.*, 2014 ; Wynholds *et al.*, 2011, 2012).

La taille compte

L'envergure des collaborations et les volumes de données manipulés varient grandement en astronomie. Jusqu'à la seconde moitié du *xx*^e siècle, la discipline s'exerçait en solitaire. Galilée possédait ses propres télescopes et avait une maîtrise totale sur ses données. Les télescopes modernes, comme Hale de l'observatoire Palomar qui, avec ses cinq mètres, a longtemps été le plus grand du monde, appartenaient à des sociétés privées et seuls quelques astronomes triés sur le volet pouvaient y accéder. Entre l'inauguration de Hale en 1948 et celle des Gemini de huit mètres en 1999, le télescope est devenu un outil de recherche bien différent. En parallèle, l'astronomie s'est mise à attirer les fonds publics, ce qui a conduit à une évolution culturelle dans la discipline. Beaucoup plus d'astronomes disposaient d'un temps

de télescope, ainsi que d'un accès aux données (McCray, 2004). La communauté mondiale des astronomes professionnels en poste à l'université et dans des centres de recherche a au moins quadruplé depuis les années 1970. On estime aujourd'hui qu'elle compte entre dix mille et quinze mille chercheuses et chercheurs actifs (International Astronomical Union, 2013 ; DeVorkin et Routly, 1999 ; Forbes, 2008). Quant à la communauté des astronomes amateurs, elle est immense et vient renforcer les rangs des connaisseurs et connaisseuses de l'astronomie.

Science lourde, science légère

L'astronomie est une mégascience, ou science lourde (*big science*), pour reprendre les termes de Weinberg (1961) et Price (1963) : vaste, ambitieuse, pratiquée sur le long terme et nécessitant des investissements économiques et sociaux substantiels. Aujourd'hui, c'est une discipline internationale où prévaut une coordination de haut niveau pour la conception des instruments et des missions. La recherche est menée par des équipes grandes et petites. Aujourd'hui, il faut dix à vingt ans pour concevoir, construire et mettre en service un télescope ; une planification à très long terme est donc nécessaire. Les contraintes de temps sont en partie sociales et politiques – le travail de formation et de coordination des équipes et la recherche de financement sur plusieurs cycles –, en partie techniques – il faut concevoir les appareils et les logiciels en anticipant l'état de la technologie au moment de la mise en service – et en partie physiques – la fabrication des instruments, y compris le moulage, le refroidissement et le polissage du verre des miroirs des télescopes, prenant en effet plusieurs années – (University of Arizona Science Mirror Lab, 2013 ; McCray, 2014). La conception des instruments et des missions influence à son tour les choix opérés sur les données résultantes, la manière dont celles-ci sont recueillies et conservées et les conditions d'accès aux instruments et aux données pour les astronomes. Chaque fois qu'un nouveau télescope est lancé dans l'espace ou qu'un observatoire voit sa « première lumière », la conception de la prochaine génération d'instruments est déjà bien entamée.

Lorsque l'astronomie s'est mise à dépendre davantage des fonds publics et de la collaboration communautaire, il a fallu établir un consensus sur les priorités de recherche (McCray, 2004 ; Munns, 2012). Depuis les années 1960, la communauté astronomique américaine mène une « étude décennale » afin d'identifier les projets prioritaires. Les catégories scientifiques des projets changent chaque décennie, illustrant les évolutions de la discipline. Parmi les neuf pôles de 2010, on trouvait « Cosmologie et physique fondamentale », « Systèmes planétaires et formation d'étoiles », « Observations électromagnétiques depuis l'espace » et « Astronomie optique et infrarouge depuis le sol » (The National Academies, 2010). L'étude décennale consiste en une série de recommandations adressées aux organismes de financement nationaux et

internationaux des régions où la communauté demande un soutien. Elle ne garantit cependant pas un financement ou des affectations prioritaires. Le financement effectif fait en effet l'objet de négociations avec les organismes, les institutions mères telles que le Congrès états-unien et les partenaires internationaux.

Big Data, longue traîne

L'astronomie relève désormais du *big data* pour ce qui est des données d'observation dont elle dispose, situant ainsi les astronomes dans la tête de la courbe de la longue traîne. Le volume absolu des données astronomiques continue de s'accroître par ordre de grandeur à chaque nouvelle génération de télescopes. Les données astronomiques sont massives par leur volume et leur vitesse, pour reprendre les termes de Laney (2001) introduits au chapitre 2. La croissance pose continuellement problème à mesure que les nouveaux instruments recueillent de plus en plus d'informations à un rythme croissant. Pratiquement toutes les données astronomiques consistent en des mesures de l'intensité des radiations électromagnétiques (par exemple, les rayons X ou la lumière visible) en fonction de la position dans le ciel, de la longueur d'onde et de l'heure.

L'adaptation de l'astronomie aux technologies numériques, qui a pris plusieurs décennies, a provoqué des changements qualitatifs dans les formes de données recueillies. Pour des générations d'astronomes, l'enregistrement des données était un processus analogique, consistant en de longues expositions continues sur des plaques de verre, des dispositifs à transfert de charge (la base de la photographie numérique) ou des spectrographes. Les astronomes passaient des nuits entières au télescope, déplaçant précautionneusement l'instrument pour obtenir une exposition continue d'un objet ou d'une région céleste sur plusieurs minutes ou plusieurs heures. Chaque nuit passée sur la montagne n'engendrait que quelques traces. Aujourd'hui, en revanche, la collecte numérique produit des images distinctes et non plus continues. Les enregistrements numériques sont bien plus faciles à copier à l'identique et à transférer sur des supports de stockage. Ils sont plus maniables et plus simples à diffuser que les données analogiques. La transition de l'analogique au numérique a démarré dans les années 1960, s'est accélérée dans les années 1970 et était presque achevée à la fin des années 1990 (McCray, 2004, 2014).

À la fin du xx^e siècle, les télescopes produisaient des données bien plus vite que les êtres humains ne pouvaient les consommer. De nombreuses phases du processus sont automatisables. Du moment que la collecte de données est définie avec précision, des robots peuvent faire fonctionner les instruments durant la nuit et envoyer les jeux de données qui seront lus au matin. Néanmoins, les instruments se trouvent souvent sous le contrôle direct de l'astronome. Recueillir ses propres données peut

toujours impliquer de veiller toute la nuit, même si c'est dans des conditions plus confortables qu'autrefois. Pour la nouvelle génération, l'analyse et la visualisation des données par ordinateur est la norme. Cependant, beaucoup d'astronomes ont acquis leur savoir-faire au temps de l'analogique et apportent cette expertise en matière d'analyse au travail d'aujourd'hui. La collecte manuelle subsiste encore dans certains cas, comme pour l'observation des taches solaires (Curwen, 2013). Les données analogiques rassemblées au cours des siècles restent des traces précieuses, car permanentes, des cieux d'autrefois. Certaines ont été numérisées et mises à disposition dans des référentiels publics. D'autres relèvent du privé et demeurent sous le contrôle des astronomes et des institutions qui les ont collectées.

Les relevés astronomiques constituent l'aboutissement de projets de recherche visant à recueillir de grands volumes de données sur une région du ciel sur un long laps de temps. Le Sloan Digital Sky Survey (SDSS), qui porte le nom de son principal financeur, l'Alfred P. Sloan Foundation, a été le premier relevé astronomique destiné à un usage public immédiat. La collecte, démarrée en 2000, a cartographié environ un quart du ciel nocturne à l'aide d'un télescope optique situé à Apache Point au Mexique. Entre 2000 et 2008, avec une série de neuf divulgations de données, le SDSS a enregistré des données de plus en plus vite et à des résolutions de plus en plus précises grâce aux nouveaux instruments ajoutés au télescope, aux avancées dans les dispositifs à transfert de charge des appareils photographiques et aux améliorations de la vitesse et de la capacité des ordinateurs (Ahn *et al.*, 2012 ; Bell *et al.*, 2009 ; Finkbeiner, 2010 ; Gray *et al.*, 2005 ; Sloan Digital Sky Survey, 2013a ; Szalay, 2011).

Pan-STARRS, la nouvelle génération des relevés astronomiques, cartographie une région plus vaste du ciel à un plus grand niveau de détails et est en outre capable d'identifier des objets en mouvement. Le télescope, basé à Hawaï, est mis en service par étapes. Les appareils photographiques à gigapixels de Pan-STARRS sont les plus grands et les plus sensibles jamais construits. Le Large Synoptic Survey Telescope (LSST)¹, un télescope terrestre destiné à établir le relevé astronomique qui supplantera Pan-STARRS, prétend être « l'œil le plus rapide, le plus large et le plus profond de la nouvelle ère numérique ». Il obtiendra chaque nuit trente téraoctets de données. Quant au Square Kilometre Array (SKA), qui doit devenir le plus grand radiotélescope au monde, il devrait enregistrer quatorze exaoctets de données par jour. Les données issues de relevés astronomiques et de grandes missions de télescopes publics sont souvent conservées dans des archives et mises à la disposition du public après des périodes d'embargo et de traitement. La gestion des données représente l'un des principaux défis de missions telles que

1. NdT : Rebaptisé Observatoire Vera-C.-Rubin en 2020.

Pan-STARRS, LSST et SKA (Pan-STARRS, 2013a ; Large Synoptic Survey Telescope Corporation, 2010 ; Square Kilometre Array, 2013 ; Shankland, 2013).

Cependant, les astronomes ne sont pas nécessairement toutes et tous des spécialistes du *big data* et toutes les données astronomiques ne sont pas diffusées au public. Certains scientifiques passent encore des nuits en montagne à recueillir de petits volumes d'observations très spécialisées. D'autres obtiennent des données « parasites » des instruments secondaires embarqués sur les télescopes spatiaux. D'autres encore fabriquent leurs propres instruments afin d'enregistrer des types d'informations précises, dont ils ont besoin pour répondre à leurs questions de recherche. Les données de ces projets plus modestes peuvent être stockées indéfiniment sur les serveurs locaux des chercheurs et chercheuses. Elles peuvent représenter des volumes importants et s'avérer très spécialisées et difficiles à interpréter. Les astronomes de l'ère analogique avaient tendance à ne se fier qu'aux données qu'ils avaient collectées eux-mêmes : en effet, nul ne connaît mieux ses données que celui qui les a recueillies, qu'elles soient analogiques ou numériques. Des artefacts tels que des défaillances informatiques ou des changements météorologiques qui occultent une image sont difficiles à consigner dans des procédures automatisées, mais peuvent constituer une information indispensable à l'interprétation des résultats.

Quant aux théoriciennes et théoriciens, ils peuvent disposer de mégadonnées ou de zéro données, selon la définition qu'ils en ont. En astronomie, les théories analytiques reposent sur des équations solubles grâce au papier et au crayon ou, aujourd'hui, avec de superordinateurs. Elles servent à modéliser des phénomènes observés ou à établir des prévisions. Dans les théories computationnelles ou numériques, les phénomènes et objets astronomiques sont simulés afin de modéliser, prévoir et expliquer des phénomènes. Les simulations computationnelles se situent à mi-chemin de la théorie et de l'observation puisqu'elles mobilisent des principes de théorie analytique pure sur des informations qui se rapprochent fort de données d'observation. La modélisation peut synthétiser ce qu'un télescope donné verrait si une source, un événement ou une région présentant les propriétés simulées venait à être observé. La plupart des modèles reposent sur des observations réelles effectuées par d'autres astronomes et utilisées comme données d'entrées. Les concepteurs et conceptrices de modèles computationnels de phénomènes astronomiques considèrent parfois les résultats de ces modèles comme leurs données. D'autres diront qu'ils n'utilisent pas de données.

Ordinairement, les simulations produisent une série temporelle qui montre l'évolution d'un phénomène donné. Les modèles sont alimentés par de multiples combinaisons de paramètres afin de simuler une gamme de conditions réalistes. Chaque combinaison de paramètres peut être exécutée à de nombreux pas de temps (appelés

« clichés »). Chaque ronde de calcul (et même chaque cliché) peut donc produire plusieurs téraoctets de données de sortie, bien plus qu'on ne pourrait conserver indéfiniment. Les modélisateurs et modélisatrices estiment parfois que leurs données consistent dans les quelques kilooctets d'observations empiriques dont ils ont besoin pour exécuter une ronde de calculs. Certains établissent des distinctions subtiles entre les simulations, le code associé aux résultats des simulations, les paramètres et les données. On peut créer des données synthétiques au même format que des données d'observation, ce qui permet des comparaisons statistiques avec le même outillage analytique.

Quand est-ce une donnée ?

Les données astronomiques sont plus difficiles à caractériser que ce qu'un amateur pourrait imaginer. En effet, de nombreuses personnes aux talents variés sont impliquées dans la conception, l'élaboration et la mise en œuvre des missions astronomiques. Des astronomes participent à ces étapes préliminaires, mais il arrive que la recherche observationnelle réalisée avec des télescopes publics dépende de données issues d'instruments conçus plusieurs dizaines d'années auparavant. Les astronomes peuvent consacrer une grande partie de leurs carrières à une collaboration à long terme, passer de projet en projet, tirer des données de plusieurs missions ou encore se concentrer sur des sujets spécialisés à l'aide de leurs propres instruments. Certaines rédigent des demandes de temps d'observation afin de recueillir leurs propres données ; certains ont recours aux données conservées dans les archives ; d'autres fabriquent leurs propres instruments ; et d'autres encore puisent dans toutes ces sources et ressources. Le moment où une entité considérée devient une donnée astronomique utile dépend de choix comme ceux-là.

Les personnes impliquées dans une étape ne connaissent souvent des étapes précédentes que le strict nécessaire pour faire leur travail. Au moment où un instrument voit sa première lumière, nul n'a une vision complète de toutes les décisions qui ont abouti aux observations réalisées. Pour le physicien ou la physicienne qui a mené la conception de l'instrument, le voltage des dispositifs à transfert de charge peut constituer des données. Pour le théoricien ou la théoricienne étudiant les origines de l'univers, il peut s'agir du résultat de simulations modélisant la formation, l'évolution et la mort d'étoiles, de galaxies et d'autres objets célestes. Pour l'astronome empirique, les données peuvent être des « cubes d'images », constitués de coordonnées célestes et spectrales. Pour l'ingénieur ou ingénieure logiciel, cela peut aussi bien être les données de sortie des dispositifs à transfert de charge que les fichiers nettoyés, étalonnés et structurés déposés dans les référentiels.

Pour citer un des astronomes interrogés, « bâtir un article d'astronomie est comme construire une maison ». Le point de départ est rarement clair : les habitants rénovent et agrandissent une maison sur des années, voire des décennies. Même quand une famille part d'un terrain vague, il peut y avoir eu une structure à cet endroit par le passé. Si on remonte encore le temps, quelqu'un a décidé de comment le terrain serait loti, ce qui a déterminé les possibilités pour la taille et l'orientation de la maison, et ainsi de suite.

Sources et ressources

Le relevé COMPLETE (COordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions [COMPLETE], 2011) considéré dans l'étude de cas ne peut être compris que si l'on appréhende plus largement les méthodes de collecte, de sélection, d'utilisation et de gestion des données en astronomie d'observation. Les sources et les ressources y sont nombreuses et diverses. Les données sont difficiles à extraire des instruments, de l'expertise disciplinaire, des principes scientifiques et des formes de représentation spécifiques au domaine. Elles consistent en des « observations du ciel », mais ces dernières reposent sur des instruments dotés de capacités de détection spécialisées. Les signaux captés par ces instruments sont nettoyés et étalonnés selon les normes disciplinaires. On a recours à des méta-données pour représenter les signaux de manière à pouvoir les mettre en correspondance avec des données d'autres instruments. Les structures des standards de données permettent aux astronomes d'appliquer des outils communs pour l'analyse, la visualisation et la communication. À ces technologies, ces outils et ces normes se superposent des institutions et des pratiques professionnelles visant à relier les données et les publications de la communauté internationale des astronomes.

Télescopes

Les télescopes constituent la technologie de base de l'astronomie tout en étant la plus complexe d'entre elles. Les technologies optiques ont fait de grands progrès depuis Galilée et enregistrent aujourd'hui des images numériques. Les télescopes modernes comportent de multiples instruments, chacun doté de capacités différentes. Ils peuvent être remplacés au fil du temps, ce qui allonge la durée de vie du télescope.

Il existe plusieurs types de télescopes, qui peuvent être placés au sol, dans les couches hautes de l'atmosphère, dans des avions et des ballons, ou encore dans l'espace. Les télescopes terrestres sont généralement situés en altitude ou dans des lieux isolés, loin des lumières des villes. Les télescopes optiques sont dotés de miroirs ou de lentilles pour concentrer la lumière, comme ceux de Palomar en Californie ou de La Silla dans le désert d'Atacama au Chili. Les radiotélescopes, quant à eux, utilisent

des antennes paraboliques pour concentrer des signaux. Le Square Kilometre Array (littéralement « réseau d'un kilomètre carré »), un ensemble de paraboles qui seront installées en Australie et en Afrique du Sud, est ainsi nommé parce que sa zone de collecte effective sera d'un kilomètre carré. Ce projet international est mis en œuvre dans l'hémisphère sud afin d'avoir la meilleure vue de la Voie lactée ; en outre, la région est moins sujette aux interférences radio que d'autres sites pressentis.

Les télescopes en orbite autour de la Terre peuvent voir bien plus loin dans l'espace. La planification de ces instruments prend des décennies et ils peuvent générer des données pour plusieurs dizaines d'années. Actuellement en orbite à 568 kilomètres de la Terre, Hubble est l'un des télescopes spatiaux les plus connus. Conçu en 1946 par Lyam Spitzer Jr, il a été lancé en 1990. Entre ces deux dates, une série de projets de conception et de développement issus de collaborations et de financements internationaux ont été menés. Le Hubble Space Telescope (HST) fournit des données scientifiques depuis l'espace depuis plus de vingt ans. Avec ses cinq ensembles d'instruments, chacun capable de capter la lumière de différentes façons et à différentes longueurs d'onde, le HST est un laboratoire astronomique en orbite. D'autres appareils alimentent le satellite, mesurent sa santé et permettent d'y apporter des ajustements. Alors que la plupart des télescopes spatiaux sont mis en orbite ou lancés dans l'espace lointain sans intervention physique ultérieure, Hubble a reçu plusieurs visites pour ajouter, réparer et remplacer des instruments. Les données qu'il a collectées ont donné lieu à plus de dix mille publications (HubbleSite, 2013b).

La collecte proprement dite de données d'instruments de télescope est souvent un processus industriel géré par la mission scientifique du projet, comme Hubble ou Chandra. Le HST, par exemple, envoie ses signaux à un satellite de poursuite et de relais, lequel les transmet à une station terrestre dans le Nouveau-Mexique, laquelle les renvoie au Goddard Space Flight Center à Greenbelt, dans le Maryland, qui les transmet à son tour au Space Telescope Science Institute (STScI) non loin de là, à Baltimore. Des validations et des vérifications sont effectuées à chaque étape de la procédure. Les astronomes utilisent les données d'observation du STScI, une fois celles-ci étalonnées et divulguées. Elles et ils peuvent aussi déposer des demandes de temps d'observation pour utiliser des instruments particuliers, pendant un laps de temps donné, afin d'étudier leurs propres questions de recherche. Environ deux cents demandes sont accordées chaque année, ce qui représente un taux d'acceptation de 20 % (HubbleSite, 2013b).

Spectre électromagnétique

Tout instrument, qu'il se trouve dans l'espace ou au sol, est conçu pour capter des signaux sur une ou plusieurs gammes de longueurs d'onde. Les radiotélescopes

détectent des signaux de basse fréquence, dits aussi à basse énergie ou de grande longueur d'onde. Les télescopes à rayons gamma captent quant à eux des signaux de haute fréquence, dits aussi à haute énergie ou de courte longueur d'onde. Les rayons gamma et X sont largement bloqués par l'atmosphère terrestre ; c'est pourquoi les télescopes à rayons gamma et X sont installés au-delà, sur des fusées ou des satellites. Le spectre électromagnétique est continu. En astronomie, il est couramment divisé, par ordre croissant d'énergie, en ondes radio, micro-ondes, infrarouges, optiques, ultraviolets, rayons X et rayons gamma. On peut également opérer des distinctions plus fines : infrarouge lointain, infrarouge moyen, infrarouge proche, rayon X mou, rayon X dur, etc. Les longueurs d'onde sont parfois désignées selon leur taille en distance métrique, en millimètres ou encore en unités moindres. La lumière visible (les couleurs perçues par l'œil humain) dont Galilée disposait ne représente qu'une bande très étroite sur ce spectre.

Objets célestes

Les objets célestes, également appelés objets astronomiques ou astres, sont des entités observables dans l'univers d'origine naturelle. Ils comprennent les étoiles, les planètes, les galaxies, les comètes, ainsi que d'autres objets moins familiers tels que les nébuleuses, les restes de supernovae et les trous noirs. La majeure partie de la recherche astronomique implique l'étude d'objets célestes, que ce soit individuellement ou en combinaison avec d'autres phénomènes. Les astres apparaissent dans des régions particulières du ciel et sont visibles à certaines longueurs d'onde. Pour étudier un objet céleste donné, l'astronome a besoin d'observations de cette région, obtenues avec des instruments capables de saisir les phénomènes étudiés. Dans certains cas, elle ou il sollicitera un temps d'observation sur un instrument de télescope spécifique sur plusieurs jours à la bonne saison, afin que le télescope soit pointé sur la région souhaitée. Comme les étoiles, les planètes et les astéroïdes se déplacent dans le ciel, les astronomes ont besoin de modèles 3D de ces trajectoires afin de pouvoir en dériver les coordonnées célestes (projetées en 2D) de ces objets à tout moment.

Produits de l'astronomie

Les observations issues d'instruments de télescope passent par de nombreuses étapes de nettoyage, d'étalonnage et de réduction avant d'être mises à disposition dans des référentiels de données astronomiques. Les étapes de ce processus sont appelées *pipelines* (European Southern Observatory, 2013 ; National Optical Astronomy Observatory, 2013b). Le traitement en *pipeline* peut prendre plusieurs mois ; ainsi, les observations sont souvent mises à disposition sous forme d'un *data release* (divulgence de données) accompagné d'un *data paper* (publication de données) qui les documente. La neuvième divulgation de données du SDSS, par exemple, comprend

toutes les données issues d'un nouvel instrument ajouté au télescope après la huitième divulgation, auxquelles s'ajoute l'astrométrie (les positions célestes) corrigée de la huitième divulgation (Ahn *et al.*, 2012 ; Sloan Digital Sky Survey, 2013b). De mêmes observations peuvent donc être divulguées plusieurs fois à mesure que des améliorations sont apportées au *pipeline*. Les astronomes veillent à identifier le *data release* utilisé dans une étude, car cela influe sur l'interprétation des résultats.

Lorsque les astronomes recueillent leurs données grâce à des demandes de temps d'observation, elles et ils peuvent réaliser eux-mêmes leur traitement en *pipeline*. Ils disposeront parfois d'un manuel recommandant des outils et des étapes à effectuer. Si leurs instruments sont bien étalonnés, ils peuvent comparer leurs données à celles de produits issus du même appareil. Utiliser son propre instrument ou un nouveau procure un avantage en matière de découverte, mais est aussi limitant, puisqu'on ne dispose pas de produits comparables pour valider ses résultats. Les astronomes concernés peuvent alors se tourner vers d'autres ressources de données pour étalonner leurs instruments, sans toujours le mentionner dans leurs publications (Wynholds *et al.*, 2012). Certains préfèrent des observations peu traitées, par exemple parce qu'ils cherchent des phénomènes nouveaux ou d'autres récurrences que les méthodes de réduction standard peuvent dissimuler.

Les observations astronomiques engendrent de nombreux autres types de produits de données, comme les éphémérides et les relevés de régions ou d'objets célestes. Les éphémérides, dont l'existence remonte aux débuts de l'observation du ciel, consistaient à retracer les positions et la brillance des astres tout au long d'une année. Ils fournissaient alors des informations essentielles pour la navigation et l'orientation. Les éphémérides contemporaines puisent les descriptions précises des connaissances sur chaque étoile dans des référentiels de données (National Aeronautics and Space Administration, Goddard Space Flight Center, 2014). Des étoiles connues peuvent être référencées par un numéro de catalogue. Les catalogues et les autres produits peuvent ainsi être interrogés pour déterminer si un objet est connu ou non. Les relevés astronomiques en temps réel utilisent ces produits pour proposer une identification quasi instantanée de nouveaux objets célestes. Les méthodes de détection de phénomènes éphémères peuvent envoyer des alertes sur des télescopes, des smartphones et d'autres appareils (Caltech, 2013b). La divulgation des données peut être immédiate ou être reportée de plusieurs années, comme nous le verrons au chapitre 8.

Les infrastructures de la connaissance

L'astronomie possède l'infrastructure de la connaissance la plus complète de tous les champs évoqués dans les études de cas. Les accords sur des standards en matière

de structures de données, de métadonnées et d'ontologies, auxquels s'ajoutent la coordination à l'échelle internationale et d'importants investissements de la communauté dans des référentiels, des outils et des ressources humaines, ont produit un réseau complexe de ressources informationnelles. Bien que la collecte de données soit largement automatisée, des parts substantielles de l'infrastructure nécessitent une expertise humaine pour l'attribution de métadonnées et l'identification de liens entre des objets voisins.

Les tâches manuelles qu'il faut accomplir pour interpréter des observations et documenter des relations entre objets numériques sont un exemple du « travail invisible » si souvent nécessaire au fonctionnement des infrastructures. Bien sûr, il est visible pour celles et ceux qui l'accomplissent, mais ceux qui dépendent de l'infrastructure ignorent généralement ces investissements, du moins tant que le système ne tombe pas en panne.

Métadonnées

Les observations astronomiques sont souvent acquises sous forme de spectres (intensité en fonction de la longueur d'onde), d'images (la distribution de l'intensité du ciel à une longueur d'onde donnée) ou de cubes (un jeu de données 3D donnant l'intensité en fonction d'une position et d'une longueur d'onde, dont on peut extraire des images et des spectres). Dans certains cas, un instrument appelé « bolomètre » sert à mesurer l'intensité à une position donnée sur une gamme très restreinte de longueurs d'onde. Les observations astronomiques sont de plus en plus souvent obtenues sous forme de série temporelle, c'est-à-dire d'une série d'échantillons au fil du temps d'un des types de données listés plus haut. Les instruments de télescope peuvent générer automatiquement les coordonnées célestes, la longueur d'onde et l'heure de l'observation sous forme de métadonnées. Les noms des objets sont entrés manuellement au moment de l'observation, car le jugement d'un être humain est nécessaire. D'autres informations utiles à l'interprétation des résultats, comme les conditions météorologiques et les erreurs instrumentales, peuvent être consignées dans des journaux d'observation.

Pour présenter des images prises à des longueurs d'onde situées bien au-delà du spectre visible, les astronomes attribuent des couleurs aux différentes parties du spectre, par exemple rouge pour les ondes radio, vert pour la fenêtre optique et bleu pour les rayons X. Cependant, il n'existe guère de standards de métadonnées pour ces fausses couleurs dans les illustrations. Les techniques pour les attribuer sont très variables et aboutissent à des images largement irreproductibles. Bien que les « jolies » illustrations composites aient les faveurs du grand public, beaucoup d'astronomes se montrent réticents à les faire figurer

dans leurs publications. La vision d'artiste doit être soigneusement corrigée par la validité scientifique (Kessler, 2012).

Les informations essentielles concernant l'instrument, les conditions d'observation, la longueur d'onde, la date, l'heure et les coordonnées célestes sont représentées dans un format de données standard appelé Flexible Image Transport System (FITS). Le FITS a été inventé dans les années 1970 et largement adopté dans les années 1980 dans le contexte de la transition numérique de l'astronomie. Les observations analogiques pouvaient ainsi être étalonnées sur les positions et les conditions de chaque télescope. La prise de vue numérique a permis de combiner des observations issues de plusieurs instruments, mais pour cela, il a fallu se mettre d'accord sur les structures de données et les systèmes de coordonnées (National Aeronautics and Space Administration, Goddard Space Flight Center, 2013a ; Hanisch *et al.*, 2001 ; McCray, 2014 ; Wells *et al.*, 1981).

La plupart des référentiels de données astronomiques fournissent aujourd'hui des ressources au format FITS ; les astronomes peuvent ainsi utiliser les métadonnées qu'elles contiennent pour localiser des données d'observation grâce à leurs coordonnées célestes, au spectre, à l'heure de l'observation et aux autres caractéristiques associées aux instruments. Un petit nombre d'astronomes préfèrent les formes de données « pré-FITS », plus brutes, situées à des niveaux de traitement plus faibles, comme nous l'avons vu au chapitre 2 (NASA's Earth Observing System Data and Information System, 2013).

Systèmes de coordonnées

L'astronomie est fondée sur le principe simple qu'il n'existe qu'un seul ciel. Il n'en a pas moins fallu des siècles d'innovation astronomique et technique pour établir un système de coordonnées qui puisse rendre compte des positions des objets célestes. La latitude (dimension nord-sud) pouvait se calculer sur terre ou en mer grâce aux étoiles. La longitude (dimension est-ouest) nécessitait des calculs temporels exacts, car la Terre se déplace sur son orbite. L'invention au XVIII^e siècle d'une horloge de précision capable de fonctionner sur un navire a transformé à la fois la navigation et l'astronomie (Sobel, 2007). Les systèmes de coordonnées astronomiques reposent sur des mesures de temps précises, car le mouvement de la Terre sur son orbite modifie subtilement la position et la longueur d'onde de l'émission d'un objet au fil du temps.

Les astronomes ont convenu, dans le cadre des normes FITS, d'un système de cartographie standard, appelé le World Coordinate System (WCS), afin de décrire les observations célestes (National Aeronautics and Space Administration, Goddard Space Flight Center, 2013b). À chaque pixel d'une image du ciel sont attribuées des

coordonnées de localisation X et Y. Celles-ci, qui correspondent généralement à l'ascension droite et à la déclinaison, sont l'équivalent de la longitude et de la latitude pour les positions terrestres. Quant au spectre électromagnétique, il constitue la troisième dimension utilisée dans les observations astronomiques. Celle-ci est souvent exprimée comme fréquence ou longueur d'onde et peut dans de nombreux cas être convertie en vitesse grâce à l'effet Doppler. Pour les objets situés hors de la Voie lactée, on incorpore la loi de Hubble au calcul de la distance d'un objet.

Des observations effectuées par différents instruments à différents moments peuvent être mises en correspondance grâce à ces systèmes de coordonnées. Des images du ciel saisies sur des plaques de verre il y a plus d'un siècle peuvent être appariées avec des clichés pris par des télescopes d'aujourd'hui (Harvard-Smithsonian Astrophysical Observatory, 2013b ; Johnson, 2007). De même, les photographies du ciel prises de nos jours peuvent être associées à leur position dans le ciel grâce au WCS, aux éphémérides et à d'autres produits de données astronomiques. La procédure est cependant imparfaite, parce que la mise en correspondance exige parfois de savoir d'où l'image a été prise (Hogg et Lang, 2008 ; Lang *et al.*, 2009).

Objets célestes

Les objets célestes et les autres phénomènes astronomiques disposent de leurs propres ensembles de métadonnées. Ils sont catalogués à la main après la publication d'articles grâce à une action internationale coordonnée. Les astres de notre galaxie sont recensés dans SIMBAD (Set of Identifications, Measurements and Bibliography for Astronomical Data, « ensembles des identifications, mesures et bibliographies pour les données astronomiques »), qui est basé au Centre de données astronomiques de Strasbourg (CDS). Les catalogues lisent les nouvelles publications en matière d'astronomie au fur et à mesure de leur parution et créent des fiches de métadonnées pour chaque objet céleste identifiable mentionné (Genova, 2013).

SIMBAD grandit rapidement au rythme des publications et des découvertes, comme le montrent les mises à jour quotidiennes de ses statistiques. Au moment d'écrire ces lignes, la base contient environ 18,2 millions d'identifiants pour 7,3 millions d'objets uniques mentionnés dans 285 000 publications, pour un total de 10 millions de références à des objets uniques (Centre national de la recherche scientifique, 2012 ; SIMBAD Astronomical Database, 2013 ; Genova, 2013). Une autre manière de présenter ces chiffres serait de dire que chacun des 7,3 millions d'astres est connu, en moyenne, sous 2,5 noms différents (les 18,2 millions d'identifiants). Chaque publication mentionne en moyenne 35 objets célestes (les 10 millions de références à des objets dans 285 000 documents). Néanmoins, ces objets ne sont pas également répartis dans la littérature astronomique. La plupart des publications décrivent quelques

astres, mais un petit nombre de documents listent un grand nombre d'objets. De même, la plupart des astres n'ont qu'un seul nom (par exemple Jupiter) alors que d'autres en ont beaucoup, parce que cela fait plusieurs siècles qu'ils sont identifiés dans des relevés et catalogues. Chaque publication est richement « taguée » de données, leur ajoutant une valeur permettant de trouver, combiner et distinguer des données sur des objets célestes.

Les astres situés hors de notre galaxie sont catalogués dans la NASA Extragalactic Database (National Aeronautics and Space Administration, Infrared Processing and Analysis Center, 2014a). Les données relatives au système solaire et aux planètes sont recensées dans un autre service (National Aeronautics and Space Administration, Jet Propulsion Laboratory, 2014). Le CDS est le pôle de coordination de nombre de référentiels de métadonnées astronomiques et fournit des outils de recherche et mise en correspondance tels qu'Aladin et VizieR (Centre national de la recherche scientifique, 2012, 2013).

Archivage des données

Des quantités massives d'observations astronomiques se trouvent dans des archives de données, que l'on appelle également référentiels, bases de données ou systèmes d'information. Si ces archives sont riches, elles ne sont pas exhaustives pour autant. Les observations issues de missions financées sur fonds publics, en particulier celles recueillies par des instruments de télescope lancés dans l'espace, sont généralement mises à la disposition du public. La plupart des référentiels sont organisés par missions ; c'est le cas pour les observations issues des missions Spitzer Space Telescope, Chandra et Hubble (Harvard-Smithsonian Astrophysical Observatory, 2013a ; NASA Spitzer Space Telescope, 2013 ; HubbleSite, 2013a).

Les données sont aussi organisées par longueur d'onde : c'est par exemple le cas dans les archives hébergées par l'Infrared Processing and Analysis Center (IPAC). L'IPAC classe les données par missions, ainsi que par types d'objets célestes, comme avec la NASA Exoplanet Archive (National Aeronautics and Space Administration, Infrared Processing and Analysis Center, 2014b). Chaque relevé astronomique majeur, comme le SDSS, Pan-STARRS et le LSST, propose son propre référentiel. Des données plus anciennes et bien conservées, comme celles du Two Micron All Sky Survey (2MASS), dont la collecte s'est achevée en 2001, gardent leur valeur indéfiniment (National Aeronautics and Space Administration, Infrared Processing and Analysis Center, 2014c).

Bien que les référentiels de données représentent des ressources précieuses, chaque archive est indépendante et possède sa propre interface utilisateur, ses

propres fonctionnalités et son propre modèle de données sous-jacent. Parce que les données astronomiques archivées tendent à être classées selon leur longueur d'onde d'observation et l'observatoire qui les a recueillies, il est nécessaire d'adopter une méthodologie cohérente (Accomazzi et Dave, 2011). Certains référentiels, comme le MAST, conservent des données issues de missions et de spectres multiples et acceptent également des contributions en données et modèles (National Aeronautics and Space Administration, Mikulski Archive for Space Telescopes, 2013). Le Data Discovery Tool, SkyView, et le WorldWide Telescope sont quelques-uns des outils – de plus en plus nombreux – permettant d'interroger de multiples archives et d'intégrer des sources de données (Goodman et Wong, 2009 ; International Virtual Observatory Alliance, 2013b).

Les astronomes disposent de plus en plus d'options pour partager des données qu'elles et ils ont recueillies eux-mêmes ou qui sont dérivées de sources publiques : archives acceptant les contributions, référentiels universitaires, sites web personnels ou de projet et échanges interpersonnels. Une étude sur un petit nombre d'astronomes indique que leurs pratiques en matière de partage ressemblent à celles d'autres disciplines. La forme la plus courante de partage de données consiste à les envoyer par e-mail aux collègues qui les demandent. Seule une faible proportion d'entre eux (environ 20 sur 175) avaient déposé des données dans une archive institutionnelle. La gestion des données reste largement le fait des personnes qui les ont collectées ou analysées. L'un des astronomes interrogés a comparé les pratiques en matière de données de son laboratoire à celles du SDSS : quand celui-ci diffuse un Data Release 1.0, 2.0 et ainsi de suite, « nous, on a plutôt : Doctorant 1, 2 et 3 » (Pepe *et al.*, à paraître).

Publications

Le contrôle bibliographique est bien plus complet en astronomie que dans la plupart des autres disciplines. Malgré son nom, l'Astrophysics Data System (ADS, « système de données de l'astrophysique ») du Harvard-Smithsonian Astrophysical Observatory–NASA consiste essentiellement en un système bibliographique. L'ADS, qui est opérationnel depuis 1993, contient des traces de publications d'astronomie fondamentale qui remontent au ^{xix}^e siècle, ainsi qu'un catalogue considérable de littérature grise de la discipline (Kurtz *et al.*, 2000, 2005 ; Harvard-Smithsonian Astrophysical Observatory, 2013c). L'ADS joue un rôle central dans l'infrastructure de la connaissance astronomique en conservant non seulement des documents bibliographiques, mais aussi les liens entre publications, traces d'objets célestes et archives de données (Accomazzi et Dave, 2011 ; Accomazzi, 2010 ; Borgman, 2013 ; Kurtz *et al.*, 2005).

Provenance

Les astronomes utilisent les riches composantes de leur infrastructure de la connaissance pour déterminer la provenance des données. Les personnels de recherche doivent pouvoir s'y fier, sachant que beaucoup de personnes, d'instruments et de logiciels différents ont touché le train binaire de l'observation. Les bits sont étalonnés, nettoyés, transformés et réduits à de nombreux stades du traitement en *pipeline*. La provenance pose plusieurs questions en fonction de la problématique de recherche et des circonstances. Une donnée obtenue dans un but particulier, dans une région céleste particulière, à certains moments et certaines longueurs d'onde, peut s'avérer utile pour d'autres fins, ou non. Par exemple, il est souvent bien plus difficile de déterminer la provenance de vieilles données saisies par un ou une astronome que de données issues de relevés astronomiques. Des données anciennes numérisées peuvent ne pas être correctement documentées pour certains types d'usages.

La provenance des relevés astronomiques et d'autres grands jeux de données est documentée dans les *data papers* (publications de données) que nous évoquions plus haut, par exemple ceux des *data releases* (divulgations de données) du SDSS (Ahn *et al.*, 2012) et de COMPLETE (Ridge *et al.*, 2006). Dans certains cas, on publiera des *instruments papers* (publications d'instruments) afin de reconnaître la contribution de celles et ceux qui les ont conçus et pour documenter plus en détail les décisions prises. Ces publications décrivent l'instrumentation, l'étalonnage et les choix de traitement. Les *data papers* figurent parmi les articles les plus cités en astronomie parce qu'ils rassemblent les références aux sources des données dans une même publication.

En astronomie, la provenance est également tracée par des outils et des services d'analyse courants. L'International Virtual Observatory Alliance (IVOA) est un organisme de coordination qui vise à élaborer des données et des outils et à les partager (Hanisch et Quinn, 2002 ; International Virtual Observatory Alliance, 2013a). Ses membres se réunissent régulièrement pour résoudre des problèmes d'interopérabilité et pour coordonner l'action des pays en faveur des infrastructures de recherche astronomique.

Bien qu'elle n'en ait pas terminé, l'astronomie a réalisé bien plus de progrès que la plupart des autres disciplines en matière de relations entre publications et données. SIMBAD fournit des liens entre les objets célestes et les articles qui relatent la recherche les concernant, lesquels sont hébergés dans l'ADS. En revanche, les métadonnées qui lient les observations aux publications et aux astres sont moins bien conservées. Des initiatives sont en cours afin de coordonner des systèmes et

des activités multiples, avec en ligne de mire une amélioration de l'interconnexion sémantique de ces ressources complémentaires (Accomazzi et Dave, 2011). Cette coordination concerne l'IVOA, l'ADS, le CDS, les bibliothèques astronomiques, les archives de données et des systèmes tels que WorldWide Telescope, qui peuvent intégrer des sources disparates (Worldwide Telescope, 2012). Nous discutons du rôle joué par ces efforts de coordination dans le partage des données au chapitre 8.

Les influences extérieures

L'astronomie ne subit pas moins que d'autres sciences l'influence de questions externes, telles que l'économie et la valeur, les droits de propriété et l'éthique. La création et l'utilisation de données en astronomie reposent sur des accords internationaux et sur une série de modèles de gouvernance. Ceux-ci sous-tendent les infrastructures de la connaissance astronomique de manière à la fois subtile et profonde.

Économie et valeur

Les données astronomiques intéressent la recherche informatique pour plusieurs raisons, la première étant qu'elles sont dépourvues de valeur financière apparente. La deuxième est que leur volume considérable et leur structure cohérente les rendent utiles à la recherche sur les bases de données. Troisièmement, il n'y a pas de sujets humains impliqués, donc très peu de contraintes déontologiques pour la réutilisation (Gray et Szalay, 2002 ; Hey, 2009b ; Lynch, 2009 ; Szalay, 2008).

Bien qu'il soit exact qu'aucun marché d'achat-vente n'existe pour les observations astronomiques ou pour les données de sorties de modèles numériques, de télescopes, d'instruments et de grandes archives de données telles que le SDSS, Hubble et Chandra, il faut envisager la plupart d'entre elles comme des réservoirs communs de ressources. Quand il y a des investissements d'organismes privés et publics dans l'infrastructure, il s'ensuit la mise en place de modèles de gouvernance afin d'assurer la qualité des ressources et un accès équitable. Le manque de pérennité et les « passagers clandestins » représentent une menace permanente pour ces ressources communes. Cependant, beaucoup d'instruments astronomiques et de ressources en données n'appartiennent pas à des réservoirs communs. Ainsi, des données sous le contrôle d'astronomes et de laboratoires peuvent être considérées, selon les circonstances, comme des informations brutes ou comme des biens privés. Certains instruments et certaines archives constituent des biens de clubs qui ne sont accessibles qu'à leurs seuls membres. Les logiciels nécessaires à l'analyse de données astronomiques et à l'interprétation de fichiers peuvent être en *open source* ou commerciaux.

Droits de propriété

Les droits de propriété des données varient selon les projets, les organismes de financement et d'autres facteurs. Que la recherche soit financée par des fonds publics ou privés, les chercheurs et chercheuses disposent généralement pour un temps de droits exclusifs sur les données. Les périodes d'utilisation exclusive, qu'on appelle également périodes d'embargo, courent de trois à dix-huit mois environ à compter de la date d'enregistrement des observations par le télescope (National Optical Astronomy Observatory, 2003, 2013a) ou à compter du moment où des « données utilisables scientifiquement » émergent du *pipeline* de traitement (Spitzer Science Center, 2013). Les chercheurs peuvent choisir de diffuser les données plus tôt ; en revanche, il faut obtenir une autorisation spéciale pour rallonger la période d'utilisation exclusive. Selon que celle-ci démarre au moment de l'observation ou lorsque les données deviennent utilisables, la différence de durée du contrôle sur les données peut se compter en mois, voire en années, pendant lesquels les chercheurs rédigeront leurs publications. Les données obtenues de télescopes privés peuvent n'être jamais divulguées, cependant des données anciennes issues de télescopes importants ont déjà été rendues publiques.

Pour les missions dont les données sont vouées à l'archivage – c'est-à-dire la majorité de celles réalisées sur les grands télescopes spatiaux –, elles sont mises à disposition dans des référentiels une fois le traitement en *pipeline* achevé et les périodes d'utilisation exclusive terminées. Lorsque les astronomes souhaitent effectuer leurs propres observations dans le cadre de ces missions, leurs droits sur les données viennent s'ajouter à la gouvernance des instruments. Les télescopes appartiennent, selon diverses modalités, aux universités, aux consortiums, aux États et aux autres organisations qui les financent. L'élaboration des politiques qui régissent l'usage des instruments, comme le Hubble Space Telescope, peut être déléguée à une mission scientifique. La majeure partie des télescopes publics sont accessibles aux astronomes qualifiés, qui peuvent soumettre une demande pour les utiliser. D'autres instruments sont réservés aux relevés synoptiques, lesquels rassemblent de vastes ensembles de données qui seront mis à la disposition de tous. Un nombre de plus en plus faible de télescopes ne sont accessibles qu'à une élite liée aux institutions privées qui les possèdent.

Les politiques en matière de droits sur les données de ces organes de gouvernance établissent des distinctions fines dans le droit d'utiliser ou de contrôler certains types de données. Par exemple, les politiques du National Optical Astronomy Observatory (NOAO) aux États-Unis distinguent entre les données scientifiques soumises à des périodes d'utilisation exclusive et les types de données accessibles à la communauté « dès l'instant du dépôt de l'exposition dans l'archive du NOAO », comme

les métadonnées d'exposition, y compris la date et l'heure, la durée, le lieu et la configuration de l'instrument. Les données d'étalonnage interne sont également considérées comme publiques. Le personnel du NOAO a accès à l'ensemble des données des instruments afin de surveiller leur santé, leur sûreté, leur étalonnage et leurs performances. Le NOAO se trouve à Kitt Peak, dans l'Arizona, et est géré par l'Association of Universities for Research in Astronomy (AURA), inclus dans le cadre d'un accord de coopération avec la National Science Foundation (National Optical Astronomy Observatory, 2013a).

Éthique

En astronomie, des questions éthiques se posent autour de l'accès aux données et aux instruments, qui représentent des ressources rares et coûteuses. L'économie morale de l'astronomie détermine qui a accès à quel télescope, quand, pour combien de temps et avec quelles ressources pour recueillir et analyser des données (McCray, 2000, 2001, 2003). Ces dernières années, l'accès est devenu plus équitable et fondé sur le mérite grâce à l'augmentation des fonds publics consacrés à la discipline, mais la déontologie et la politique joueront toujours un rôle. Les télescopes sont financés par un ensemble complexe de partenaires, dont des universités, afin que leurs membres puissent accéder aux équipements dont ils ont besoin.

L'accès aux données astronomiques peut être retardé à cause de problématiques liées au traitement en *pipeline*, aux périodes d'utilisation exclusive, à la gouvernance, etc. Les données de la mission Planck, qui étudie le fond diffus cosmologique, ont ainsi été divulguées bien plus que tard qu'initialement prévu. Les personnels de recherche ont préféré diffuser une trentaine d'articles d'un seul coup, accompagnés des données, au motif que ces dernières n'étaient pas utilisables par d'autres avant d'être complètement étalonnées (Planck Collaboration *et al.*, 2013).

Les observations astronomiques peuvent constituer des informations sensibles en raison de leur valeur pour la navigation et la défense. Pan-STARRS, par exemple, est en partie financée par l'US Air Force pour surveiller des objets proches de la Terre ; les données jugées sensibles ne sont pas mises à la disposition de la recherche astronomique. Pan-STARRS distingue sa mission scientifique première de son rôle dans la défense des États-Unis (Pan-STARRS 2012, 2013b). Travailler de manière ouverte signifie généralement que les données sont divulguées au moment de la parution de l'article et non que les activités quotidiennes sont soumises en permanence au regard du public.

Mener des recherches en astronomie

La plupart des astronomes vivent dans un monde où les données abondent, de même que les outils et les services pour les sélectionner et les analyser. Elles et ils vivent également dans un monde de contraintes, où les missions se planifient sur le temps très long et où les financements, l'infrastructure et l'accès aux instruments et aux données dépendent de la coordination internationale. L'accès aux télescopes et au temps d'observation est plus équitable que par le passé, mais les scientifiques appartenant à des institutions riches, membres de grands consortiums d'instruments, disposent toujours de plus de ressources que celles et ceux travaillant dans des universités plus modestes ou des pays pauvres. Les décisions sur les entités qui deviendront des données et sur les données qui seront partagées, réutilisées ou conservées sont influencées par l'accès à ces ressources et par des contraintes temporelles, technologiques et infrastructurelles.

La présente étude de cas suit un laboratoire du Harvard-Smithsonian Center for Astrophysics (CfA) tout au long du relevé COMPLETE, un projet de plusieurs années, tandis que son équipe élabore des questions de recherche, recueille et analyse des données et publie ses découvertes dans plus de quarante articles. Notre étude se penche sur les infrastructures de la connaissance dont les scientifiques dépendent, sur la manière dont elles et ils représentent les données, sur les parties prenantes et sur les pratiques de publications.

Le relevé COMPLETE

Le relevé COMPLETE (abréviation de COordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions), basé au CfA, est un grand jeu de données créé à partir de référentiels publics d'observations astronomiques et de nouvelles observations des mêmes régions célestes. Il a cartographié intégralement trois immenses régions de formation stellaires de notre galaxie. Les observations effectuées couvrent le spectre électromagnétique des rayons X aux ondes radio. Le laboratoire a ensuite exploré ces données pour répondre à une série de questions de recherche. Le relevé COMPLETE est apprécié pour son exhaustivité, la diversité de ses sources et sa taille, que l'on estime être mille fois plus grande que les ressources coordonnées disponibles dix ans plus tôt (COordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions [COMPLETE], 2011).

Pendant environ sept ans, de nombreuses personnes, dotées de toutes sortes de savoir-faire, ont été impliquées dans la réalisation du relevé. La composition de l'équipe a varié d'une douzaine de membres à vingt-cinq, dont des professeurs et professeures, des directeurs et directrices de recherche, des postdoctorantes et

postdoctorants et des étudiantes et étudiants de premier et deuxième cycle². Le jeu de données de COMPLETE sert toujours à la recherche, surtout à des travaux d'observation et de statistique visant à comprendre la physique des régions de formation stellaire.

Les questions de recherche

Les questions de recherche du laboratoire du relevé COMPLETE s'intéressent à la manière dont le gaz interstellaire forme de nouvelles étoiles. La principale problématique de la recherche en formation stellaire est de comprendre quelles distributions d'étoiles se forment au fil du temps lorsque du gaz interstellaire se trouve dans certaines conditions. Cette question a été subdivisée en problèmes plus modestes, dont certains doivent être résolus avant de pouvoir traiter le reste du puzzle. Parmi les conclusions majeures du domaine à l'heure où nous écrivons, on trouve la découverte du « rayonnement de nuage » dans l'infrarouge moyen (Foster et Goodman, 2006), le développement et l'implémentation d'un algorithme d'identification de structure pour décrire des régions de formation stellaire réelles (Pineda *et al.*, 2009 ; Rosolowsky *et al.*, 2008) et simulées (Beaumont *et al.*, 2013), une réinterprétation du sens des températures sur les cartes de gaz interstellaire (Schnee *et al.*, 2006) et une évaluation du rôle de l'autogravitation dans la formation stellaire (Goodman *et al.*, 2009b).

La collecte de données

Pour identifier les données dans les archives existantes, le laboratoire a effectué des recherches basées sur les coordonnées et les noms d'objets dans des référentiels afin d'en extraire les données correspondant aux trois régions de formation stellaire étudiées (Persée, Ophioclus et Serpent). Il a également fait appel aux métadonnées de SIMBAD et ADS pour identifier d'anciens articles traitant des objets et des phénomènes célestes de ces régions. Cependant, les métadonnées contenues dans ces archives sont connues pour être incomplètes ; le laboratoire s'est donc fié au savoir-faire astronomique de son personnel et sur les sources mentionnées dans les articles pour savoir où chercher. Plus de la moitié des données du relevé COMPLETE sont nouvelles, issues de multiples demandes de temps d'observation sur des instruments de télescope. Ces nouvelles données ont été traitées dans les *pipelines* associés à chaque instrument. Bien que la procédure soit complexe, il faut noter que l'infrastructure de la connaissance de ce domaine de l'astronomie facilite la mise en correspondance d'observations anciennes et nouvelles issues d'instruments multiples.

2. NdT : Aux États-Unis, le premier cycle d'études supérieures dure quatre ans. Le deuxième cycle permet d'obtenir des diplômes de niveau master ou doctorat.

L'analyse des données

Les premières étapes de l'analyse de données ont consisté à convertir tous les jeux, qu'ils proviennent d'archives ou de demandes de temps d'observation, au format FITS. Combiner ces fichiers requiert une connaissance intime des normes FITS. Il a fallu faire des choix dans la manière de fusionner les fichiers. Les jeux de données ne couvrent pas exactement les mêmes zones du ciel, c'est pourquoi il y a davantage de données disponibles sur certaines régions de formation stellaire que sur d'autres. Si les métadonnées disponibles sur les positions célestes et les spectres sont essentielles à la fusion des jeux, une expertise considérable n'en est pas moins nécessaire pour corriger les différences d'étalonnage, de caractéristiques des instruments, de modèles de données et autres (Goodman *et al.*, 2009). Lorsque les jeux ont été fusionnés dans un même fichier, le laboratoire a pu employer une série d'outils, certains *open source* et d'autres commerciaux, pour utiliser des fichiers FITS et d'autres formats courants comme données d'entrée.

Les astronomes créent parfois leurs propres logiciels ou rédigent leurs propres scripts dans des outils existants pour obtenir un avantage concurrentiel. L'invention de nouveaux outils et la validation de nouvelles méthodes peuvent constituer des contributions scientifiques en soi (Rosolowsky *et al.*, 2008). Par exemple, dans leur article dans *Nature* sur le rôle de la gravité à différentes échelles, (Goodman *et al.*, 2009b) ont expliqué comment elles et ils ont utilisé des dendrogrammes de façon novatrice pour mesurer des structures sur une gamme d'échelles spatiales. Leurs visualisations 3D séparent dans le dendrogramme les régions en autogravitation de celles qui ne le sont pas pour montrer la supériorité de l'algorithme du dendrogramme sur un autre plus ancien, CLUMPFIND. Leur article a été le premier à paraître sous forme de PDF 3D permettant de manipuler et faire tourner des vues tridimensionnelles dans l'article (voir figure 2 dans Goodman *et al.*, 2009b) grâce à une version spéciale d'Adobe Reader.

La publication des résultats

COMPLETE a produit un corpus de recherche cumulatif pour le laboratoire, chaque étude obtenant des résultats qui venaient éclairer les suivants. Les articles individuels se fondent sur des sous-ensembles du relevé rassemblés dans un but spécifique, par exemple l'exploration du rôle de l'autogravitation dans la formation stellaire évoquée plus haut. Parce que plusieurs articles exploitent le même grand jeu de données, ils ne décrivent des protocoles de recherche que ce qui est nécessaire pour comprendre les spécificités de chacun. Les informations de provenance nécessaire à l'élaboration du relevé sont fournies dans le *data paper*, qui compte dix-sept auteurs et autrices d'institutions partenaires (Ridge *et al.*, 2006). L'article dans *Nature* ne fait que quatre pages parce que c'est le maximum autorisé par la revue ; douze pages de contenus

supplémentaires sont publiées en ligne. Cependant, on trouve une documentation complémentaire sur le site web du projet et dans les références des articles.

Les articles fondés sur le relevé COMPLETE ont été publiés dans des revues d'astronomie ; par conséquent, ils sont catalogués dans SIMBAD et ADS, ce qui les rend découvrables par objet, par région et par caractéristiques bibliographiques.

La conservation, le partage et la réutilisation des données

L'équipe du relevé COMPLETE est répartie entre plusieurs institutions de différents pays, chacune ayant ses propres pratiques de partage, de conservation et de réutilisation des données. Le noyau du laboratoire, basé au Harvard-Smithsonian Center for Astrophysics (CfA), assure l'entretien des jeux de données et du site web. Il met à disposition les données du relevé dans de nombreux formats et sous-ensembles à télécharger, toujours accompagné d'une riche documentation (COordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions [COMPLETE], 2011). Une suggestion de citation des jeux de données est fournie, mais le laboratoire ne cherche pas à traquer l'usage ou les renvois. Les jeux dérivés sont diffusés sur Dataverse, qui a récemment été adapté aux données astronomiques. Le jeu de données reste actif : on peut donc y ajouter de nouvelles divulgations, de la documentation et des corrections au fur et à mesure. L'équipe n'a pas déposé les données du relevé dans MAST ou dans d'autres référentiels, qui l'auraient déchargée de la conservation à long terme.

Le laboratoire du CfA qui a effectué le relevé COMPLETE et l'a exploité dans une série de projets de recherche est bien plus attentif à la conservation des données que la moyenne des scientifiques dans la plupart des disciplines. Il s'implique activement dans leur partage et dans le développement de l'infrastructure astronomique. Son équipe a créé le site Dataverse pour déposer et partager des données astronomiques, est responsable de l'intégration de données et de publications disciplinaires dans le relevé All-Sky de l'ADS, se montre très active sur le projet WorldWide Telescope et a étudié le partage et la réutilisation de données en astronomie (Goodman et Wong, 2009 ; Goodman, 2012 ; Goodman *et al.*, 2013 ; Pepe *et al.*, 2011).

Malgré la vaste documentation du relevé COMPLETE, le laboratoire reconnaît être toujours confronté au problème du « doctorant 1, 2, 3 », endémique dans la recherche collaborative à long terme (Edwards *et al.*, 2011 ; Pepe *et al.*, 2013). Lorsque se posent des questions autour de délicats choix d'étalonnage, de transformations ou d'autres procédures analytiques survenus des années plus tôt, il faut parfois retrouver l'étudiant ou la chercheuse postdoctorale qui était au cœur de ces décisions. Ces dernières ont pu, elles aussi, dépendre de choix effectués auparavant par d'autres

parties pendant le développement des instruments, des *pipelines* et des produits. C'est le problème de la « construction de la maison » : la provenance des données ne peut être retracée plus loin que leur naissance, or les données peuvent être nées plusieurs fois.

Comme la plupart des données numériques, les données astronomiques sont indisociables du code informatique utilisé pour les nettoyer, les réduire et les analyser. Celles sous forme de fichiers FITS, qui ont déjà été réduites par un traitement en *pipeline*, peuvent être analysées avec des suites logicielles courantes, qu'elles soient commerciales ou *open source*. Cependant, de nombreux types de codes associés aux données astronomiques peuvent ne pas être diffusés. Ce peut être parce que les astronomes élaborent leurs propres *pipelines* pour les données qu'elles et ils ont collectées avec leurs propres instruments ou grâce à une demande de temps d'observation. Ils peuvent rédiger leurs propres scripts ou outils spécialisés pour analyser des données publiques. Dans d'autres cas, le code des simulations informatiques est strictement protégé, mais celui lié à l'analyse et à l'interprétation des données de sortie est divulgué. Dans d'autres cas encore, les résultats des simulations peuvent ne pas être diffusés s'ils ne sont pas considérés comme des données. Ces exemples ne représentent que quelques-unes des nombreuses formes de travail scientifique des données que l'on trouve en astronomie.

La science et la technologie des réseaux de capteurs

À la fin des années 1990, les spécialistes des sciences exactes et des technologies se sont mis à déployer des réseaux de capteurs afin de recueillir des données de recherche en plus grandes quantités, à de plus hautes densités et de meilleure qualité qu'il n'était possible jusque-là. Ainsi, dans les sciences de l'environnement, les réseaux de capteurs sont utilisés pour étudier la croissance des plantes, la situation météorologique, la configuration des vents, l'activité chimique et biologique, l'activité animale et les sources de contamination du sol et de l'eau. Dix années de recherche collaborative sur les pratiques en matière de données au sein du Center for Embedded Networked Sensing (CENS), où des systèmes de détection sont conçus, testés et déployés dans le cadre d'applications scientifiques, ont produit de riches ressources pour la présente étude de cas. Les méthodes propres aux sciences sociales qui ont permis de la rédiger sont développées au chapitre 6.

Le CENS, qui dépendait de la National Science Foundation, a été financé de 2002 à 2012 et se consacrait à la conception de systèmes de détection à des fins scientifiques et sociales grâce à la collaboration de spécialistes de l'ingénierie, de

l'informatique et des sciences. Parce que leur travail transcendait les frontières disciplinaires, les participantes et participants devaient expliciter leurs pratiques, leurs méthodes et leurs exigences de recherche.

La taille compte

La science des réseaux de capteurs, dans les domaines du CENS, est un exemple emblématique de la « science légère » de Price (1963). Ce dernier la jugerait peut-être moins mûre que la mégascience qu'est l'astronomie en raison de l'hétérogénéité de ses méthodes de recherche, mais il vaut mieux la voir comme étant adaptative. Les méthodes utilisées sont rigoureuses et répondent aux critères de preuve de leurs disciplines respectives. Les données ainsi obtenues peuvent être hautement valides, mais elles sont difficiles à répliquer.

Le CENS a été lancé en 2002 par un noyau de chercheurs et chercheuses basés dans quatre, puis cinq universités californiennes. Ces chercheurs avaient chacun des collaborateurs et collaboratrices dans d'autres institutions. La composition du centre a varié d'année en année au gré des projets et à mesure de l'évolution de la liste des étudiantes, étudiants, professeures, professeurs, postdoctorantes, postdoctorants et personnels. À son apogée, le centre comptait trois cents participants et des pratiques variées en matière de données. En moyenne, environ 75 % de participants du CENS travaillaient à la conception et au déploiement de technologies de détection ; le reste se consacrait aux domaines d'application scientifiques, médicaux et sociaux. Les spécialistes des technologies (technologues) développaient de nouvelles applications – scientifiques ou autres – tandis que les scientifiques cherchaient de nouvelles technologies pour faire progresser leurs méthodes de recherche.

Les données recueillies dans les déploiements sur le terrain du CENS étaient massives au point de vue de la variété, à défaut de l'être en volume absolu ou en vitesse. Cependant, les équipes scientifiques ont été confrontées à des problèmes d'échelle à mesure que ces deux dernières dimensions augmentaient. Les réseaux de capteurs produisaient bien plus de données que les méthodes de prélèvement à la main qui avaient cours dans ces domaines. Dans une étude biologique de la croissance racinaire, par exemple, des scientifiques avaient collecté et encodé à la main une centaine d'images en sept ans, grâce à des appareils photo placés dans des tubes transparents et plantés dans le sol près des plantes étudiées. En automatisant les appareils et en envoyant les images sur le réseau de capteurs, elles et ils pouvaient prendre jusqu'à six mille clichés par jour, représentant environ dix gigaoctets de données (Vargas *et al.*, 2006). Passer des méthodes manuelles au codage automatisé s'est avéré problématique, pour plusieurs raisons. Le codage manuel reposait sur

le savoir-faire d'étudiantes et d'étudiants, dont certains encodaient pendant des heures. L'opération était difficile, car les racines étaient très petites et poussaient lentement. Lorsque les racines touchaient le tube, elles apparaissaient sous la forme de minuscules points. Une fois qu'elles avaient suffisamment grandi pour être visibles sur les parois du tube, les codeurs et codeuses étudiaient les images précédentes pour tenter de déterminer le moment de l'apparition des racines. Il fallait parfois, pour retracer l'origine de ces observations, fouiller des carnets de terrain, des cassettes vidéo de croissance racinaire et d'autres documents. On effectuait des tests de concordance entre codeurs, mais les marges d'erreur rendaient difficile une codification algorithmique des pratiques. Dans les études de biologie marine, les équipes scientifiques prélevaient généralement des échantillons d'eau trois ou quatre fois par période de vingt-quatre heures. Ces observations étaient corrélées en série temporelle. De leur côté, les réseaux de capteurs analysaient l'eau toutes les cinq minutes. À un tel rythme, les corrélations et les séries temporelles ne suffisaient plus, d'où l'adoption de techniques de modélisation complexes (Borgman *et al.*, 2012 ; Borgman *et al.*, 2007 ; Wallis *et al.*, 2007).

Quand est-ce une donnée ?

Tout au long des dix années d'existence du CENS, la notion de donnée y a été mouvante. Chaque chercheur, chercheuse, étudiante, étudiant et membre du personnel interrogé nous a décrit ses données à sa façon. Ces descriptions ont évolué au fil des entretiens suivants, des visites sur site et de l'examen de leurs publications. Des personnes appartenant à la même équipe proposaient des explications différentes de leurs données selon leur rôle au sein de l'équipe, leur expérience et les phases de l'activité de recherche. La notion de donnée a également évolué à mesure du progrès de la technologie des réseaux de capteurs et de la maturation des collaborations et des méthodes de recherche (Borgman *et al.*, 2012 ; Mayernik *et al.*, 2012 ; Wallis *et al.*, 2008).

La recherche du CENS était, dans l'ensemble, un mélange de méthodes exploratoires, descriptives et explicatives. Les scientifiques formulaient une hypothèse en laboratoire et la testaient sur le terrain ou vice-versa. Les spécialistes des technologies pouvaient éprouver certaines théories en laboratoire et d'autres sur le terrain. De même, les théoriciens, par exemple les ingénieures et ingénieurs en électricité qui modélisaient les réseaux de capteurs, pouvaient tester leurs théories lors du déploiement des réseaux de capteurs sur le terrain (Cuff *et al.*, 2008 ; Wallis *et al.*, 2007).

Les biologistes, les sismologues, les environnementalistes et autres spécialistes d'un domaine arrivaient au CENS en apportant leurs questions de recherche et

leurs méthodes. Leurs pratiques scientifiques, en particulier pour les applications biologiques et environnementales, présentaient les caractéristiques de l'écologie identifiées par Bowen et Roth (2007) : 1) le protocole de recherche est éminemment émergent, 2) les outils et les méthodes sont développés *in situ*, souvent à partir de matériaux disponibles sur place et sont très spécifiques au contexte, 3) les études sont difficilement répliquables en raison de la nature dynamique des systèmes écologiques et 4) les interactions sociales entre membres de la communauté de recherche sont extrêmement importantes.

Les spécialistes de l'informatique et de l'ingénierie ont eux aussi apporté leurs questions et leurs méthodes au CENS. Peu d'entre eux avaient de l'expérience dans la conception de matériel ou de programmes à des fins scientifiques, en particulier pour de la recherche de terrain dans les conditions imprévisibles du monde réel. La conception des technologies s'annonçait pour eux particulièrement éprouvante, puisque les exigences et les critères d'évaluation s'adaptaient avec fluidité aux questions de recherche (Borgman *et al.*, 2012). Il fallait en apprendre suffisamment sur le domaine de leurs partenaires pour concevoir, déployer et évaluer les technologies et les données qui en résulteraient.

La recherche au sein du CENS combinait du matériel acheté dans le commerce avec celui fabriqué sur place. Les décisions prises pour la conception des technologies de détection – matérielles comme logicielles – influaient sur les types de données pouvant être obtenues. Comme pour les instruments de télescope en astronomie, les choix de conception, faits bien avant le travail de terrain, pouvaient déterminer ce qui deviendrait une donnée.

Bien que les équipes scientifiques et techniques travaillassent ensemble sur le terrain, aucune ne maîtrisait pleinement les origines, la qualité ou les applications des données collectées ensemble. Certaines données recueillies lors des déploiements, comme les échantillons d'eau, n'intéressaient que les scientifiques, tandis que d'autres ne présentaient d'intérêt que pour les équipes techniques, par exemple les données proprioceptives des robots. Les deux parties avaient donc un intérêt commun dans les données issues des capteurs, mais l'appliquaient à des questions de recherche différentes. Parce qu'il s'agissait d'applications scientifiques nouvelles pour des technologies expérimentales, confronter les données à la réalité du terrain a représenté un défi permanent.

Sources et ressources

Les chercheurs et chercheuses du CENS recueillaient la plupart de leurs données eux-mêmes, ne disposant guère de référentiels ou d'autres sources externes où

puiser. Les équipes rassemblaient diverses sortes de données issues de réseaux de capteurs et d'échantillons physiques. Le code et les modèles logiciels étaient essentiels à la conception des instruments et à l'interprétation des données. D'ailleurs, ils étaient parfois traités comme des données.

Réseaux de capteurs intégrés

Même à l'époque de la fondation du CENS, en 2002, les réseaux de capteurs n'étaient pas des technologies nouvelles. On les utilisait pour faire fonctionner des processus industriels, par exemple dans des usines pétrochimiques, et pour surveiller le débit et la qualité de l'eau. Les nouveautés apportées par le CENS étaient l'utilisation de réseaux de capteurs intégrés pour poser de nouvelles questions scientifiques, ainsi que la collaboration de spécialistes des sciences et des technologies pour concevoir des technologies utilisables dans le monde réel (Committee on Networked Systems of Embedded Computers, 2001). Le CENS a su combiner des technologies du commerce avec de nouveaux appareils et de nouveaux protocoles de recherche afin de recueillir des données d'un genre inédit.

Une génération plus tôt, les technologies de télédétection révolutionnaient les sciences de l'environnement (Kwa et Rector, 2010 ; Kwa, 2005, 2011). La capacité à voir la Terre depuis des satellites, à des niveaux de granularité qui s'améliorent toujours, a permis une vision bien plus globale des phénomènes environnementaux.

Les réseaux de capteurs pouvaient être déployés sur terre et dans l'eau, selon la technologie employée. Comme le montre la figure 5.1, les capteurs pouvaient être enterrés dans le sol, accrochés à des bouées ou des bateaux, fixés à des poteaux ou d'autres installations, ou suspendus à des câbles que l'on déplace sur trois dimensions au-dessus du sol. Ils étaient utilisés pour détecter des indicateurs de nitrates dans l'eau, d'arsenic dans des rizières, de la vitesse et de la direction du vent, des niveaux de luminosité, de mouvements physiques de la terre ou d'animaux et divers autres phénomènes. Les données étaient extraites des capteurs soit à la main, par exemple en les copiant sur une clé USB, soit en les envoyant à un nœud doté d'un accès internet. L'utilisation des données de ces capteurs et les méthodes d'obtention variaient en fonction des applications, des technologies choisies et de la distance du lieu. Certains déploiements prenaient place dans des zones urbaines disposant d'ores et déjà de réseaux sans fil, mais beaucoup se faisaient dans des montagnes, des îles et des déserts isolés.

Figure 5.1. Science des réseaux de capteurs

Technologies de réseaux de capteurs employées au CENS

Crédit : Jason Fisher.

Les technologues du CENS utilisaient les données des capteurs en leur attribuant plusieurs rôles : 1) observations de phénomènes physiques et chimiques, y compris des sons et des images, 2) observations de phénomènes naturels utilisées pour actionner des capteurs robotiques ou les guider à un certain endroit dans l'environnement, 3) données de performance issues des capteurs à leur propre sujet, comme les heures où ils sont activés ou en sommeil, les défauts qu'ils détectent, la tension de la batterie et les tables d'acheminement du réseau et 4) données proprioceptives recueillies par les capteurs, c'est-à-dire des données pour guider les robots, comme la vitesse du moteur, le cap, le roulis, le tangage, le lacet et l'angle de barre (Borgman *et al.*, 2012).

Les différences les plus visibles entre les disciplines participantes se manifestaient dans les critères de preuve. Les biologistes, par exemple, mesuraient des variables telles que la température conformément aux normes de leur domaine. En général, les spécialistes de l'ingénierie et de l'informatique ignoraient l'existence de ces standards internationaux ou ne s'en préoccupaient guère. Pour leur travail, une référence locale de mesures constantes suffisait à l'étalonnage. Questionné sur ses pratiques de mesure, un technologue a simplement répondu : « une température est une température ». Quand nous avons demandé à un biologiste partenaire comment mesurer une température, il nous a donné une réponse longue et nuancée où il abordait le type d'instrument, où et quand la mesure était effectuée, le degré de contrôle sur l'environnement, la précision de l'instrument et les registres d'étalonnage. Cet homme avait d'ailleurs installé trois types de thermomètres côte à côte sur un terrain de recherche et avait noté les mesures effectuées pendant une année entière avant de se fier aux instruments et aux données qu'ils produisaient (Wallis *et al.*, 2007).

Échantillons physiques

Les équipes scientifiques du CENS ont continué à prélever physiquement des échantillons d'eau, de sable et de sol. Elles effectuaient ainsi des observations d'organismes vivants, comme la distribution de phytoplancton et de zooplancton dans un lac. Les échantillons étaient testés dans des laboratoires expérimentaux sur place et, parfois, examinés plus précisément dans le centre de recherche, après le déploiement.

Logiciels, code, scripts et modèles

Un capteur ne mesure pas directement le vent, l'arsenic, les nitrates ou d'autres variables scientifiques ; en réalité, il mesure le voltage et d'autres indicateurs détectables. Les sorties de capteurs sont généralement constituées de signaux binaires qui doivent être interprétés grâce à des modèles statistiques. Certaines consistent en des photographies. On utilise des modèles statistiques de phénomènes physiques ou chimiques pour interpréter ces indicateurs (Batalin *et al.*, 2004 ; Deshpande *et al.*, 2004 ; Hamilton *et al.*, 2007 ; Wallis *et al.*, 2007).

Les équipes techniques avaient parfois recours à des ressources informationnelles externes, comme des référentiels de code. Le code, le logiciel et les modèles tendaient à être considérés comme des synonymes de la donnée pour les spécialistes de l'informatique et de l'ingénierie interrogés (Wallis *et al.*, 2013).

Données de base

Les équipes scientifiques utilisaient des données issues de ressources externes pour planifier la collecte de nouvelles sources sur des terrains particuliers. Parce qu'elles retournaient généralement sur les mêmes, elles avaient besoin de beaucoup d'informations de base et de contexte sur ces terrains. Les données recueillies par des organismes publics comme le Department of Fish and Game de Californie ont constitué des ressources essentielles, de même que les données collectées par l'équipe lors de précédentes visites au lac. Parmi les informations contextuelles importantes sur le lac, on citera les mois où les algues prolifèrent le plus, la topologie du fond du lac, les espèces de phytoplanctons et de zooplanctons susceptibles d'être trouvées, la présence de nutriments et leur concentration. Les équipes techniques ont, elles aussi, parfois tiré leurs données d'étalonnage de sources externes (Wallis *et al.*, 2013).

Les infrastructures de la connaissance

Alors que l'astronomie a accumulé une infrastructure de la connaissance sophistiquée pour coordonner données, publications, outils et référentiels au cours de longues décennies de coopération internationale, le CENS se situe à l'autre bout du spectre. Le centre lui-même remplissait une fonction fédératrice essentielle, afin de réunir des

chercheurs et chercheuses autour d'intérêts communs. Il fournissait l'infrastructure technique sous forme de matériel, de réseaux et de personnel, mais a peu investi dans des ressources informationnelles partagées. Les publications étaient versées dans le système *eScholarship* de l'université de Californie, formant l'un de ses plus grands dépôts (Center for Embedded Networked Sensing, 2013 ; Pepe *et al.*, 2007).

À l'origine, la recherche du CENS reposait sur les technologies de réseaux de capteurs autonomes et de « poussière intelligente ». Si cette orientation s'était maintenue, mettre en place des structures normalisées de données et métadonnées aurait été bien plus faisable. À mesure que le centre mûrissait, ses participantes et participants ont mieux compris les problèmes scientifiques et technologiques posés par ce type de recherche exploratoire. Les technologies expérimentales ont alors été jugées trop fragiles et capricieuses pour être laissées sans surveillance sur le terrain. La recherche du CENS a par la suite évolué vers des méthodologies « mettant l'humain dans la boucle » et adaptables sur le terrain.

Métadonnées

Les travaux scientifiques employant des réseaux de capteurs constituent un domaine de recherche et non une discipline qui serait dotée d'une histoire comparable à celle de l'astronomie. Chaque partenaire apporte à la collaboration ses propres pratiques disciplinaires, notamment son usage des métadonnées. À part pour la sismologie et la génétique, peu de standards de métadonnées existaient pour les données produites par les équipes du CENS. Quand il y en avait, ils n'étaient pas pour autant forcément adoptés par la communauté scientifique ou par les laboratoires locaux – ce qui est d'ailleurs caractéristique de la science légère – (Cragin *et al.*, 2010 ; Millerand et Bowker, 2009). Ainsi, des standards XML officiels existaient, par exemple, pour les données environnementales et les données de capteurs, mais ils n'ont pas été utilisés (Knowledge Network for Biocomplexity, 2010 ; Open Geospatial Consortium, 2014 ; Higgins *et al.*, 2002 ; Ruixin, 2002).

Certaines équipes utilisaient cependant des métadonnées pour leurs besoins propres, mais pas celles des standards XML. Elles créaient des registres qui décrivaient le contexte de la collecte de données, comprenant l'heure précise, le lieu, les conditions locales, la position des capteurs, les caractéristiques de ces derniers (fabricant, modèle, numéro de série et autres). Les conventions de nommage représentaient la forme plus commune de métadonnées. Les équipes étaient rarement satisfaites de la qualité des métadonnées utilisées et se rendaient souvent compte que leurs anciennes données étaient difficiles à retrouver ou à réutiliser.

Provenance

Chaque équipe du CENS gérait ses propres registres de traçabilité. Elles avaient fréquemment recours aux données tirées de déploiements antérieurs ou d'études en laboratoire pour comparaison. Les scientifiques du CENS tendaient à retourner régulièrement sur un même site de recherche pendant de nombreuses années. Elles et ils accumulaient connaissances et données sur leur terrain, ce qui leur permettait d'effectuer des comparaisons longitudinales. En revanche, les spécialistes des technologies dépendaient bien moins des sites en tant que terrain concret ou des comparaisons longitudinales : les essais initiaux de leurs instruments pouvaient être aussi bien conduits dans des baignoires, des piscines, des jardins et des espaces verts du centre.

Ne disposant pas, contrairement aux astronomes, d'un réservoir commun de données, les scientifiques des champs associés au CENS n'avaient pas d'équivalent des *data papers* pour établir la provenance d'un jeu. À la place, elles et ils conservaient les informations de traçabilité en local, comme les données. Les personnes désireuses de réutiliser des données du CENS contactaient généralement les auteurs et autrices des publications où elles étaient rapportées (Wallis *et al.*, 2013).

Les influences extérieures

La science des réseaux de capteurs attire des chercheurs et chercheuses issus de parcours académiques très variés. Chaque personne et chaque groupe a apporté à la collaboration ses propres valeurs économiques, ses préoccupations en matière de propriété et ses problématiques éthiques. C'est à l'intersection des domaines de recherche que quelques-uns des problèmes les plus intéressants et les plus inattendus se sont posés.

Économie et valeur

Les façons d'obtenir et d'échanger des données au sein du CENS variaient en fonction des domaines et des circonstances de chaque projet. Les sismologues et les biologistes marines et marins disposaient de quelques réservoirs communs de ressources. Les environnementalistes puisaient des informations météorologiques, hydrologiques ou concernant d'autres aspects des conditions de terrain dans des registres à l'échelle locale, étatique et fédérale. Dans la mesure où ces chercheurs et chercheuses travaillaient sur le territoire des États-Unis, la plupart de ces documents sont considérés comme des biens publics. Leur usage peut également être autorisé hors des États-Unis, auquel cas ils peuvent devenir des biens de club ou des réservoirs communs de ressources, selon la gouvernance en vigueur. Les scientifiques du CENS recueillaient des données d'observation dans beaucoup d'autres pays et l'accès aux données de base sur les conditions locales variait en fonction. Des référentiels de

logiciels ouverts comme GitHub et Source Forge servent de réservoirs communs de ressources pour la communauté des spécialistes de l'informatique et de l'ingénierie, qui utilisaient parfois des logiciels qui y sont développés, ou y déposaient du code (Boyle et Jenkins, 2003 ; Kelty, 2008 ; Uhler, 2006 ; Wallis *et al.*, 2013). Dans d'autres cas, comme nous le verrons ci-dessous et au chapitre 8, les échanges de données consistaient en des accords informels entre individus.

La communauté sismologique états-unienne est soutenue par les Incorporated Research Institutions for Seismology (IRIS), un consortium d'universités qui gèrent des établissements consacrés à l'acquisition, la gestion et la diffusion de données sismologiques (Incorporated Research Institutions for Seismology, 2013). Les données sismiques sont utilisées non seulement pour la recherche scientifique et l'éducation, mais aussi pour la prévention des risques liés aux tremblements de terre et pour la vérification du traité d'interdiction complète des essais nucléaires (IRIS Data Management Center, 2013). Les données issues de subventions octroyées par la National Science Foundation doivent être mises à disposition dans le référentiel de l'IRIS dans un délai particulier à compter du retrait du terrain du dernier élément de matériel d'un projet financé. Si les règles concernant les périodes d'utilisation exclusive étaient respectées, il n'en reste pas moins que les chercheurs et chercheuses bénéficiaient d'une grande marge de manœuvre pour décider quand ils retireraient les équipements sismiques du site. Ils pouvaient choisir de temporiser le retrait des capteurs afin de gagner du temps pour analyser leurs données (Wallis *et al.*, 2013).

Le CENS a recueilli très peu de données génomiques, mais certaines recherches sur la prolifération d'algues nuisibles et la qualité de l'eau comprenaient des analyses ADN. Lorsque les organismes de financement ou les revues l'exigeaient, ces données étaient versées à la GenBank, la Protein DataBank ou d'autres archives (Wallis *et al.*, 2013). Certaines données environnementales recueillies par télédétection de satellites possèdent une grande valeur commerciale, comme le suivi de bancs de poissons ou de conditions météorologiques permettant de prévoir les récoltes (Kwa, 2005). La plupart des projets en sciences de l'environnement recueillaient de petites quantités de données sur des sites de recherche particuliers. Rares étaient les données organisées de manière à pouvoir être facilement combinées ou comparées. Mais prises dans leur ensemble, ces données avaient le potentiel d'être précieuses pour d'autres (Borgman *et al.*, 2007b ; Wallis *et al.*, 2010b).

Droits de propriété

Comme en astronomie, les droits de propriété en science et technologie des réseaux de capteurs concernent davantage les instruments que les données. Alors que les astronomes partagent des instruments de grande taille, les scientifiques et les

technologiques du CENS avaient plutôt tendance à acheter ou fabriquer leurs propres petits appareils pour recueillir des données. Le matériel acquis grâce à des fonds publics devient généralement la propriété de l'université subventionnée.

Plusieurs entreprises se sont associées au CENS, apportant leur expertise, du matériel et des fonds supplémentaires. De petites sociétés ont été formées pour vendre certains des équipements, des algorithmes et des méthodes élaborés au CENS. Aucune n'a connu de réel succès commercial ; leur objectif était surtout le transfert de technologies. La démarche globale du centre tendait vers la science ouverte, privilégiant la diffusion de logiciels sous forme de code *open source*. L'un des projets commerciaux les plus réussis a consisté en une organisation à but non lucratif fondée par des anciennes et anciens du CENS, qui conçoit encore aujourd'hui des technologies de capteurs en réseau pour des applications environnementales, médicales et économiques (Nexleaf, 2013).

Éthique

Dans la création de données scientifiques et techniques issues de réseaux de capteurs, les problèmes éthiques se posent au moment de décider quels phénomènes étudier, où et comment, et avec quel degré de précision rapporter les résultats. Par exemple, des scientifiques du CENS étudient des espèces ou des habitats menacés. Les résultats publiés incluent suffisamment de détails pour valider les travaux, mais pas assez pour repérer la localisation exacte des sites. En effet, les recherches effectuées au sein de réserves naturelles protégées étaient souvent sensibles. Ces dernières se trouvent généralement dans des endroits isolés, bien que certaines puissent être ouvertes au public pour des activités éducatives. Les visiteurs et visiteuses, qu'ils viennent pour se divertir ou effectuer des recherches, sont tenus de respecter les habitats et les écosystèmes afin que la flore, la faune et les phénomènes puissent être étudiés dans des conditions naturelles.

Les spécialistes de l'informatique et de l'ingénierie du CENS étaient supposés se conformer aux normes éthiques de leurs disciplines. Le code de déontologie de l'Association for Computing Machinery (ACM), la principale organisation professionnelle d'informaticiennes et informaticiens, couvre les impératifs moraux généraux (respecter autrui, ne pas nuire, se montrer juste et honnête, respecter la vie privée, honorer la confidentialité, etc.), la responsabilité professionnelle (travail de qualité supérieure, connaissance et respect des lois en vigueur, évaluation des systèmes et des risques, etc.), l'encadrement du personnel et le respect du code (Association for Computing Machinery, 1992). L'IEEE, une grande organisation professionnelle du génie, dispose d'un code de déontologie similaire, quoique moins détaillé, qui mentionne des devoirs tels que « prendre des décisions conformes à la sécurité, à la

santé et au bien-être du public », « reconnaître et corriger ses erreurs » et éviter de blesser autrui (Institute of Electrical and Electronics Engineers, 2013).

Selon ces principes, les chercheurs et chercheuses doivent recueillir les données de manière responsable, mais la définition de la « responsabilité » varie selon les domaines. Par exemple, des spécialistes de l'ingénierie et de la biologie ont travaillé ensemble pour adapter des algorithmes à visée militaire à la localisation de chants d'oiseaux. Des scientifiques qui s'étaient jusqu'alors tenus à l'écart des applications militaires se sont ainsi retrouvés à déployer des technologies d'armement à des fins pacifiques. Une équipe d'informaticiens et informaticiennes, qui adaptaient des capteurs photographiques pour visualiser les mouvements des animaux sur le terrain, ont installé leurs appareils dans les couloirs du campus pour les tester, avant d'être accusés d'enregistrer le comportement de sujets humains sans leur consentement (Wallis *et al.*, 2007).

Alors que le CENS étendait ses activités à d'autres applications des réseaux de capteurs, les spécialistes de l'utilisation scientifique de ces technologies se sont tournés vers les usages sociaux des appareils mobiles. Les téléphones portables sont devenus une plateforme importante de recueil des données et d'étude de la topologie de réseau. Lorsque les participantes et participants se sont mis à suivre leur propre comportement – alimentation, itinéraires, usage du vélo et autres – grâce à des applications mobiles, les inquiétudes quant à la vie privée ont atteint leur paroxysme. Les spécialistes de l'informatique et de l'ingénierie ont dû prendre des décisions difficiles sur les données qu'elles et ils pouvaient collecter, par rapport à celles dont ils avaient besoin. Le CENS est devenu le lieu d'une étude pluriannuelle sur la construction de valeurs au sein de la conception de technologies mobiles (Shilton, 2011).

Mener des recherches avec des réseaux de capteurs intégrés

Le CENS a soutenu de nombreux projets indépendants en même temps, même si une partie du personnel, du matériel et des pratiques se partageaient entre plusieurs projets. Ces collaborations tiraient les chercheurs et chercheuses hors de leur zone de confort : les technologues devaient tester du nouveau matériel dans des conditions de terrain imprévisibles et les scientifiques devaient faire confiance aux technologues pour la réussite de leurs excursions (Mayernik *et al.*, 2012).

De petites équipes de recherche – constituées d'un mélange d'étudiantes, étudiants, professeurs, professeurs et personnels de recherche issus de domaines variés – travaillaient ensemble sur un site donné sur des périodes allant de quelques heures à deux semaines. On parlait alors de « déploiement », puisque les capteurs étaient

déployés pour recueillir des données de diverses sortes. La participation variait selon les jours, s'élevant jusqu'à une vingtaine de personnes.

Nous vous présentons ci-après un scénario composite d'un déploiement typique au sein du CENS, que nous décrivons plus en détail dans une autre publication (Borgman *et al.*, 2012). Il illustre une série d'activités communément associées à la collecte, à la gestion, à l'utilisation et à la conservation de ces types de données. Ce scénario concerne une prolifération d'algues nuisibles (ou HAB, pour Harmful Algal Bloom), un phénomène où un type d'algue particulier devient soudainement dominant dans l'eau et qui peut se produire en eau douce comme dans les océans. La prolifération crée une toxicité qui tue les poissons et d'autres animaux comme les lions de mer, soit en consommant l'oxygène dont les poissons ont besoin pour vivre, soit en relâchant de l'acide domoïque, une neurotoxine qui nuit aux grands mammifères. Il est important d'étudier les HAB, car elles peuvent causer de grands dommages, jusqu'à tuer des dizaines de milliers de poissons par jour. Les déploiements HAB ont été réalisés dans un lac connu pour ses proliférations estivales.

Les réseaux de capteurs permettent aux spécialistes de la biologie marine d'étudier davantage de variables qu'en prélevant des échantillons à la main et d'effectuer bien plus d'observations. La collecte de données peut s'adapter aux conditions locales en choisissant judicieusement les capteurs et leur placement. Les études des HAB au moyen de réseaux de capteurs permettent aux spécialistes de l'informatique et de l'ingénierie de tester les capacités de capteurs physiques et biologiques à recueillir un grand nombre de variables. Les roboticiennes et roboticiens jugent les HAB particulièrement intéressantes parce que le flux des observations peut servir à déclencher des systèmes de détection sur des bateaux, des bouées, des hélicoptères, des appareils photo et des véhicules autonomes.

Les questions de recherche

L'objectif général du CENS était le développement conjoint – ou co-innovation – de nouveaux instruments qui permettraient de nouvelles formes de science (Center for Embedded Networked Sensing, 2012 ; Committee on Networked Systems of Embedded Computers, 2001 ; Estrin, 2008). La science et la technologie progressaient en symbiose, comme dans la métaphore de la guêpe et du figuier de Licklider (Borgman, 2011 ; Licklider, 1960 ; Waldrop, 2001). L'une ne pouvait avancer sans l'autre ; elles étaient interdépendantes et s'influençaient mutuellement.

En dépit de l'interdépendance des équipes scientifiques et techniques, leurs objectifs à long terme étaient plus proches des buts de leurs disciplines respectives que de ceux du centre. Les biologistes ont poursuivi leur étude des phénomènes biologiques

liés aux HAB avant, pendant et après le CENS et les technologues ont continué d'améliorer leurs instruments, leurs algorithmes et leurs modèles pour cibler des phénomènes avant, pendant et après le CENS. Leur sélection des données, leurs manières de les gérer et les endroits où ils les publiaient correspondaient aux programmes de leurs disciplines respectives, malgré leur forte implication dans la collaboration (Borgman *et al.*, 2012 ; Mayernik *et al.*, 2012 ; Wallis *et al.*, 2013).

Lors de la recherche de terrain sur les HAB, l'équipe scientifique a étudié la distribution des phénomènes dans le lac, tandis que l'équipe technique se concentrait sur la vision robotique (Borgman *et al.*, 2012). Les exigences des scientifiques ont permis de mener des recherches technologiques sur des algorithmes de guidage robotique, de santé du réseau, de détection des défaillances du capteur et de conception d'interfaces de technologies des capteurs. Les spécialistes de l'informatique et de l'ingénierie se sont appuyés sur les discussions avec l'équipe scientifique pour choisir leur matériel et les capteurs spécifiques, ainsi que l'heure, le lieu et la durée du déploiement de chacun.

La collecte de données

D'un déploiement à l'autre, le nombre de personnes et la distribution des compétences variaient considérablement. Lors des quatre jours du déploiement au lac pour étudier les HAB, la participation a fluctué au quotidien. Le premier jour, les étudiantes, étudiants et personnels de recherche sont arrivés pour installer le matériel. Le deuxième jour, les professeures et professeurs sont venus diriger la collecte de données. Dans cet exemple, une vingtaine de personnes sont allées et venues en quatre jours. Parmi elles, il y avait huit à dix spécialistes en génie électrique qui ont fabriqué le système de détection, quatre ou cinq membres de l'équipe robotique, deux statisticiennes et statisticiens et six à huit chercheurs et chercheuses en biologie marine. Les responsabilités se chevauchant, les chiffres de participation sont approximatifs.

Bien que toutes les parties soient venues sur le terrain avec une série de questions de recherche et l'instrumentation correspondante, la collecte de données dépendait grandement des conditions locales. Les chercheurs et chercheuses ont sélectionné et positionné les capteurs avec soin, puisque chacun d'eux était adapté à certains types de conditions et pouvait enregistrer des données à des densités particulières (Akyildiz *et al.*, 2002). Le placement des capteurs constitue d'ailleurs un sujet de recherche à part entière (Younis et Akkaya, 2008). En effet, des facteurs comme l'humidité du sol ou le pH du lac influencent l'endroit où on les placera. Les capteurs pouvaient être déplacés plusieurs fois au cours du déploiement en fonction des résultats intermédiaires. Certains placements étaient fondés sur des probabilités

conditionnelles, comme la capacité des roboticiens et roboticiennes à déplacer automatiquement les capteurs aux endroits où des HAB étaient prévus.

Pour la recherche sur les HAB, les deux équipes avaient besoin d'observations de phénomènes chimiques et physiques (par exemple, les concentrations de nitrate par heure, emplacement et profondeur du lac) et de phénomènes naturels (par exemple, la distribution du phytoplancton et du zooplancton) qui puissent être utilisées pour guider les capteurs robotiques. Les scientifiques avaient aussi besoin d'échantillons d'eau contenant des organismes vivants ; ceux-ci étaient testés dans des laboratoires expérimentaux sur place et d'autres étaient testés plus tard sur le campus, après le déploiement. Les technologues avaient en outre besoin de données proprioceptives et de performances sur les capteurs (Borgman *et al.*, 2012).

La fiabilité des capteurs variait beaucoup, ce qui a représenté une grande source de frustration dans les premières années du CENS. Les capteurs fonctionnaient de manière erratique, s'arrêtaient ou redémarraient au hasard, semblait-il. Dans ce dernier cas, les horloges intégrées repartaient à zéro, ce qui rendait impossible la mise en correspondance des données recueillies sur l'ensemble du réseau. Les réseaux de capteurs pour la recherche du terrain telle que la pratiquait le CENS s'avéraient bien moins fiables et autonomes qu'attendu. À la suite d'une désastreuse perte de données sur un site positionné sur un autre continent, le centre est passé à des méthodes à intervention humaine. Cette approche était plus adaptée à l'évaluation de la qualité des données en temps réel (Mayernik *et al.*, 2012).

En pratique, la technologie des capteurs était toujours expérimentale et s'adaptait donc en continu aux nouvelles questions de recherche et aux nouvelles possibilités techniques. Elle n'a jamais été suffisamment stabilisée pour standardiser les procédures de collecte de données. C'est une des caractéristiques qui font de la recherche du CENS une science légère (Borgman *et al.*, 2006, 2007a).

L'analyse des données

Les données étaient traitées avant, pendant et après les déploiements sur le terrain. Les équipes du CENS consacraient des efforts considérables à vérifier les instruments et les données des capteurs pour s'assurer de leur fiabilité. La méthode utilisée, la vérification au sol, consiste à utiliser des techniques de mesure connues pour éprouver la validité des nouvelles méthodes.

Les équipes scientifiques avaient la parfaite maîtrise des échantillons physiques d'eau, de sol et d'autres matériaux prélevés sur le terrain. Certains étaient traités sur place, d'autres dans des laboratoires du campus. Les équipes techniques,

en revanche, n'avaient guère d'intérêt pour ces données. Rendre intelligibles des données de capteurs suppose des modèles scientifiques sous forme d'algorithmes statistiques. Ces modèles scientifiques, développés conjointement par les équipes scientifiques et techniques, étaient considérés comme les produits les plus importants de la recherche du CENS.

Au cours des déploiements, par exemple celui concernant la prolifération d'algues nuisibles, les données des capteurs étaient envoyées aux ordinateurs des équipes techniques responsables des réseaux. Celles-ci les étalonnaient et les nettoyaient, ce qui impliquait de mettre en correspondance les différentes références temporelles des capteurs, de supprimer les artefacts tels que le redémarrage d'un appareil dû à une défaillance informatique et d'ajouter des notes concernant les décisions de terrain sur le moment où un capteur était déplacé, la manière et l'endroit (Wallis *et al.*, 2007). Une fois nettoyées et étalonnées, les données étaient fournies aux équipes scientifiques participantes. Les scientifiques les comparaient alors à leurs modèles et d'autres sources de données de confiance, par exemple les courbes d'étalonnage établies en laboratoire (Borgman *et al.*, 2012). La plupart des données n'étaient retenues que par l'équipe qui les avait recueillies, quand elles l'étaient.

La différence d'objectifs à long terme entre les équipes scientifiques et techniques se manifestait surtout dans le traitement des données. Des désaccords sont apparus autour des normes incompatibles, des pratiques de partage et de l'assistance à la gestion des données (Edwards *et al.*, 2011 ; Mayernik *et al.*, 2011 ; Mayernik, 2011). Dans les déploiements purement scientifiques ou purement techniques, les équipes qui recueillaient les données se chargeaient aussi de leur traitement. Dans les déploiements conjoints, la gestion et le traitement final variaient selon le type de données.

La publication des résultats

Les recherches issues des déploiements sur le terrain du CENS ont été publiées dans les revues et les colloques de chaque discipline participante. Dans de nombreux cas, les résultats scientifiques et technologiques des déploiements ont été publiés séparément, dans des publications visant le lectorat de leurs domaines respectifs. Dans d'autres, les résultats ont été publiés conjointement par des auteurs et autrices de différents champs de recherche. Une étude des schémas d'auctorialité et d'interconnaissance parmi les chercheurs et chercheuses du CENS a montré comment de nouvelles collaborations se sont formées et ont évolué au cours des dix années d'existence du centre (Pepe, 2010, 2011).

Les participantes et participants aux activités de recherche comprenaient des professeures et professeurs, des chercheurs postdoctoraux et chercheuses postdoctorales, des étudiantes et étudiants de premier et deuxième cycle et des personnels de recherche à plein temps. Les étudiants et le personnel impliqués dans la conception et le déploiement du matériel de recherche étaient considérés comme faisant partie de l'équipe, cependant le personnel n'a pas toujours été crédité dans les publications. Le statut auctorial était d'ailleurs un sujet très sensible pour celles et ceux qui concevaient et entretenaient les instruments nécessaires au projet. L'instrumentation du CENS évoluait continuellement parce que la co-innovation de la science et de la technologie était essentielle à la poursuite des objectifs du centre.

La conservation, le partage et la réutilisation des données

À la fin du déploiement sur le terrain, les équipes se dispersaient, chaque membre emportant les données dont il avait la responsabilité. Par conséquent, les données issues de ces déploiements en commun étaient disséminées et peu susceptibles d'être rassemblées de nouveau. Il n'existait que peu de registres de provenance utilisables pour reconstituer ou répliquer un déploiement donné (Borgman *et al.*, 2012 ; Wallis *et al.*, 2013).

Chaque équipe avait documenté ses données de manière suffisante pour ses propres besoins. Certaines, notamment dans les sciences exactes, les tenaient à jour pour des comparaisons futures. D'autres, surtout dans l'ingénierie, n'avaient guère l'usage des données des déploiements une fois les articles publiés. Les équipes du CENS ont peu investi dans la création de métadonnées (Mayernik, 2011). Les conventions de nommage des fichiers représentaient souvent la forme de gestion la plus sophistiquée employée par les équipes de recherche. Les tableurs étaient généralement le plus petit dénominateur commun pour l'échange de données au sein d'un groupe et entre plusieurs groupes.

L'adaptabilité des méthodes de collecte de données au CENS a produit des jeux utilisés localement et difficiles à combiner avec d'autres. Peu de chercheurs et chercheuses ont déposé leurs données dans des bases, en partie à cause de l'absence de standardisation et en partie à cause du manque de référentiels susceptibles de les accueillir. La responsabilité des données allait généralement de pair avec les rôles dans la rédaction des publications (Wallis, 2012), comme nous le verrons plus avant au chapitre 9 (Chang *et al.*, 2006 ; Wallis *et al.*, 2010b ; Wallis, 2012).

En principe, les standards de métadonnées existants auraient pu être utilisés – individuellement ou combinés entre eux – pour décrire une grande partie des données du CENS. Ainsi, les observations écologiques recueillies par les capteurs ou à la main

auraient pu être décrites avec une structure et un vocabulaire communs. De même, si ces standards XML avaient été intégrés aux algorithmes de collecte, les caractéristiques des capteurs auraient pu être enregistrées automatiquement.

Cependant, les structures de métadonnées formelles s'intègrent difficilement dans des activités de recherches aussi locales et adaptatives (Aronova *et al.*, 2010 ; Mille-
rand et Bowker, 2009 ; Ribes *et al.*, 2005). Les langages de métadonnées conçus pour les besoins des catalographes, des indexeurs et des indexeuses s'adaptent difficilement à une utilisation légère par des personnels de recherche. L'Ecological Metadata Language, par exemple, s'accompagne d'un manuel de plus de deux cents pages (Knowledge Network for Biocomplexity, 2013). Un lexique pour la recherche aquatique y comporte plus de dix mille entrées, avec quatre cents termes rien que pour les nitrates. Les chercheurs et chercheuses du CENS ont trouvé l'envergure de ces langages de métadonnées décourageante : ils ne pouvaient justifier l'effort nécessaire à leur mise en œuvre, même à petite échelle. Le centre ne disposait pas du personnel suffisant pour accompagner un tel niveau de gestion professionnelle des données. Les tentatives de construction d'un référentiel pour soutenir la recherche interdisciplinaire au moyen des réseaux de capteurs ont été peu fructueuses, en grande partie du fait de l'hétérogénéité des données et de la diversité des pratiques locales de gestion (Wallis *et al.*, 2010b).

Bien que la gestion des données soit informelle dans la plupart des équipes du CENS, les chercheurs et chercheuses étaient généralement désireux de les partager. Les conditions auxquelles ils se sentaient prêts à le faire variaient cependant considérablement, allant de la divulgation immédiate de toutes les données brutes à l'exigence d'être coauteur ou coautrice des publications reposant sur celles-ci. La plupart d'entre eux diffusaient volontiers leurs données une fois les articles correspondants publiés (Borgman *et al.*, 2006). Une partie des données et du code logiciel était versée dans des référentiels publics, mais la forme la plus courante de partage pratiquée par les chercheurs était les échanges interpersonnels à la demande (Wallis *et al.*, 2013).

Conclusion

Les cas de l'astronomie et des réseaux de capteurs intégrés illustrent, par leurs différences, la diversité de la recherche, des pratiques et des données. L'astronomie est une discipline bien établie disposant d'un noyau de revues et de colloques. Les applications scientifiques des réseaux de capteurs intégrés constituent davantage un domaine problématique qu'une discipline, mais disposent aussi de plusieurs

revues et colloques. Les chercheurs et chercheuses en astronomie et en réseaux de capteurs travaillent grâce à une instrumentation commune. Néanmoins, les télescopes et les archives de données constituent d'importants investissements infrastructurels, régis dans l'optique d'engendrer des milliers de publications, du savoir partagé et des données qui peuvent potentiellement conserver leur valeur indéfiniment. En comparaison, les investissements dans l'infrastructure de la science et de la technologie des réseaux de capteurs sont minimes.

L'utilisation de technologies de détection intégrées en réseau pour étudier des phénomènes émergents sur le terrain représente l'extrême opposé du travail scientifique des données en astronomie. Les réseaux de capteurs déployés par le CENS étaient des technologies largement expérimentales. Certains instruments étaient trop fragiles pour être laissés sans surveillance ; d'autres étaient déplacés fréquemment pour s'adapter aux conditions du terrain, selon un protocole de recherche reposant sur l'intervention humaine. La fonction fédératrice du CENS, qui réunissait les savoir-faire scientifiques et techniques nécessaires à la résolution de problématiques de recherche, faisait partie intégrante de son infrastructure de la connaissance. La fonction fédératrice consiste, entre autres, à fournir une expertise technique supplémentaire, de l'aide administrative et des espaces de collaboration. Pour le reste, les participantes et participants se reposaient sur les infrastructures de leurs domaines ou de leurs départements universitaires.

Malgré les investissements infrastructurels du CENS, les personnels de recherche manquaient de standards de données, d'archives et de mécanismes de classification pour faciliter et intégrer les échanges de ressources. La responsabilité de la gestion des données incombait aux chercheurs et chercheuses, sans guère d'appareil sur lequel s'appuyer. Il s'agit là, cependant, du problème de l'œuf et la poule. Les projets fondés sur les réseaux de capteurs tendent à être exploratoires, à aborder des problématiques émergentes et à être confrontés à des situations de terrain dynamiques par essence. Les équipes de recherche n'avaient pas nécessairement besoin de mettre les données de différents déploiements en correspondance, puisque chaque excursion visait généralement à répondre à de nouvelles questions avec de nouveaux instruments. Même si elles effectuaient des comparaisons entre plusieurs lieux et moments, elles n'étaient que rarement confrontées à des exigences d'intégration de données semblables à celles du relevé COMPLETE.

Ces contrastes correspondent parfaitement aux provocations exposées au chapitre 1. Les divergences en matière de travail scientifique des données de ces deux champs contribuent largement aux différences quant à qui y possède les données de la recherche, les contrôle, les pérennise ou y accède. Les réservoirs communs

de ressources dominant l'astronomie alors que les données privées resteront, dans les années à venir, la norme de la recherche sur les réseaux de capteurs. Une demande de réservoirs communs de ressources n'existe, pour ce qui est des domaines du CENS, que pour les champs spécialisés que sont la sismologie et la génétique.

Ces différences en matière de travail scientifique des données soulignent par ailleurs la divergence des méthodes d'échanges de données, quand il y en a. Les réservoirs communs de ressources, qu'ils consistent en des systèmes d'instrumentation ou d'information, reposent sur des normes communes et sur des institutions qui soutiennent l'interopérabilité, comme l'Astrophysics Data System, le CDS, SIMBAD et NED. Les objets célestes sont liés aux publications où ils sont mentionnés ; en revanche, l'établissement de liens entre articles et jeux de données est bien moins fréquent. Dans la recherche en réseaux de capteurs, les échanges de données reposent largement sur les contacts interpersonnels et il n'existe pas de méthode formelle pour lier des publications, des données et d'autres objets de recherche. Néanmoins, les deux domaines investissent des moyens humains considérables pour faire fonctionner ces infrastructures de la connaissance.

Les caractéristiques temporelles de ces domaines influencent aussi l'évolution de leurs infrastructures de la connaissance et les relations entre parties prenantes. L'astronomie est ainsi le plus ancien et le mieux établi des champs des sciences exactes.

Les infrastructures d'aujourd'hui transcendent le temps et l'espace, permettant d'intégrer des données anciennes à des observations contemporaines et futures. De nos jours, les astronomes dépendent bien moins du mécénat qu'autrefois, mais le soutien financier de riches donateurs et donatrices est toujours recherché pour les grands projets d'instrumentation. L'augmentation des fonds publics consacrés à l'astronomie a contribué à son internationalisation, aux investissements dans des réservoirs communs de ressources et à une plus grande équité d'accès aux instruments et aux données. Cette discipline est l'une des rares à parler d'une seule voix, à travers son relevé décennal.

À l'inverse, la science et la technologie des réseaux de capteurs sont un carrefour où des scientifiques en mal de technologies rencontrent des technologues en quête de domaines d'application intéressants. Les participantes et participants viennent de champs nouveaux comme anciens et gardent un pied dans chaque monde le temps de la collaboration. Le domaine est passionnant et a atteint une masse critique en matière de colloques et de revues, mais les perspectives d'embauches, de

postes et de promotions sont ailleurs. Il n'y a pas encore de motivation à bâtir des réservoirs communs de ressources ou une infrastructure de la connaissance plus complète pour la science et la technologie des réseaux de capteurs.

6. Le travail scientifique des données dans les sciences sociales

Introduction

Les sciences sociales recouvrent la recherche sur le passé, le présent et l'avenir du comportement humain. En dépit de leur longue histoire, ces champs sont attaqués dans les universités comme au-dehors. Le débat sur les « deux cultures » des sciences et des lettres lancé par l'essai de C. P. Snow (1956) a fait rage tout au long des années 1960, mais a largement laissé les sciences sociales à la marge de cette dichotomie. Malgré la montée en puissance des collaborations et des parcours interdisciplinaires, les divisions entre domaines se sont creusées avec la transformation des structures de financement et de rémunération (Hollinger, 2013). Dernièrement, la politologie est devenue la science sociale la plus politisée, au point que le Congrès américain a suspendu le financement public de la discipline sauf pour les travaux qui « promeuvent la sécurité nationale ou les intérêts économiques des États-Unis » (Prewitt, 2013). Les réactions, tant aux États-Unis qu'à l'étranger, expriment une profonde inquiétude quant à l'immixtion de la politique dans l'attribution de subventions individuelles, quelle que soit la discipline. L'évaluation par les pairs (*peer review*) et l'expertise scientifique sont en jeu et la pression pour des résultats immédiats s'accroît au détriment de la construction théorique à long terme (P. Boyle, 2013 ; Prewitt, 2013).

Certains des problèmes rencontrés par la recherche en sciences sociales sont attribuables aux difficultés dans la gestion des données. Par exemple, en tentant de répliquer une étude économique influente, un étudiant y a découvert des erreurs de calcul. Celles-ci – et la réponse que leur ont apportée les auteurs – ont fait la une des revues scientifiques comme de la presse économique, contribuant à interroger la fiabilité de la recherche (Marcus, 2013 ; Monaghan, 2013 ; Wiesenthal, 2013). Un cas de fraude par un professeur de psychologie sociale néerlandais est passé inaperçu des années durant, soulevant des inquiétudes semblables sur la recherche et l'efficacité de l'évaluation par les pairs dans cette discipline (Enserink, 2012a ; Shea, 2011). Le taux de rétractation en économie et dans les autres champs des sciences sociales est certes nettement moindre que dans les sciences exactes, mais cela n'indique pas pour autant une plus grande intégrité. Les différences seraient plutôt liées aux manières dont les maisons d'édition de chaque discipline gèrent les propositions comportant des erreurs, des plagats ou de la fraude (Karabag et Berggren, 2012 ; Oransky, 2012).

Les méthodes traditionnelles d'échantillonnage, comme le recours au courrier postal ou aux appels aléatoires sur des téléphones fixes, sont moins fiables depuis que les activités de communication se font en ligne. D'aucuns affirment que l'intérêt des sondages et des autres méthodes conventionnelles de recherche en sciences sociales touche à ses limites et que de nouvelles approches sont nécessaires (Savage et Burrows, 2007, 2009). Les politiques de protection des sujets de recherche sont elles aussi en pleine transformation. Un placard fermé à clé ne suffit plus à protéger les dossiers numériques recueillis sur des personnes ; les nouvelles méthodes doivent répondre à la réalité de la protection de la vie privée, du *data mining* et de la réidentification. La recherche sur des sujets sensibles tels que le terrorisme et les conflits est essentielle à la politique publique, mais la mener tout en préservant la confidentialité – et la vie – des sujets oblige à des choix difficiles (Jackson *et al.*, 2013). Étant donné la complexité du comportement humain et des institutions sociales, il est souvent très difficile d'établir une causalité dans les sciences sociales. Ces disciplines sont en quête de nouvelles méthodes et sources et, ce faisant, se confrontent aux promesses et aux écueils du travail scientifique des données.

Les méthodes de recherche et les pratiques en matière de données

Les pratiques en matière de données d'un domaine font partie intégrante de ses méthodes de recherche ; or, les sciences sociales explicitent bien plus leurs méthodes que les autres champs. Les manuels méthodologiques abondent pour guider les programmes universitaires sur le protocole de recherche, les statistiques, la recherche quantitative et qualitative et la visualisation. L'équilibre entre une description la plus riche possible du comportement humain et la nécessité de respecter les droits des individus, des groupes et des institutions étudiés est au cœur de ces méthodes.

Quelques dichotomies de base – quoique réductionnistes – suggèrent la palette des méthodes possibles. Ces dimensions ne sont pas incompatibles et peuvent se combiner de différentes manières. La première dichotomie met en regard l'explication idiographique et l'explication nomothétique. Les études idiographiques sont particulières à un lieu, à une condition ou à un événement. Cette démarche cherche à décrire et expliquer un cas donné de la manière la plus complète possible. À l'inverse, les études nomothétiques identifient des facteurs causaux qui influencent une classe d'événements ou de conditions (Babbie, 2013). La deuxième dichotomie distingue les méthodes reposant avant tout sur des techniques de décompte et de calcul pour étudier les problèmes sociaux, dites « quantitatives », de celles qui ont recours à l'interprétation, dites « qualitatives ». La troisième différencie les méthodes interventionnistes (*obtrusive*) et non interventionnistes (*unobtrusive*).

Les méthodes interventionnistes sont celles de travaux où il y a un certain degré d'interférence : le sujet de recherche est conscient d'être étudié, de préférence avec son consentement. Les méthodes non interventionnistes, quant à elles, ne supposent pas d'intervention : le chercheur ou la chercheuse travaille sur des traces d'activité humaine ou observe un comportement sans interférer.

Les notions de fiabilité et de validité transcendent ces dichotomies. La fiabilité correspond à la constance, à la probabilité que des observations répétées d'un même phénomène produisent le même résultat. La validité est la valeur de vérité, le degré auquel une mesure saisit le concept qu'elle est censée mesurer (Babbie, 2013 ; Shadish *et al.*, 2002).

Les scientifiques arbitreront différemment entre ces dimensions pour chaque étude. Les sondages, par exemple, sont généralement nomothétiques, quantitatifs et exigent un vaste échantillon pour être suffisamment fiables. À l'inverse, les ethnographies sont le plus souvent idiographiques, qualitatives, interventionnistes et s'inquiètent davantage de leur validité que de leur fiabilité. Les études ayant recours au *big data* nécessitent des méthodes statistiques et, parfois, des modélisations informatiques. Ces méthodes sont à même de produire des données anonymisables et réutilisables par d'autres. En revanche, les données qui reposent sur une analyse fouillée sont propres à engendrer des descriptions fécondes de phénomènes, mais les données peuvent s'avérer impossibles à anonymiser ou à partager. Le plus souvent, il faut faire des choix difficiles quant aux données pouvant être obtenues ou non, comment et pourquoi, ainsi que sur les façons de les rapporter et de les diffuser.

Études de cas en sciences sociales

Le travail scientifique des données est tout aussi divers dans les sciences sociales que dans les sciences exactes ; nous n'essayerons donc pas d'être exhaustifs. Les études de cas du présent chapitre abordent les raisons et les manières dont les gens utilisent les technologies de l'information, illustrant les dimensions des méthodes expliquées plus haut. La première série de cas, sur les enquêtes sur Internet et les médias sociaux comme c'est d'usage, compare des sondages, où les enquêtés évoquent leur usage d'Internet, avec des traces de leur comportement réel sur des technologies telles que Twitter. OxlS, petit nom de l'Oxford Internet Survey of Britain, est menée deux fois par an par l'Oxford Internet Institute depuis 2003 au moyen d'entretiens personnels. Plusieurs études qui utilisent les contenus de Twitter et d'autres services de microblogage comme ressources en données éclairent la manière dont les décisions relatives aux preuves et aux méthodes influent sur les pratiques en matière de données et les résultats.

La seconde étude de cas de ce chapitre s'intéresse à la manière dont les technologies de l'information sont conçues, mises en place et utilisées dans la recherche en science et en technologie. Les méthodes sociotechniques sont employées pour étudier les pratiques en matière de données du Center for Embedded Networked Sensing (Centre pour la télédétection intégrée en réseau). Les résultats de ces travaux sont rapportés dans l'étude de cas sur les pratiques en matière de données du CENS au chapitre 5. Dans le présent chapitre, nous résumons les méthodes employées dans une recherche sociotechnique sur dix ans et nous examinons leurs implications dans le travail scientifique des données. Ces méthodes sont largement idiographiques, mais comportent également une analyse de réseaux sociaux, ainsi que l'élaboration et l'évaluation de technologies. Ensemble, ces deux études de cas examinent la manière dont les spécialistes des sciences sociales abordent des méthodes et des problèmes nouveaux et les implications qui en découlent pour les infrastructures de la connaissance.

Les enquêtes sur Internet et l'étude des médias sociaux

Les études d'Internet tirent leurs méthodes d'autres domaines des sciences sociales. Les sondages sont une façon habituelle de poser les mêmes questions à un grand nombre de gens. Ils peuvent être menés en personne, par courrier, par e-mail ou en utilisant d'autres technologies comme des questionnaires sur le Web et des applications mobiles. L'analyse de réseaux sociaux, une méthode couramment employée dans la recherche sur Internet, est bien antérieure au Web et aux médias sociaux d'aujourd'hui comme Twitter, Facebook, LinkedIn, Flickr et Pinterest. Les sociologues modélisent des relations entre individus et groupes depuis les années 1920 environ, en ayant recours à tous les indicateurs disponibles : cartes postales, appels téléphoniques, listes de membres et autres liens sociaux (Freeman, 2004 ; Wellman et Haythornthwaite, 2002).

Pour toute méthode de recherche, il est nécessaire de disposer d'un savoir-faire scientifique pour concevoir l'étude, sélectionner la population à étudier et les méthodes d'échantillonnage et nettoyer, analyser et interpréter les données obtenues. Le danger de l'utilisation du *big data* dans les sciences sociales et ailleurs consiste à croire qu'on analysera les données aussi facilement qu'on les a obtenues. Or, mener des enquêtes et des études sur les médias sociaux est bien plus difficile qu'il n'y paraît. Certains travaux sur Internet sont très sophistiqués et prennent soigneusement en compte les limites de ces données en matière de fiabilité et de validité. D'autres sont naïfs et exploitent des flux de données comme d'intéressantes sources de preuves sans bien en comprendre les limites (Boyd et Crawford, 2012).

La taille compte

La mégascience ou science lourde (*big science*) au sens de Weinberg (1961) et Price (1963) se caractérise par la maturité du champ de recherche et par la sophistication des méthodes. La recherche par sondage, qui documente les tendances de la société depuis fort longtemps, est sans doute l'un des domaines les plus mûrs des sciences sociales. Des archives telles que l'Inter-University Consortium for Political and Social Research (2013) recueillent des données de sondage depuis plus de cinquante ans, originellement sur papier, puis sous forme numérique. Les archives de données des sciences sociales ont différentes priorités selon leur source de financement, le type d'étude, la région et d'autres critères. Les centres de recherche universitaires comme l'Institute for Quantitative Social Science à Harvard et la Science Data Archive à l'université de Californie à Los Angeles (UCLA) financent des référentiels de données, des didacticiels et instructions, le développement d'outils et d'autres services (Institute for Quantitative Social Science, 2013 ; Social Science Data Archive, 2014).

On peut effectuer des sondages de manière ponctuelle, mais ils s'avèrent particulièrement utiles quand ils sont menés sur de longues périodes. Des sondages d'opinion sur la politique, l'enseignement supérieur et les mentalités en général sont réalisés à intervalles réguliers depuis des décennies. En posant les mêmes questions clés à chaque fois, la comparabilité est maintenue. Les enquêtes sont adaptées pour répondre aux problématiques sociales du moment en ajoutant ou modifiant chaque fois quelques questions. Après qu'elles ont été déposées dans des archives, d'autres scientifiques peuvent en réanalyser les données, comparer les différentes études ou répliquer – entièrement ou partiellement – le protocole de recherche sur des populations différentes. Cependant, réutiliser ces données suppose un investissement considérable pour comprendre comment et pourquoi chaque enquête a été menée, en interpréter les résultats et déterminer quelles composantes peuvent être réutilisées ou comparées.

Les données des médias sociaux sont massives par leur volume absolu et peuvent produire dix fois, cent fois, mille fois, voire des millions de fois plus d'observations que les sondages, les entretiens ou les études en laboratoire. Suivre les communications quotidiennes de personnes au moyen de journaux intimes et d'autres méthodes manuelles, par exemple, ne fournit qu'un mince filet de traces. En revanche, s'il est difficile d'obtenir des chiffres fiables sur les transactions dans les médias sociaux, les volumes peuvent être colossaux. Ainsi, un des principaux fournisseurs de données de médias sociaux pour les entreprises affirme disposer d'un flux de trois milliards d'activités par jour (Gnip, 2013b).

Quand est-ce une donnée ?

Le déluge de données tiré de l'enregistrement numérique de l'activité humaine peut s'avérer une mine d'or pour les chercheurs et chercheuses en sciences sociales. La surabondance de sources et de ressources, à la fois contemporaines et historiques, amène toutefois son lot de difficultés. Découvrir ces richesses revient parfois à suivre une carte au trésor de pirate sur un chemin semé d'embûches, de chausse-trapes, de faux indices et d'indications trompeuses. La quête scientifique consiste à déterminer quelles entités peuvent valablement mettre en évidence quels phénomènes. Souvent, les chercheurs et chercheuses sont déchirés entre les données qu'ils veulent et les données qu'ils peuvent obtenir. Les meilleurs travaux de recherche sont ceux qui adaptent au mieux leurs méthodes aux questions posées. Des protocoles novateurs peuvent mener à des découvertes sensationnelles, mais s'avérer difficiles à répliquer ou à expliquer à celles et ceux qui n'appartiennent pas à la discipline. Des méthodes nouvelles peuvent aussi produire des données plus difficiles à documenter, partager, réutiliser et conserver comparées à celles issues de protocoles traditionnels.

Sources et ressources

Lorsque des scientifiques recueillent leurs propres données, elles et ils ont davantage de contrôle sur leurs protocoles de recherche. Pour mener des travaux à plus grande échelle, ils devront souvent rassembler des données tirées de ressources externes. Beaucoup combineront différentes méthodes, en réalisant des entretiens ou des ethnographies qu'ils compléteront de traces issues de médias sociaux, de transactions économiques, de recensement ou autres.

Les sondages sont généralement nomothétiques par nature : ils mesurent un petit nombre de variables de manière à pouvoir en tirer des comparaisons sur de vastes populations. Les scientifiques contrôlent le plus de sources de variances possible en indiquant précisément les questions à poser, la population à interroger et le plan de sondage. Elles et ils savent ce qu'ils demandent, à qui et ce qu'ils comptent apprendre de chaque question. Ils contrôlent également la manière dont les questions sont posées en formant les sondeurs et sondeuses à s'adresser aux participants de manière cohérente, à formuler précisément les questions et à noter les réponses. Le protocole de recherche doit équilibrer la validité interne et externe, c'est-à-dire respectivement le degré auquel l'étude contrôle les variables pour isoler les phénomènes et le degré auquel l'étude est généralisable à des populations plus vastes et diverses (Shadish *et al.*, 2002). Ces choix méthodologiques influent sur la capacité à se fier aux sources de données, à déterminer la provenance, à employer des outils d'analyse et à interpréter les résultats.

Les sondages en ligne sur le Web peuvent toucher des populations bien plus importantes que les entretiens personnels, mais obtenir un échantillonnage cohérent est difficile et les taux de réponse sont souvent bas. La fiabilité et la validité tendent à être plus élevées dans les enquêtes qui ont recours à des enquêteurs et enquêtrices humains pour s'adresser en personne aux participants, mais elles sont aussi bien plus coûteuses. Lorsque les chercheurs et chercheuses recueillent leurs propres données, ils peuvent parfois contrôler la variance en comparant leurs sources aux ressources d'autrui. Les instituts de sondage interrogent généralement 1 500 personnes environ, en privilégiant le téléphone plutôt que le face-à-face. Les enquêtes peuvent être étalonnées grâce à de telles sources, comme le sont les instruments de télescopes et de réseaux de capteurs.

Twitter a pris une place importante dans la recherche en technologie en raison de son usage international et de ses années d'existence, qui rendent possibles des analyses longitudinales. La plateforme permet d'envoyer de brefs messages – jusqu'à 140 caractères¹ – appelés « tweets ». Ces 140 signes peuvent contenir des informations supplémentaires grâce à des abréviations et des liens, lesquels peuvent être condensés grâce à des services de réduction d'URL. Les tweets peuvent également comporter des informations temporelles et des coordonnées géospatiales s'ils ont été envoyés depuis des appareils mobiles où ces fonctions sont activées. On peut y joindre des photos et d'autres images. Les comptes Twitter peuvent être nommés ou anonymes. Beaucoup appartiennent à des personnalités publiques, à des entreprises ou à des organisations grandes ou petites. Il n'y a pratiquement pas de limite à ce qu'un tweet peut dire : les particuliers évoquent leur journée, les scientifiques commentent les découvertes du moment, les entreprises présentent leurs nouveaux produits, les activistes motivent leurs troupes et les bibliothèques annoncent les modifications de leurs horaires et de leurs services. Les données de Twitter sont utilisées pour étudier les relations de communications, la santé publique, les événements politiques, les phénomènes linguistiques et le cours de la bourse, entre autres sujets (Bollen *et al.*, 2010 ; Bruns et Liang, 2012 ; Collins, 2011 ; Eysenbach, 2011 ; Murthy, 2011 ; Ozsoy, 2011 ; Shuai *et al.*, 2012 ; Simonite, 2013 ; Thelwall *et al.*, 2013 ; Zappavigna, 2011).

La variété des entités pouvant servir de données dans les flux Twitter et dans chaque tweet en fait une ressource attractive. Cependant, sa faiblesse réside dans la difficulté d'obtenir des données valides et fiables. À mesure que la valeur marchande des contenus de Twitter est apparue et que les préoccupations pour la vie privée ont émergé, les flux ont été rendus moins accessibles aux scientifiques. L'accès à l'ensemble du « flot » est limité, ce qui rend l'échantillonnage difficile. La distribution

1. NdT : 280 caractères depuis 2017.

des comptes Twitter et de leur activité par âge, sexe, race, pays, orientation politique, revenu et autres paramètres n'est pas homogène ; un échantillon d'utilisateurs ou utilisatrices ou de tweets ne constitue donc pas nécessairement une représentation fidèle de la population que le chercheur ou la chercheuse souhaite étudier. Cependant, la plus grande menace à la validité des tweets comme indicateurs d'activité sociale est l'évolution de l'usage des services en ligne. Une part croissante des comptes Twitter est constituée de robots servant à influencer la communication publique. Un compte peut ainsi attirer un grand nombre d'abonnés moyennant finance. Seulement 35 % des abonnés Twitter seraient de vraies personnes et jusqu'à 10 % de l'activité des médias sociaux proviendrait de comptes robotiques (Furnas et Gaffney, 2012 ; Urbina, 2013).

Les infrastructures de la connaissance

La diversité des ressources en données et de l'instrumentation dans les sciences sociales rend peu probable l'émergence d'une infrastructure de la connaissance comparable à celle de l'astronomie. Le comportement humain ne se prête pas à des descriptions normalisées comme le font les astres ou le spectre électromagnétique ; par ailleurs, l'échelle de coordination des télescopes spatiaux n'est pas transposable aux projets de ces disciplines. Les infrastructures de la connaissance de ces champs n'en comportent pas moins des outils et une expertise partagés. Une formation méthodologique, dispensée dans les cours de sciences sociales du second cycle, est essentielle au partage des méthodes et des données. Les cours de méthodologie abordent généralement l'utilisation d'outils analytiques, comme les progiciels statistiques, les logiciels de cartographie et les systèmes de codage qualitatifs.

Les archives des sciences sociales, que nous avons évoquées plus haut, comptent parmi les ressources les plus riches en données dont dispose la recherche par sondage. La collecte et la conservation de traces sociales sont une pratique bien antérieure à la recherche sur Internet (Boruch, 1985). Les registres de recensement, par exemple, prennent de la valeur avec les années, car ils constituent des observations sur un lieu et une époque. Le *Domesday Book*, le registre du recensement mené en Angleterre en 1085 et 1086 par Guillaume le Conquérant, est aujourd'hui considéré comme « le trésor le plus précieux de Grande-Bretagne » par les National Archives (2013). En 2011, le livre a été mis en ligne et doté d'une fonction de recherche des noms et des lieux mentionnés. Les registres de recensement contemporains, ainsi qu'une vaste gamme de traces sociales et institutionnelles, peuvent servir à la recherche.

Archiver les médias sociaux est un défi majeur, comme la bibliothèque du Congrès des États-Unis l'a appris à ses dépens. En 2010, elle a accepté de se charger de

la conservation de l'ensemble des archives de Twitter. En octobre 2012, environ cinq cents millions de tweets étaient envoyés quotidiennement. À l'heure où nous écrivons, la bibliothèque possède 133 téraoctets de données, qu'il faut deux jours pour interroger (Alabaster, 2013 ; Allen, 2013). Elle n'a pas encore réussi à rendre les données interrogeables ; les rendre interprétables est plus difficile encore. Les formes, les contenus et les usages des médias sociaux changent bien plus vite que les méthodes scientifiques pour les utiliser. Trouver un moyen d'incorporer ces nouveaux médias dans les infrastructures de la connaissance est une vraie gageure. Ils représentent des ressources précieuses, mais la question des moyens, du comment et du pour qui reste ouverte.

Métadonnées

La diversité des données et des questions de la recherche sur Internet rend difficile la mise en place des schémas de métadonnées communs, propres à permettre l'interopérabilité et l'échange de données. Pour les sondages, le standard de métadonnées le plus consensuel est la Data Documentation Initiative (DDI), qui est exprimée en XML. Comme mentionné au chapitre 2, la DDI a été largement adoptée par les archives de données des sciences sociales et s'applique à tout objet numérique que l'utilisateur ou l'utilisatrice considère comme une donnée (Data Documentation Initiative, 2012). L'un de ses objectifs est de faciliter la réutilisation de métadonnées (Vardigan *et al.*, 2008). La DDI est plus souple et moins prescriptive que les fichiers FITS utilisés en astronomie. Néanmoins, la DDI s'accompagne, comme la plupart des standards, d'une importante documentation. Les utilisateurs doivent investir du temps dans son apprentissage et son implémentation pour structurer et documenter leurs données. Une fois cette opération réalisée, les jeux de données sont prêts à être proposés à des archives ou échangés avec des collaborateurs et collaboratrices qui emploient aussi la DDI.

Parce qu'un sondage peut être très différent d'un autre, la DDI ne normalise pas les noms de variables, les guides de codification et les autres formes de documentation essentielles à l'analyse ou à l'interprétation des données. Dans les sciences sociales, les conventions de nommage et les pratiques documentaires sont souvent élaborées en interne. Même des conventions de base, comme la manière de codifier le sexe et l'âge dans une enquête, varient entre les projets de recherche et les jeux de données, y compris ceux déposés dans des archives. Un sondage peut codifier les hommes et les femmes comme 0 et 1 respectivement, un autre fera l'inverse et d'autres encore utiliseront 1 et 2 ou 2 et 1. De même, l'âge peut être codifié en années ou par l'année de naissance, qui sera notée avec deux ou quatre chiffres. Ainsi, une entrée « 45 » dans le champ « âge » peut signifier « le sujet a 45 ans » dans une étude et « le sujet est né en 1945 » dans une autre. Un jeu de

données n'a guère de valeur s'il lui manque des noms de variables, des guides de codification et d'autres formes de documentation.

Les données issues de médias sociaux sont particulièrement complexes, compte tenu de la variété des sources et la forme libre et dynamique des communications qui y sont effectuées. Les données tirées de services commerciaux tendent à être plus standardisées, surtout s'ils fournissent des API (Application Programming Interface, interface de programmation d'application). Les sources commerciales de données Twitter, par exemple, publient le format dans lequel les flux Twitter sont fournis (Gnip, 2013a). Ces formats sont eux-mêmes fondés sur des standards publiés d'échange de données, comme le JavaScript Object Notation (JavaScript Object Notation, 2013). Comme la DDI, ces formats ne sont qu'un point de départ pour les scientifiques qui sélectionnent des unités d'activité sur les médias sociaux pour s'en servir comme données.

Provenance

Établir la provenance dans les études sur Internet est difficile en raison de l'adaptabilité des méthodes, de la variété des ressources, des origines multiples des données et du manque de documentation sur les décisions prises à chaque étape du traitement. Ce sont généralement les enquêtes qui prévoyaient d'emblée une réutilisation qui présentent les meilleures informations de provenance. Le General Social Survey (GSS), qui fait partie de l'International Social Survey Programme (ISSP), est en cela exemplaire. Les données, les guides de codification et le reste de la documentation de ce sondage, qui est mené depuis 1972, sont à la disposition du public (General Social Survey, 2013 ; International Social Survey Programme, 2013). Pour les personnes formées à la recherche par sondage et à l'utilisation de ces données, repérer la formulation d'une question à une année considérée et la distribution des résultats est simple. Des usages plus complexes nécessiteront un examen plus poussé pour déterminer quelles questions – sur des sujets comme l'orientation politique – ont varié d'année en année, comment les variables ont été codifiées, comment elles ont été combinées dans des index, etc. Reportez-vous à la figure 6.1 tirée du General Social Survey pour en voir un exemple.

Malgré la documentation explicite de la provenance dans le General Social Survey, il faut beaucoup de savoir-faire et de travail pour utiliser ces données de manière fiable, comme le montrent les références croisées dans la légende du tableau. Les questions ont légèrement évolué de sondage en sondage, s'adaptant à leur époque. Les regroupements ont également changé, modifiant l'empreinte de chaque jeu de données, ce qui n'est pas sans rappeler les problèmes cartographiques du relevé COMPLETE, que nous avons vus au chapitre 5. Les choix

préliminaires de réduction et de nettoyage des données – que ce soit dans les sondages, les indicateurs économiques ou les observations astronomiques – peuvent influencer profondément sur la capacité à retracer la provenance ou à réinterpréter des jeux de données (Blocker et Meng, 2013).

56. Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?

[VAR: PARTYID]

RESPONSE	PUNCH	YEAR										COL 240	
		1972-82	1982B	1983-87	1987B	1988-91	1993-96	1998	2000	2002	2004	2006	ALL
Strong Democrat	0	2197	143	1271	151	884	1050	370	414	408	455	700	8023
Not very strong Democrat	1	3482	109	1655	89	1282	1542	597	507	515	504	736	11018
Independent, close to Democrat	2	1768	44	904	51	578	887	349	325	267	281	527	5981
Independent (Neither, No response)	3	1736	30	855	32	721	1031	477	568	528	471	997	7444
Independent, close to Republican	4	1106	8	743	9	571	698	244	261	199	239	327	4405
Not very strong Republican	5	2011	8	1259	15	1170	1318	484	399	449	425	637	8175
Strong Republican	6	1009	8	751	2	662	808	239	285	315	396	495	4970
Other party, refused to say	7	243	0	75	1	44	104	63	48	48	29	65	720
Don't know	8	10	0	0	0	0	0	0	0	0	0	0	10
No answer	9	64	4	29	3	15	64	9	12	36	12	26	274

REMARKS: See Appendix D: Recodes, for original question format and method of recoding. See Appendix N for changes across surveys. If planning to perform trend analysis with this variable, please consult GSS Methodological Report No. 56.

Figure 6.1. Tableau du General Social Survey comparant l’affiliation à un parti politique au fil du temps

Crédit : National Opinion Research Center.

La documentation de provenance est essentielle aux données de médias sociaux et, pourtant, très difficile à établir. L’origine d’un tweet, d’un billet de blog ou d’une autre communication peut constituer un indicateur de sa valeur, de sa fiabilité ou de sa validité. La provenance peut inclure du contexte sur les émetteurs, les destinataires, le contenu de la communication et la chaîne de relations associées à chacun de ces éléments. Certaines études se préoccupent des relations tandis que d’autres s’intéressent à la teneur des communications. Les tweets mentionnent fréquemment des ressources en ligne et fournissent des liens pour y accéder. Ces derniers se brisent vite et les ressources disparaissent, rendant la provenance des tweets difficile à retracer (Salaheldeen et Nelson, 2013).

Le souci de la provenance et de l’interopérabilité a conduit à la diffusion d’outils ouverts pour la recherche sur les médias sociaux (Social Media Research Foundation, 2013). Ces outils dépendent parfois eux-mêmes de standards techniques pour documenter les relations de provenance dans le Web sémantique (Groth et Moreau, 2013). Cependant, la plupart des médias sociaux ne sont pas encore conçus d’après les technologies du Web sémantique, aussi les chercheurs et chercheuses utilisant ces données ont-ils toujours besoin d’élaborer leurs propres relations de provenance

(Barbier *et al.*, 2013). Étant donné l'évolution rapide des médias sociaux, des outils, des méthodes et des savoir-faire, il est peu probable que nous pourrions à court terme déterminer la provenance à un niveau suffisant pour pouvoir comparer ou réutiliser des jeux de données.

Les influences extérieures

Parce que les sciences sociales étudient le comportement humain, elles subissent davantage de contraintes extérieures qu'aucun autre domaine. Bien des problèmes de vie privée, de confidentialité et de propriété sont insolubles. D'autres peuvent être réglés ingénieusement dans le cadre de ce qui peut être demandé ou observé, des coûts d'acquisition des données et de l'usage de biens détenus par autrui (Brady, 2004).

Économie et valeur

La valeur des données des sciences sociales en général et de la recherche sur Internet en particulier réside souvent dans leur « emballage ». Les enquêtes sociales peuvent constituer des réservoirs communs de ressources si elles se trouvent dans des archives ouvertes à la plupart des scientifiques, comme la UK Data Archive. D'autres peuvent être des biens de club lorsque les archives où elles sont déposées ne sont accessibles qu'aux membres d'un consortium, comme c'est le cas pour l'ICPSR (Inter-University Consortium for Political and Social Research, 2013 ; UK Data Archive, 2014). Elles peuvent constituer des biens privés si elles ne sont échangées que de manière interpersonnelle ou des biens publics si elles sont publiées ouvertement pour être utilisables par quiconque. Acquérir des données de sondage est souvent coûteux. Par exemple, envoyer des enquêteurs et enquêtrices formés sur le terrain suppose un investissement considérable en argent, en temps et en expertise. Réutiliser des sondages permet d'économiser et d'étendre la gamme des ressources disponibles pour la recherche comparative et longitudinale.

De même, l'économie des données de médias sociaux dépend des sources et de l'emballage. Les sociétés commerciales comme Facebook, Twitter et Google possèdent les ressources les plus importantes. Ces dernières ont une immense valeur marchande et sont vendues à des entreprises pour leurs analyses d'affaires. Pour acquérir des jeux de ces données, les scientifiques doivent souvent payer : ce sont donc des biens de club. Certaines données peuvent être obtenues en copiant des sites web et par d'autres moyens. L'activité sur des *chat rooms* et d'autres sources publiques peut être accessible gratuitement, mais imposer d'autres contraintes. En effet, indépendamment d'éventuels frais, les contenus de médias sociaux sont généralement soumis à des restrictions de licences sur ce qu'on peut en faire et sur le degré auxquels ils peuvent être diffusés à d'autres.

Les sondages, les médias sociaux et d'autres observations de l'activité sur Internet peuvent être analysés à l'aide de logiciels. Beaucoup de ces outils sont commerciaux, quelques-uns sont *open source*. Une fois que les données sont entrées dans l'outil, on peut leur appliquer des routines prédéfinies et des scripts sur mesure. Analyser des données de médias sociaux peut nécessiter beaucoup de programmation *ad hoc*. Ces outils produisent généralement des jeux de données. L'utilisation ultérieure de ces derniers dépend de l'accès à un logiciel analytique, voire à une version particulière de celui-ci, auquel s'ajoute l'accès aux programmes ou aux scripts sur mesure. Quelles que soient les conditions de licence qui s'appliquent à un jeu de données, l'accès aux outils peut constituer une contrainte et déterminer s'il peut être utilisé et comment.

Droits de propriété

Les droits se rapportant aux données et aux outils – logiciels, par exemple – sont particulièrement complexes dans le cadre de la recherche sur Internet. Lorsque les données sont disponibles, elles peuvent avoir été diffusées sans documentation spécifiant les logiciels associés et leurs versions. Les outils sur mesure ne sont pas forcément divulgués. Les logiciels nécessaires à la réutilisation et à l'interprétation de jeux de données peuvent être coûteux, incompatibles avec le matériel informatique et les systèmes d'exploitation utilisés, n'être plus disponibles ou n'être pas intégralement diffusés. Lorsque des scientifiques recueillent des informations sur des personnes, les droits sur les données peuvent appartenir aux chercheurs et chercheuses, aux sujets de recherche ou aux fournisseurs d'informations. La possibilité, pour des sociétés, de vendre des données sur leur clientèle et la capacité de cette dernière à contrôler les données qui la concernent dépendent des pays et des juridictions. Le pouvoir de publier des données et la responsabilité afférente varient d'autant. Ces considérations de propriété influent sur les choix des études sur Internet et sur bien d'autres sortes de recherches en sciences sociales. Les scientifiques préfèrent souvent recueillir leurs propres données afin d'éviter ces problèmes.

Éthique

Les considérations éthiques dans la recherche en sciences sociales, y compris dans la recherche sur Internet, concernent le respect des personnes étudiées et leurs droits, que ce soit individuellement ou en tant que groupe. Les données personnelles sont soumises à de nombreuses réglementations quant aux données pouvant être recueillies et aux conditions de collecte. Aux États-Unis, les principes éthiques fondamentaux sur le traitement des sujets humains sont codifiés dans le rapport Belmont pour ce qui est du respect des personnes, de la bienfaisance et de la justice (US Department of Health and Human Services, 1979). L'application de ces règles varie selon les disciplines, les organismes de

financement, les juridictions et d'autres facteurs. La plupart des pays disposent de principes similaires pour la recherche sur des sujets humains.

Les données sur les sujets humains donnent lieu à un imbroglio de conflits. D'un côté, nous avons une prolifération d'informations personnelles divulguées par les internautes sur les médias sociaux et à travers leurs autres transactions sociales et économiques. Ces informations suscitent des débats sur la vie privée et la propriété des données. De l'autre côté, nous avons des principes établis pour protéger l'identité de sujets de recherche en raison du risque physique, psychologique ou économique que la divulgation leur ferait encourir. Entre les deux se trouvent des interrogations sur les règles autour du consentement éclairé qui seraient trop rigides et empêcheraient le partage de données susceptibles de bénéficier aux individus concernés. Les données biomédicales sont en première ligne de ces conflits. Certaines associations de patientes et patients souhaitent des contrôles plus étroits pour empêcher la ré-identification, alors que d'autres veulent octroyer aux malades le droit de consentir au partage de données les concernant (Field *et al.*, 2009).

Les règles éthiques sur les sujets humains sont plus claires en ce qui concerne les méthodes traditionnelles, telles que les sondages. Les chercheurs et chercheuses déterminent quelles questions poser, de quelle manière et comment rapporter les données de façon à maintenir la confidentialité. La diffusion des données tirées de sondages dépend souvent de la capacité à les codifier de manière à préserver l'anonymat des sujets. Les archives de données disposent de procédures et de politiques pour garantir que les données sont fournies et publiées sous des formes qui respectent la confidentialité. Cependant, à mesure que le volume de données en ligne sur les sujets humains augmente, la capacité à identifier des personnes en combinant des jeux de données progresse. Il est parfois nécessaire d'ajouter des niveaux de contrôle, par exemple en ne donnant accès qu'aux métadonnées, et ce uniquement aux chercheurs certifiés. Il y a de plus en plus de données personnelles identifiables recueillies et conservées sous forme numérique et les risques changent en conséquence, de même que les moyens de les écarter (National Research Council, 2013).

Selon la plupart des directives en matière de sujets humains, les données publiques et les comportements adoptés en public, comme le fait de tweeter, de bloguer et de publier, ne sont pas librement utilisables par la recherche. L'Association of Internet Researchers occupe un rôle majeur dans ce domaine en établissant des principes pour la collecte de données issues d'activités en ligne (Association of Internet Researchers, 2012). En utilisant les services des médias sociaux, les personnes ne consentent pas nécessairement à devenir des sujets de recherche. Bien qu'il soit

difficile d'obtenir un consentement éclairé pour de telles études, d'autres moyens existent pour protéger les droits et la vie privée des sujets. Les politiques en matière de vie privée des sociétés de médias sociaux sont soumises à une surveillance étroite, ce qui a réduit d'autant plus l'accès aux données de Twitter, Facebook et autres sources (Bruns et Liang, 2012 ; Schroeder, 2014).

Une fois les résultats publiés, des questions peuvent se poser autour de l'usage éthique des données. Les scientifiques s'inquiètent de la façon dont leurs découvertes seront utilisées, potentiellement à mauvais escient. Ces questions se compliquent lorsque l'exactitude des articles est remise en cause, qu'ils soient rétractés plus tard ou non. L'erreur de tableur dans l'article économique de Reinhart-Rogoff n'a été décelée que lorsqu'un tiers a tenté d'en répliquer les résultats ; le coauteur et la coautrice ont alors dû déployer des efforts considérables pour retrouver l'erreur et l'étape du traitement où elle était apparue. D'autres questions ont été soulevées autour de la sélection des données, de leur poids et de l'utilisation de tableaux plutôt que d'un progiciel d'analyse statistique (Marcus, 2013 ; Wiesenthal, 2013). Un auteur influent a d'ailleurs réprimandé les décideurs et décideuses politiques pour s'être fondés sur un unique indicateur statistique et les encourage à examiner une vaste gamme de données, de méthodes et de théories avant de prendre des décisions importantes (Summers, 2013).

Mener des enquêtes sur Internet et des recherches sur les médias sociaux

Des individus et des équipes restreintes peuvent réaliser des entretiens, une expérience, un sondage en ligne ou une analyse de médias sociaux à petite échelle. En revanche, des enquêtes de l'ampleur d'OxIS supposent des ressources humaines et un capital financier substantiel. L'étude des médias sociaux nécessite un cadre théorique et méthodologique, une expertise computationnelle et statistique, la disponibilité de ressources informatiques et l'accès à des données.

L'enquête Oxford Internet Survey (OxIS) a été conduite pour la première fois en 2003 dans le cadre d'une étude internationale sur l'usage d'Internet (World Internet Project, 2013). Sur la base de cette réussite, OxIS a été menée de nouveau en 2005, 2007, 2009, 2011 et 2013 (Dutton *et al.*, 2013 ; Dutton *et al.*, 2005 ; Dutton *et al.*, 2009 ; Dutton et Blank, 2011). Le responsable de la recherche d'OxIS a assuré la continuité du projet pendant les six éditions de l'enquête, tandis que le personnel changeait après un ou deux cycles. OxIS a ainsi accumulé plus d'une décennie d'observations comparatives sur qui utilise Internet en Grande-Bretagne, comment, avec quels appareils et en quoi l'usage d'Internet est comparable à celui d'autres

médias. L'enquête pose également des questions sur les normes sociales, comme l'intérêt pour la politique ou l'opinion sur le gouvernement, qui permettent des comparaisons avec d'autres études.

En dépit de leur ressource commune, les études qui exploitent les contenus de Twitter diffèrent tellement en matière de théories et de méthodologies qu'il est difficile de les comparer. Les choix de phénomènes à étudier conduisent à sélectionner différentes entités. Par conséquent, un même contenu peut engendrer de nombreuses sortes de données. Meyer, Schroeder et Taylor (Meyer *et al.*, 2013), par exemple, ont comparé la théorie, la méthode et les résultats de trois études de Twitter où des informaticiennes et informaticiens évoquaient des questions sociales. L'une de ces études demandait si Twitter était utilisé plutôt comme un moyen de communication ou comme un réseau de relations, avec 1,47 milliard de relations Twitter (Kwak *et al.*, 2010) ; deux d'entre elles s'interrogeaient sur qui influence qui (Bakshy *et al.*, 2011 ; Cha *et al.*, 2010), s'appuyant respectivement sur 1,6 million d'utilisateurs et utilisatrices et sur 1,7 milliard de tweets de 54 millions d'utilisateurs. Ralph Schroeder (2014), dans un examen ultérieur des aspects sociaux de l'article de Kwak (*et al.*, 2010), remarque que l'envergure des données Twitter dont ils disposaient n'est plus accessible aux scientifiques aujourd'hui.

Les questions de recherche

Dans les enquêtes sur Internet comme dans les médias sociaux, il est difficile de séparer les questions de recherche de la théorie et de la méthodologie. Les études sont conçues pour saisir des informations sur des phénomènes particuliers, dans des populations particulières, à des moments – et parfois dans des endroits – spécifiques, qui seront mesurées d'une manière conforme aux théories et aux hypothèses.

Les sondages OxIS portent sur un large périmètre, à savoir l'usage d'Internet au sein d'un pays sur une longue période, mais sont en partie conçus pour être comparés à des questions semblables posées par des institutions partenaires dans le cadre du World Internet Project. Les chercheurs et chercheuses de l'Oxford Internet Institute (OII) élaborent les enquêtes avec une grande minutie, travaillant soigneusement chaque item du questionnaire, leur enchaînement, le plan de sondage et les consignes aux sondeurs et sondeuses. Afin d'assurer la continuité, une série de questions clés est répétée à chaque étude et parfois affinée pour s'adapter à l'évolution technologique. Les choix d'analyse de données font partie intégrante du protocole de recherche, déterminant à l'avance la codification de chaque item pour atteindre la plus grande puissance statistique. Certains items du sondage fournissent des données démographiques et descriptives sur la population des

internautes ; d'autres sont conçus pour explorer des questions théoriques sur qui utilise Internet, quand, comment et pourquoi.

Les manières dont les questions de recherche façonnent le choix des preuves sont particulièrement notables dans l'étude des médias sociaux. Par exemple, Kwak et ses collaborateurs (2010) s'intéressent à la topologie du réseau. Pour savoir si Twitter représente un réseau social ou un média d'actualité, elles et ils ont comparé la distribution abonnements/abonnés de l'ensemble du site à la mi-2009 à des schémas d'activités dans d'autres réseaux humains. En revanche, une étude très différente est celle de la censure dans les réseaux de microblogage chinois (King *et al.*, 2013). Ces chercheurs et chercheuses, bien connus dans les sciences sociales quantitatives, ont combiné analyse de réseau et entretiens.

La collecte de données

La part des ressources et du travail de recherche consacré à la collecte de données varie aussi considérablement. Dans les enquêtes, le travail de terrain pour mener des entretiens peut prendre des semaines, voire des mois, suivi du même temps d'analyse des données. Chacun des six sondages OxlS cite des réponses issues d'entretiens avec plus de deux mille individus (Dutton *et al.*, 2013). Bien qu'ils soient préparés et analysés par des scientifiques de l'Oxford Internet Institute, les entretiens sont délégués à une société spécialisée dans la recherche de terrain. L'OII reçoit alors des fichiers de données anonymisées codifiés selon ses spécifications.

Dans les études sur les médias sociaux, déterminer quelles données sont souhaitées, lesquelles peuvent être collectées et comment les recueillir peut s'avérer plus laborieux qu'obtenir le jeu de données. En effet, la collecte peut se faire en quelques heures ou jours à l'aide d'un algorithme. Elle est précédée par le développement et le test des algorithmes et peut être suivie de longues périodes d'analyse. Les données Twitter présentent, entre autres attraits, une structure relativement simple, un format standard et un vaste champ d'applications, malgré le peu d'informations transmises dans chaque tweet. Si l'on utilise des API pour collecter les données, les traces sont enregistrées avec l'heure, la date, le lieu et d'autres variables en syntaxe normée, qui peuvent être traités en métadonnées. Notons cependant qu'apprendre à utiliser l'API de Twitter représente un investissement non négligeable, à en juger par l'existence de manuels d'utilisation de plus de quatre cents pages (Makice, 2009).

L'étude de Kwak (*et al.*, 2010) a utilisé l'API de Twitter pour collecter les profils de tous les utilisateurs et utilisatrices pendant trois semaines. Au cours des deux mois suivants, elles et ils ont relevé les profils d'utilisateurs évoquant des sujets en vogue. Une série d'informations annexes sur ces thèmes et les tweets associés

ont également été recueillies. De son côté, l'étude de King (*et al.*, 2013) utilise une méthodologie très différente pour l'acquisition de ses données, obtenant 11,3 millions de publications de plus de 1 400 plateformes de microblogage et autres médias sociaux en Chine. Elles et ils ont conçu leurs propres algorithmes pour recueillir et codifier les données parce que leur but était d'enregistrer les publications rapidement avant que la censure ne les repère et ne les supprime.

L'analyse des données

Les méthodes quantitatives produisent des données complexes et difficiles à nettoyer, étalonner, interpréter et conserver. Un grand nombre de petites décisions sont effectuées lors de leur traitement et de leur analyse. La capacité à réutiliser des données peut absolument dépendre de la correction des erreurs et des anomalies, de la cohérence des décisions prises, de leur adéquation avec les buts de l'étude et de leur documentation claire et complète.

L'OII reçoit les jeux de données des entretiens sous forme de fichiers dans STATA, un progiciel d'analyse de données et de statistique largement utilisé dans les sciences sociales (STATA, 2013). Selon les consignes transmises par l'OII à l'entreprise chargée de conduire les entretiens en son nom, chaque réponse de chaque question est codifiée avec des noms de variables spécifiques dans le jeu de données. Les analystes de l'OII traitent ces noms comme des métadonnées pour analyser les données. La première passe sert à nettoyer le jeu de données, à repérer d'éventuelles anomalies et erreurs et, si nécessaire, à poser des questions à l'institut de sondage. Dans l'enquête de 2011, le personnel de l'OII a échangé à trois reprises avec l'institut avant de publier son rapport préliminaire. Afin d'analyser les données de manière plus poussée pour des articles de revue ultérieurs, le personnel a encore contacté l'entreprise plusieurs fois avec des questions plus précises.

La plupart des anomalies ne pouvaient être repérées que par des personnes dotées d'une connaissance intime des relations et des tendances attendues dans les données. Des relations surprenantes pouvaient indiquer des découvertes importantes ou bien être le fruit d'une codification incohérente. Ainsi, en voyant des jeunes déclarer des comportements plus typiques des personnes âgées et vice versa, les chercheurs et chercheuses de l'OII se sont interrogés sur la codification du champ de l'âge. Il se trouve que, bien que l'OII ait spécifié aux sondeurs et sondeuses de demander la date de naissance plutôt que l'âge des personnes interrogées, un examen attentif a révélé que certains avaient demandé aux enquêtés leur âge et l'avaient inscrit tel quel. Certaines personnes de 22 ans étaient donc codifiées par un 22, ce qui, dans le fichier STATA, voulait dire qu'elles étaient nées en 1922.

Kwak et ses collaborateurs (2010) ont formulé leur problématique de recherche en termes de topologie de réseau et utilisé les données de médias sociaux comme un banc d'essai. En toute logique, leur collecte et leur analyse de données se sont concentrées sur les liens relationnels. Elles et ils ont recueilli le plus d'informations contextuelles possible sur chaque tweet. Ces scientifiques ont expérimenté différentes techniques pour repérer les spams et plusieurs seuils pour les considérer comme de mauvaises données. Au bout du compte, ils ont supprimé plus de vingt millions de tweets et près de deux millions de comptes utilisateurs de leur jeu de données initial. Les flux d'informations étaient mesurés par le nombre d'abonnés de chaque utilisateur ou utilisatrice et par le degré de réciprocité de ces relations. L'équipe de recherche en a conclu que Twitter se rapprochait davantage d'un mécanisme de diffusion d'informations que d'un outil pour bâtir et entretenir un réseau de relations sociales.

L'étude de King (*et al.*, 2013) est, quant à elle, fondée sur des théories de la censure et de l'action gouvernementale. Les explications sur la théorie, la collecte de données, le codage et l'analyse y sont étroitement mêlées et occupent près de la moitié de l'article. Chaque décision de codage est minutieusement expliquée au regard des théories sociales. Certaines publications ont été codifiées à la main pour interpréter la teneur et l'intention du message. L'article se termine par une explication des implications sociales et politiques des schémas de censure identifiés.

La publication des résultats

Les résultats de ces études sur Internet sont publiés dans divers médias en fonction de leurs différences en matière de théorie, de méthode et de lectorat. Les rapports de chaque enquête OxIS sont diffusés à un très large public et notamment aux décideurs et décideuses politiques lors d'événements. Le rapport d'ensemble, disponible en ligne et sur papier, comprend des données récapitulatives sur chaque variable importante, des tableaux croisés et un résumé des méthodes de collecte. Des détails méthodologiques supplémentaires ainsi que l'instrument d'entrevue sont publiés en ligne. Une fois le rapport divulgué, les analystes d'OxIS explorent plus profondément les données de l'enquête aux fins de la recherche théorique et politique. Les articles de revue sont un endroit plus adéquat pour aborder les détails de la méthodologie et effectuer des comparaisons avec d'autres études (Di Gennaro et Dutton, 2007 ; Dutton et Shepherd, 2006). On peut les trouver dans les revues et les actes de recherche en communication, de science politique, de sociologie et de domaines proches.

Les études sur les médias sociaux sont rapportées dans tous les domaines, des sciences sociales à l'informatique en passant par les sciences exactes, la médecine,

les sciences humaines et la presse grand public. Les deux études évoquées ici (King *et al.*, 2013 ; Kwak *et al.*, 2010) ont été publiées dans des médias reconnus dans leurs disciplines respectives et toutes deux sont largement citées. Cependant, les différences théoriques et méthodologiques sont telles qu'effectuer des comparaisons entre travaux pourrait être vain – ou bien mener à des découvertes sensationnelles. Des études comme celles-là puisent leurs données dans des ressources communes, mais les entités choisies deviennent des preuves très différentes de phénomènes distincts. Ces domaines savants se recoupent de façon nouvelle. Néanmoins, il est rare qu'un politologue ou une informaticienne recherchent des éléments théoriques ou méthodologiques dans la littérature de l'autre.

La conservation, le partage et la réutilisation des données

Le cas des études sur Internet illustre les compromis entre méthodes conventionnelles et novatrices. L'innovation peut engendrer des découvertes importantes, mais entraîner des difficultés dans la documentation, le partage et la réutilisation des données. Les pratiques de conservation, de partage et de réutilisation semblent varier d'un projet à l'autre. La taille et la longueur des études, le renouvellement du personnel, les exigences externes de normalisation ou de dépôt et les attentes liées à la réutilisation sont autant de facteurs qui influent sur les pratiques en matière de données.

Les enquêtes OxIS sont conçues pour consigner l'usage d'Internet en Grande-Bretagne de manière à se comparer à des travaux analogues dans d'autres pays. Bien que beaucoup des questions clés soient les mêmes, le plan de sondage, les questions et l'interprétation des réponses différeront significativement. Alors qu'OxIS couvre des zones urbaines et rurales de la Grande-Bretagne, les études de pays en développement sont largement urbaines ; les pays riches disposent de plus de connexion haut débit alors que les pays pauvres utilisent essentiellement les réseaux mobiles, et ainsi de suite.

Les attentes en matière de réutilisation influent sur la capacité à interpréter les données ultérieurement. La première édition d'OxIS, en 2003, a été lancée rapidement en raison du financement et des collaborateurs et collaboratrices disponibles. Bien que le projet ait dès l'origine été pensé comme une étude continue, le personnel de recherche n'a d'abord constitué qu'une documentation minimale sur la collecte, la manipulation, la codification et l'analyse des données. Le questionnaire et le jeu de données de l'enquête de 2003 ont servi de base à l'étude OxIS de 2005. Au fil des enquêtes, le personnel a élaboré des procédures de documentation de plus en plus détaillées et complexes pour permettre les comparaisons longitudinales. Les jeux de données cumulés ont gagné en valeur avec chaque nouvelle série d'observations.

L'OII diffuse les données d'OxIS et documente ses jeux de manière très complète pour leur réutilisation. Il s'agit cependant de deux procédures distinctes. Comme les organismes de financement n'exigent pas de l'OII qu'il dépose ses données dans un référentiel, celui-ci diffuse ses jeux directement auprès de scientifiques qualifiés à des fins non commerciales après une période d'embargo de deux ans. Par exemple, c'est au moment de la publication du rapport de l'édition 2011 que le jeu de données de 2009 a été divulgué. Les conditions d'accès et de licence sont expliquées sur son site web. Le jeu de données mis à disposition correspond à celui reçu de l'entreprise qui conduit les entretiens, sous forme de fichier STATA. La documentation en ligne comprend les noms des variables et les questions dont celles-ci sont dérivées, ainsi que la méthodologie de base et les statistiques descriptives publiées dans le rapport d'enquête. Bien que le personnel d'OxIS divulgue volontiers, sur demande, des informations supplémentaires sur les jeux de données, il a reçu peu de requêtes des deux cents parties environ qui ont fait l'acquisition du jeu. Les réutilisateurs et réutilisatrices peuvent opérer leurs propres transformations et codifications sur le jeu de données STATA « brut ». L'équipe d'OxIS connaît la DDI, mais ne l'a pas mise en œuvre : son utilisation du jeu de données ne justifierait pas les frais induits et aucun facteur externe ne lui impose d'implémenter cette norme.

Les résultats rapportés par l'OII sont consignés selon les standards des différents médias de publication. La documentation rédigée en interne pour maintenir la fiabilité, la validité et la continuité au fil des enquêtes est considérée comme un contenu exclusif à son propre usage, ce qui n'est pas sans rappeler la façon dont on traite beaucoup de logiciels développés localement. Les informations internes comportent des détails sur les choix de transformation des variables, de combinaison des variables en index, de vérification des items et d'autres opérations de nettoyage et d'analyse effectués par les chercheurs et chercheuses. Cette documentation interne favorise la continuité et fournit à l'équipe un avantage concurrentiel.

La spécificité et la temporalité des études sur les médias sociaux constituent leur force, mais aussi leur faiblesse pour ce qui est de la réutilisation et de la répliquabilité. La nature temporelle des données de médias sociaux, à laquelle s'ajoute la dégradation rapide des sources auxquelles elles font référence, limite leur intérêt pour une réutilisation. En théorie, les études sur les médias sociaux pourraient être répliquées à intervalles réguliers. Néanmoins, les méthodes de collecte et d'analyse des données présentent une continuité moindre que la démarche des sondages, car elles reposent sur des outils qui évoluent rapidement. À mesure que ces médias deviennent des formes de communication courantes, leurs caractéristiques se transforment. L'apparition de robots sociaux, de messages qui s'autodétruisent et d'autres innovations rend les comparaisons difficiles. Faute d'indicateurs stables, valider et

étalonner ces méthodes pose problème. L'article d'informatique (Kwak *et al.*, 2010) et celui de science politique (King *et al.*, 2013) se concluent tous deux par des remarques comme quoi leurs résultats pourraient être largement appliqués à d'autres domaines. Ils n'explicitent cependant pas si leurs données sont disponibles, ni où ni comment. Que les jeux de données et les algorithmes puissent être réutilisés ou les conditions répliquées ou non, chacune de ces études suscite des questions de recherche qui peuvent être explorées dans différents domaines.

La sociotechnique

La sociotechnique, l'objet du second cas du présent chapitre, s'intéresse aux problèmes qui sont en partie sociaux et en partie techniques. L'étude de cas développée ici – qui correspond à une recherche en cours depuis 2002 sur les pratiques en matière de données et les infrastructures de la connaissance au CENS – combine des approches idiographique et nomothétique et applique un vaste éventail de méthodes de recherche.

Les méthodes idiographiques applicables à la sociotechnique comprennent les ethnographies, les entretiens, les récits oraux, les observations formelles et informelles et les analyses de traces d'activité humaine. Les technologies peuvent être étudiées du point de vue de l'interaction personne-machine, que ce soit pour explorer la cognition ou des phénomènes associés à des contextes spécifiques. Elles peuvent aussi être observées en tant qu'interventions dans les pratiques professionnelles, comme dans le domaine du travail coopératif assisté par ordinateur. Les projets dans l'un ou l'autre de ces champs peuvent être modestes, locaux et courts ou bien vastes, globaux et à long terme.

Les explications idiographiques sont unies par la place qu'elles accordent à l'examen minutieux et à l'interprétation. Les démarches interprétativistes font le lien entre les sciences sociales et les sciences humaines, qui explorent et comparent de multiples points de vue. Bien que ces méthodes aient une longue histoire, elles font toujours l'objet de polémiques sur des questions d'épistémologie, de critères de preuves et de philosophies du savoir. Nous prenons ici acte de ces débats, sans tenter de les traiter ou de les trancher (Garfinkel, 1967 ; Geertz, 1973 ; Glaser et Strauss, 1967 ; Latour et Woolgar, 1986 ; Latour, 1987 ; Lévi-Strauss, 1966 ; Lofland *et al.*, 2006 ; Roth et Mehta, 2002).

La taille compte

La plupart des travaux sociotechniques correspondent à de la science légère, puisqu'ils s'intéressent à des problématiques émergentes et confuses. Ils sont menés par de petites équipes de recherche ou même par un chercheur ou une chercheuse isolés, selon l'ampleur du projet. Il faut parfois des mois, voire des années de labeur pour transcrire, compiler, codifier, analyser et interpréter des données qualitatives. Les réseaux sociaux des communautés de recherche peuvent, grâce à l'analyse de publications ou d'autres indicateurs sociométriques, fournir des comparaisons quantitatives. De même, la fouille de texte (*text mining*) dans les publications peut révéler des récurrences dans les objets de recherche et leur évolution. Le but est généralement de combiner plusieurs méthodes. En appliquant des démarches présentant divers degrés de fiabilité, de validité interne et externe ainsi que différentes échelles, on peut entériner les résultats par triangulation.

Quand est-ce une donnée ?

Les travaux du CENS rapportés dans la présente étude de cas se fondent sur une littérature sociotechnique sur les pratiques de terrain, notamment dans les sciences de l'environnement, de plus en plus foisonnantes. Il s'agit essentiellement d'études qualitatives ayant recours à une combinaison d'ethnographie, d'entretiens, d'observation participante et d'analyse de documents (Aronova *et al.*, 2010 ; Cragin *et al.*, 2010 ; Cragin et Shankar, 2006 ; Jackson et Buyuktur, 2014 ; Jackson *et al.*, 2011 ; Karasti *et al.*, 2006 ; Olson *et al.*, 2008 ; Zimmerman, 2007). Comme pour la recherche sur Internet, les scientifiques peuvent puiser dans une pléthore de ressources en données. Les spécialistes de la sociotechnique doivent se montrer ingénieuses et ingénieux pour identifier les sources de données et évaluer leur crédibilité, leur indépendance et leurs relations avec d'autres formes de données.

Sources et ressources

En dépit du déluge de données dont elles disposent sur la pratique scientifique, les études sociotechniques sur les pratiques en matière de données tendent à se faire *in situ*. C'est pourquoi les chercheurs et chercheuses sont davantage susceptibles de recueillir leurs propres observations que de se fonder sur des sources externes. Néanmoins, les traces produites par les sujets de recherche, dont leurs publications, peuvent constituer des sources auxiliaires. Les catégories de sources et de ressources que nous évoquons ci-après sont courantes dans la recherche sur les pratiques de terrain citées plus haut, mais ne représentent pas une liste exhaustive de toutes les sources de preuves potentielles.

Observations de terrain et ethnographie

La recherche sociotechnique prend place sur un site de recherche, qui n'est pas nécessairement un lieu physique unique. On appelle ethnographies multisites des travaux qui effectuent des comparaisons entre différents sites (Marcus, 1995). Les observations peuvent également s'effectuer en ligne, grâce aux méthodes de « l'ethnographie virtuelle » (Hine, 2000). Les scientifiques peuvent observer les activités qui les intéressent de loin et mener des entretiens à l'aide d'une communication vidéo ou audio. Si elles augmentent l'échelle et la distance du projet de recherche, ces démarches renoncent néanmoins à un certain nombre d'informations contextuelles.

Toute recherche qui étudie des individus ou des communautés *in situ* nécessite leur permission. C'est ce qu'on appelle « l'entrée ». Une fois arrivé sur le site de recherche, y rester peut s'avérer tout aussi difficile. Dans les sciences sociales, le but de l'observation sur le terrain est d'étudier des phénomènes sans perturber l'environnement plus qu'absolument nécessaire. Une fois sur place, les chercheurs et chercheuses se présentent et expliquent la raison de leur présence plus en détail, puis s'efforcent d'être discrets et de perturber le moins possible le milieu. Ils cherchent ainsi à éviter « l'effet Hawthorne », qui tire son nom d'une célèbre étude sur une centrale de Western Electric où la productivité a augmenté à la suite d'expérimentations sur la luminosité, avant de décliner après la fin de l'étude. Le simple fait d'être observé modifie le comportement des personnes (Landsberger, 1958).

Une manière de minimiser les risques pour la validité des observations de terrain est de rester suffisamment longtemps pour que la présence du chercheur ou de la chercheuse fasse partie de l'environnement quotidien. Ce processus peut prendre des semaines, des mois ou des années. Les anthropologues peuvent ainsi passer une carrière entière à étudier une communauté particulière, qu'il s'agisse de physiiciens ou de gangsters. Une manière de se fondre dans le décor consiste à participer activement et ainsi à en apprendre plus sur le site. Par exemple, s'ils étudient les pratiques scientifiques en matière de données, les chercheurs peuvent contribuer à la collecte ou à l'analyse des données de l'équipe, se charger des commissions pour les fournitures et le matériel ou aider à des tâches physiques comme installer des batteries sur le terrain. La participation aux activités étudiées comporte cependant ses propres risques. Les chercheurs peuvent craindre de se montrer trop subjectifs, de « se convertir » et de parler au nom de leurs sujets de recherche plutôt que de parler d'eux.

Les chercheurs et chercheuses appliquent de nombreuses techniques pour rassembler des observations et les transformer en données. Sur le terrain, il vaut souvent mieux participer d'abord et prendre des notes plus tard. Selon les circonstances, les

chercheurs prennent des notes, enregistrent de l'audio ou des vidéos ou prennent des photos, avec la permission des personnes observées. Les enregistrements sonores peuvent être retranscrits, mais cette tâche est longue et coûteuse. Les enregistrements eux-mêmes restent précieux en raison des nuances de la voix et des indices visuels qu'on y trouve. Les notes, les transcriptions, les entretiens et les autres documents seront codifiés selon les thématiques, les événements et d'autres indicateurs des phénomènes étudiés. La codification peut être simple et manuelle ou complexe et assistée par des technologies. Ainsi, des outils d'analyse commerciaux et *open source* existent pour encoder des événements, des personnes, des thématiques et d'autres catégories. Une fois les documents codifiés, ces outils simplifient l'agrégation et la visualisation.

Entretiens

Dans la recherche sociotechnique idiographique, les entretiens consistent généralement en des questions ouvertes. Une entrevue de cinq à dix questions peut susciter une conversation d'une heure ou plus. Bien que les questions soient générales, elles peuvent produire des explications des activités : par exemple, comment les participantes et participants sélectionnent les terrains ou les technologies, ou quelles difficultés ils rencontrent dans la conservation de leurs données. En posant les mêmes questions à plusieurs personnes, les scientifiques peuvent comparer les pratiques d'individus, d'équipes, de domaines, de terrains et autres.

Dans le cadre d'une recherche qualitative, un échantillonnage aléatoire est rarement réalisable. Si le site est petit, il peut être possible d'interviewer chaque participante et participant. Le plus souvent, la population du site est trop diverse pour qu'un même ensemble de questions soit pertinent pour tous les enquêtés. Ainsi, un sujet de préoccupation des cheffes et chefs d'équipe peut ne pas concerner celles et ceux qui fabriquent les instruments, et vice versa. Une démarche alternative consiste à stratifier l'échantillon par catégories théoriques du phénomène étudié, en équilibrant le nombre et la distribution des participants dans chacune.

Traces et documents

Les traces et les autres formes de documents associées à un site de recherche peuvent être des ressources en données très utiles, mais elles ne constituent pas des preuves en elles-mêmes, comme le savent bien les spécialistes de l'histoire et de l'archivistique. La compréhension des traces d'activité humaine dépend de la date, du lieu et des circonstances où elles sont utilisées (Furner, 2004a). Les personnels de recherche doivent se renseigner le plus possible sur le contexte de chaque trace et ainsi tisser une trame à partir de ces informations rassemblées. Par exemple, des différences entre les recueils de données de deux laboratoires peuvent tenir aussi

bien de leurs habitudes de conservation que des données qu'elles collectent. De même, des différences dans les éléments biographiques sur des personnes reflètent la façon dont elles choisissent de se présenter et la fréquence à laquelle elles le font.

Les spécialistes de la sociotechnique ont recours à des sources publiques pour en apprendre autant que possible sur les sites potentiels. Il est souvent facile de trouver des informations sur le financement, le personnel, les équipements, les événements, les activités de recherche et les publications. S'il faut généralement vérifier les détails des documents publics avec les enquêtés, recueillir ces informations à l'avance permet aux chercheurs et chercheuses de mieux employer leur temps sur place. Lorsqu'ils observent les participants et participantes ou s'entretiennent avec eux, les chercheurs rassemblent souvent autant de documents internes pertinents que les enquêtés veulent bien leur en communiquer. Ainsi, des traces telles que les carnets de laboratoire peuvent fournir un éclairage sur la conduite de la recherche. Certaines représentent des contenus sensibles, que les chercheurs protégeront dans le cadre d'accords de confidentialité.

Concevoir et évaluer des technologies

Un élément de l'étude d'un site de recherche consiste à observer comment les enquêtés utilisent les technologies dans leur travail. Des travaux sur les bibliothèques numériques, les archives de données, les technologies éducatives, les outils collaboratifs, les modèles climatiques, les systèmes de traitement de texte, le courriel et bien d'autres technologies nous ont renseignés sur la manière dont nous travaillons (Blomberg et Karasti, 2013 ; Bowker et Star, 1999 ; Edwards *et al.*, 2007 ; Edwards, 2010 ; Karasti *et al.*, 2006 ; Olson *et al.*, 2008 ; Ribes et Finholt, 2009).

Certains spécialistes sociotechniques bâtissent des systèmes dans le cadre d'interventions dans les communautés de pratique. Dans les bibliothèques numériques, on peut concevoir de petits systèmes pour tester de nouvelles interfaces utilisateur, des fonctionnalités de recherche d'information, des théories pédagogiques ou des processus cognitifs. L'Alexandria Digital Earth Prototype Project, par exemple, était une initiative conjointe de géographes, d'informaticiennes et d'informaticiens, de psychologues et de spécialistes des sciences de l'information pour utiliser des données des sciences de la terre dans des cours de premier cycle (Borgman *et al.*, 2000 ; Janee et Frew, 2002 ; Smith et Zheng, 2002). Le Science Library Catalog consistait en une interface utilisateur graphique qui servait à étudier les compétences cognitives et développementales d'enfants de huit à douze ans lorsqu'elles et ils cherchaient des informations scientifiques. Le système a été bâti progressivement et amélioré en fonction des capacités des enfants. Ainsi, au cours de plusieurs années de recherche, des hypothèses sur

les différences d'âge dans la capacité à manier les hiérarchies, à catégoriser des informations, à chercher par ordre alphabétique ou sur un graphique et à persévérer dans l'accomplissement de tâches informationnelles ont pu être éprouvées (Borgman *et al.*, 1995 ; Hirsh, 1996).

Les infrastructures de la connaissance

La sociotechnique est une problématique plus qu'une discipline. Les chercheurs et chercheuses convergent vers un problème présentant des aspects sociaux et techniques, comme les pratiques en matière de données, chacun apportant ses propres théories, ses méthodes et ses perspectives. Ils amènent également avec eux, si elles existent, les infrastructures de la connaissance de leurs disciplines respectives. Comme pour la recherche sur Internet, l'expertise partagée, les méthodes et les outils sont plus essentiels à l'infrastructure que les archives. Les données résultant de ces méthodes mixtes sont particulièrement difficiles à organiser, partager ou conserver. Il n'existe que peu de référentiels ou de standards communs qui puissent être appliqués aux notes de terrain, aux sites web, aux enregistrements audio et vidéo, aux fichiers, aux logiciels, aux photographies, aux échantillons physiques et aux myriades d'autres types de données pouvant être recueillies dans ce genre de travaux. Diffuser les données peut revenir à les disperser par genre et par matière, ce qui est contre-productif pour la réplcation.

Métadonnées

Étant donné la combinaison des sources de données, on tend à élaborer les conventions de nommage et les pratiques documentaires en interne. Certains personnels de recherche rédigent des guides de codification pour favoriser un encodage cohérent à long terme et par des codeurs et codeuses multiples. En principe, la DDI peut servir à structurer des données et métadonnées qualitatives (Data Documentation Initiative, 2012 ; Vardigan *et al.*, 2008). Dans la pratique, seuls les grands projets comptant de multiples collaborateurs et collaboratrices ont suffisamment intérêt à investir dans des pratiques formelles en matière de métadonnées.

Il est difficile de normaliser les pratiques d'encodage de données qualitatives parce que les sources de données varient beaucoup d'un projet à l'autre et d'un site à l'autre. La nature itérative des observations sur le terrain et de la codification décourage aussi la normalisation : ainsi, la théorie ancrée encourage les scientifiques à enrichir et réviser leur structure de codification à mesure qu'elles et ils collectent des données et découvrent davantage d'informations contextuelles sur celles qu'ils possèdent déjà (Anderson, 1994 ; Glaser et Strauss, 1967 ; Star, 1999). Les hypothèses sont élaborées à travers la codification, puis testées sur d'autres parties du corpus.

L'itération améliore la cohérence et donc la fiabilité des méthodes. Néanmoins, la validité externe peut diminuer à mesure que la validité interne augmente.

Provenance

Consigner la provenance pour des méthodes mixtes nécessite des registres pour chaque type de données et pour les relations qui les unissent. Imaginons, par exemple, qu'une photographie soit prise à un moment et dans un lieu donné dans le cadre d'un entretien donné. Un ensemble de documents est également acquis à cette occasion. La photographie présente un intérêt si elle est comparée à d'autres prises par le même laboratoire à d'autres moments ou à des photographies d'autres laboratoires prises à peu près au même moment. Chaque série d'entretiens, de photographies, d'enregistrements, de documents et d'autres formes de données peut avoir de multiples lignages. La provenance, dans le cadre des méthodes mixtes et des autres formes de recherche interprétative, peut comprendre de multiples réseaux de relations.

La réplication, en soi, représente rarement un problème, car l'explication idiographique est ancrée dans des personnes, des lieux, des moments et des situations. En revanche, la sociotechnique s'inquiète tout autant de la véracité que les autres champs. Les scientifiques acquièrent, quand c'est possible, des informations de sources plurielles et indépendantes. L'information de provenance peut répondre aux besoins du chercheur ou de la chercheuse plutôt qu'à ceux de la réutilisation par d'autres. Il peut conserver ses données sa vie durant, s'assurant ainsi que ses observations puissent être vérifiées. Dans d'autres cas, les comités de protection des personnes exigent que les données soient détruites à la fin du projet. En effet, la confidentialité des traces et l'anonymat des individus dans la présentation des données sont primordiaux. Plus la provenance d'une trace quelconque sur un enquête est détaillée, plus il est difficile de diffuser les jeux de données.

Les influences extérieures

La recherche sociotechnique implique des sujets humains et est donc contrainte par une série de facteurs économiques, éthiques et de propriété comparables à ceux des autres domaines des sciences sociales. La démarche idiographique étant généralement fondée sur une relation de travail étroite entre les scientifiques et les sujets, ces types de données peuvent être plus sensibles que celles issues des enquêtes sur Internet ou des études des médias sociaux, qui saisissent des traces publiques d'activité.

Économie et valeur

L'« emballage » des données influe sur leur valeur. Les données issues de travaux à méthodes mixtes peuvent être combinées de tant de façons qu'il est difficile de les situer dans les quadrants des biens économiques. Les spécialistes de la sociotechnique peuvent puiser dans des réservoirs communs de ressources associés à leurs sites de recherche, comme des archives de données ou des dépôts de publications, dont les enquêtés sont les auteurs et les autrices. Les informations concernant les sujets de recherche peuvent être publiques ou privées, selon la manière dont on les a obtenues. Cependant, une fois possédés, les produits de données sociotechniques deviennent parfois des biens privés qui ne peuvent pas être partagés sous peine de révéler l'identité des sujets. Par exemple, une liste des pages web des sujets de recherche peut être constituée de liens vers des informations publiques, mais la liste elle-même représente le groupe des participantes et participants et doit être tenue confidentielle.

Droits de propriété

Les personnels de recherche en sociotechnique accèdent à des informations qu'ils peuvent fouiller pour découvrir des éléments probants ; cependant, y accéder ne signifie pas qu'ils peuvent les diffuser ou les reproduire. Par exemple, les carnets de laboratoire demeurent généralement la propriété de l'équipe de recherche et les publications, celle des ayants droit. Toute source de données des études sociotechniques peut être grevée de droits de propriété, que ce soit du matériel, des logiciels, des documents, des échantillons ou d'autres matériaux.

Éthique

Les travaux sociotechniques sont généralement assujettis aux politiques en matière de sujets humains. Néanmoins, les règles de consentement, d'anonymat, de confidentialité et de contrôle des registres peuvent varier légèrement selon les méthodes appliquées. Les risques et les responsabilités ne sont pas toujours les mêmes pour les ethnographies, les observations sur le terrain, les entretiens, les évaluations de technologie, etc. Des problèmes éthiques se posent dans la manière de recueillir les données, de protéger les registres et de rapporter les résultats. Par exemple, lorsqu'on publie des travaux, le niveau de détail fourni dépend parfois des risques encourus par les sujets. Les pratiques d'identification varient selon les revues et les communautés. Dans la plupart des cas, les sites de recherche sont dissimulés en supprimant des détails locaux et en attribuant des pseudonymes aux enquêtés ; mais dans d'autres, les sites et les personnes sont nommés. Les résultats sont agrégés dans des groupes suffisamment grands pour qu'il soit difficile d'identifier les individus.

Les observations peuvent comprendre de longs récits à travers les mots des participantes et participants, des éléments biographiques et des vidéos de visages. Tout travail de recherche qui repose sur un grand niveau de détail concernant des individus et des groupes suppose un investissement considérable dans la sauvegarde de la confidentialité et des droits des personnes. Parmi les propositions de nouvelles règles nationales pour la protection des sujets humains aux États-Unis, on trouve un cadre en quatre parties sur la confidentialité : « données sûres, endroits sûrs, personnes sûres et livrables sûrs » (National Research Council, 2013, p. 50). Dans le modèle proposé, les études sont conçues pour une meilleure sécurité des données avant, pendant et après la collecte. Une méthode consiste à recueillir des données sur plusieurs sites, pour que le lieu soit plus difficile à déceler. Les données confidentielles peuvent être partagées dans des « endroits sûrs », comme des « enclaves de données virtuelles » : elles sont conservées dans un *data center* où les scientifiques peuvent, sur demande, manipuler les données, mais pas les transférer sur leurs propres ordinateurs. Les personnels de recherche qui diffusent des données et ceux qui les réutilisent peuvent devenir des « personnes sûres » grâce à des formations et des certifications. Les livrables peuvent être rendus plus sûrs en catégorisant les données selon leur nocivité. Ainsi, pour les données qui risquent peu de nuire, de simples accords sur les conditions d'utilisation peuvent suffire. Pour les données « véritablement radioactives », des règles plus strictes prévaudraient (National Research Council, 2013, p. 52). Quand bien même ces règles s'appliqueraient à toutes les catégories de données sur les sujets humains, c'est la recherche qualitative qui connaîtrait les changements les plus marqués. Certains des amendements proposés visent à favoriser le partage et la réutilisation des données sur les sujets humains.

Mener des recherches sociotechniques au CENS

Le Center for Embedded Networked Sensing offrait une occasion rare d'étudier, sur le long terme, la formation de collaborations et la collecte, l'analyse, la gestion et la publication des données. Comme nous l'avons expliqué au chapitre 5, le CENS était un centre scientifique et technologique relevant de la National Science Foundation, fondé en 2002 avec un financement pour cinq ans. Une deuxième subvention pour cinq années supplémentaires a prolongé l'existence du CENS jusqu'en 2012. La responsable de la recherche sur les pratiques en matière de données était également l'une des cofondatrices du centre. L'équipe sociotechnique chargée des pratiques en matière de données, qui rejoindrait plus tard l'équipe des statistiques, était une unité de recherche du CENS. Elle comptait également, comme beaucoup des autres unités, des collaborateurs et collaboratrices en dehors du centre.

Une grande partie de la recherche sur les pratiques concernant les données au sein du CENS consistait en de l'observation participante, puisque l'équipe était intégrée au centre. Les professeures et professeurs, les étudiantes et étudiants et le personnel de l'équipe prenaient part à des activités formelles et informelles et passaient le plus de temps possible sur le site. Au cours de l'existence du CENS, l'équipe a réalisé des ethnographies et des entretiens ouverts, observé des essais pilotes en laboratoire et sur le terrain, participé à des déploiements sur le terrain de quelques heures à plusieurs semaines, assisté à d'innombrables réunions d'équipe, présenté ses résultats lors de colloques, de journées d'actualités et de retraites scientifiques et analysé des documents tels que des carnets de laboratoire et de terrain, des fichiers de données de formats divers et des publications. La collecte de données se poursuit aujourd'hui encore, car l'équipe étudie la postérité du CENS dans les années suivant sa fermeture officielle (Borgman *et al.*, 2014).

Les observatrices et observateurs participants sont rarement des membres officiels de l'organisation qu'ils étudient, mais dans le cas du CENS, l'équipe de recherche en sciences sociales avait des responsabilités administratives analogues à celles des autres équipes. Elle a également contribué à l'activité du centre en concevant des technologies pour la collecte et la gestion de données scientifiques, dont un site de dépôt en accès ouvert pour les publications.

Comme pour toute méthodologie de recherche, il fallait en permanence trouver des compromis en matière d'échelle, d'envergure et d'objectivité. En tant que membre à part entière du CENS, l'équipe disposait d'un bien meilleur accès aux individus, aux sites et à la documentation que la plupart des spécialistes des sciences sociales. Elle était aussi consciente du risque de subjectivité. Être conscient de ces risques contribuait déjà à les atténuer, de même qu'avoir des partenaires extérieurs qui posaient des questions critiques sur les méthodes et l'interprétation. Généralement, on demandait aux spécialistes des sciences exactes et des technologies du CENS de vérifier les descriptifs de leurs activités avant qu'ils ne soient publiés par l'équipe sociotechnique. Elles et ils les ont corrigés et améliorés de bonne grâce et sans essayer d'altérer les interprétations effectuées.

Les questions de recherche

L'équipe chargée des pratiques en matière de données a débuté son étude du CENS en se demandant comment les chercheurs et chercheuses individuels et leurs équipes créaient des données et comment ces pratiques varient selon les équipes et les domaines. Étant d'entrée de jeu des membres de la communauté, les spécialistes sociotechniques avaient la possibilité de suivre les données de leur conception initiale au traitement final en passant par les étapes de nettoyage, d'analyse et

de présentation. À mesure que le contexte du travail scientifique des données était mieux appréhendé, les questions de recherche ont évolué pour s'intéresser au travail collaboratif des équipes, au contrôle et à la propriété des données, aux métadonnées et à la provenance, au partage et à la réutilisation et aux applications éducatives des données du CENS. Plusieurs thèses de doctorat, un mémoire de master et de nombreux projets étudiants sur les pratiques en matière de données et les collaborations au sein du centre ont encore élargi les questions de recherche traitées depuis 2002.

Les questions de recherche et le financement ont été réaffirmés tout au long des projets de l'équipe. Les fonds de départ fournis par le CENS ont financé le traitement de la problématique initiale s'agissant des données et des pratiques. Ces résultats ont mené à des demandes de subvention, qui ont financé de nouvelles découvertes qui ont à leur tour conduit à d'autres demandes et à des publications. La plupart des bourses étaient attribuées à plusieurs collaborateurs et collaboratrices, permettant des comparaisons des pratiques en matière de données avec d'autres sites de recherche (Borgman *et al.*, 2006, 2007a, 2007b, 2012 ; Borgman, 2006 ; Edwards *et al.*, 2011 ; Mandell, 2012 ; Mayernik *et al.*, 2007, 2012 ; Mayernik, 2011 ; Pepe *et al.*, 2007, 2010 ; Pepe, 2010 ; Shankar, 2003 ; Shilton, 2011 ; Wallis *et al.*, 2007, 2008, 2010a, 2010b, 2012, 2013 ; Wallis, 2012).

La collecte de données

Le terrain des spécialistes sociotechniques englobait tout endroit où se trouvaient les chercheurs et chercheuses du CENS. Les équipes étaient basées dans l'une des cinq universités participantes et avaient parfois des partenaires dans d'autres institutions. Les données étaient recueillies dans les laboratoires et les espaces publics du campus, ainsi que dans des sites de recherche aux États-Unis et ailleurs dans le monde. Différentes méthodes étaient appliquées simultanément par une équipe allant de deux à huit membres selon les périodes. Certains chercheurs et chercheuses se concentraient sur l'observation des terrains, d'autres sur la conduite d'entretiens, l'analyse de documents, la conception et l'évaluation de technologies et l'identification des réseaux sociaux au sein du CENS. Chaque personne de l'équipe a pris part à au moins deux méthodes de collecte pour bénéficier d'une formation polyvalente et faciliter la combinaison de données à partir de méthodologies multiples.

L'une des premières activités de l'équipe sociotechnique a consisté à identifier des standards et des formats de métadonnées adaptés aux activités de recherche de cette communauté interdisciplinaire naissante. La tâche s'est avérée bien plus difficile que prévu et a orienté les premières années du projet. Les membres du CENS n'ayant guère d'expérience des standards de métadonnées, leur demander directement leurs préférences n'a pas abouti à des renseignements utiles. L'étape

suivante dans ce but a donc consisté à identifier la gamme des données recueillies. Les observations sur le terrain, les entretiens et l'analyse des publications du CENS ont été les principales sources d'information sur ce sujet. L'équipe a repéré quelles données les scientifiques et technologues acquéraient et lesquelles elles et ils conservaient le plus souvent en vue d'une analyse et d'une réutilisation ultérieure.

Une fois les besoins généraux en métadonnées identifiés, on les a comparés aux standards existants ou émergents. Plusieurs standards pour les données environnementales et pour les données de capteurs constituaient des candidats prometteurs. En effet, ils pouvaient, en principe, être utilisés seuls ou combinés pour décrire la majeure partie des données du CENS. Les observations écologiques recueillies par les capteurs ou à la main pourraient être décrites avec une structure et une terminologie communes. Par ailleurs, les caractéristiques des capteurs seraient enregistrées automatiquement si ces standards XML étaient intégrés dans les algorithmes de collecte.

Les standards de métadonnées ont alors été présentés aux chercheurs et chercheuses du CENS en expliquant leurs avantages et leurs inconvénients. Bien que reconnaissant des efforts de l'équipe sociotechnique pour les aider à gérer leurs données, les membres du CENS ont estimé que les standards existants ne répondaient pas à leurs besoins. Comprendre les raisons de cette non-adoption a nécessité encore plusieurs années de recherche approfondie, comme nous le verrons plus en détail au chapitre 8 (Borgman *et al.*, 2006, 2007a ; Borgman, 2006 ; Pepe *et al.*, 2007 ; Shankar, 2003 ; Wallis *et al.*, 2006).

À mesure que le CENS voyait le nombre de ses participantes et participants passer de quelques dizaines à plusieurs centaines, le besoin de meilleurs registres de provenance s'est fait sentir. En effet, la culture orale ne fonctionnait plus. Les équipes de recherche partaient pour un site de déploiement, souvent à plusieurs heures de route, pour se rendre compte une fois sur place qu'il leur manquait une pièce matérielle essentielle ou un savoir-faire particulier détenu par un ou deux participants – typiquement, un étudiant ou une étudiante de second cycle. L'équipe sociotechnique s'est efforcée de combler cette lacune grâce à un logiciel simple, baptisé « CENS Deployment Center » (Mayernik *et al.*, 2007). CENS DC, comme on l'appelait, était rempli de descriptifs du matériel et du personnel présents lors des déploiements précédents – données que l'équipe sociotechnique avait recueillies en amont. Le système comportait des fonctions de modèles pouvant servir à générer un projet de déploiement et à créer des registres sur le terrain, par exemple pour indiquer ce qui avait fonctionné ou non, ce qui manquait ou s'avérait particulièrement utile et ainsi de suite. Ces fonctionnalités visaient à rendre les

déploiements plus efficaces et productifs grâce à l'observation des précédents. Le système enregistrait également des catégories d'information fréquentes dans les publications scientifiques : il pouvait ainsi aider à la rédaction d'articles sur la recherche de terrain. Fondé sur le travail de certaines équipes, CENS DC a été testé par elles avant d'être mis en service. Il a partiellement rempli ses fonctions chez les quelques équipes qui l'ont intégré dans leurs activités (Mayernik, 2011 ; Wallis, 2012).

L'analyse des données

Comme les enquêtes OxIS, la recherche sur les pratiques en matière de données du CENS a démarré de manière modeste. À mesure que l'équipe grandissait et qu'il devenait possible de réaliser des études plus longues et importantes, l'analyse de données s'est faite plus formelle. Les entretiens étaient enregistrés et transcrits de façon professionnelle. On a rédigé des guides de codification pour maintenir une cohérence entre les multiples enquêtes et les nombreuses observations sur le terrain. NVivo, un progiciel commercial d'analyse de données pour la recherche qualitative et à méthodes mixtes, a été utilisé pour encoder les personnes, les événements, les thématiques et d'autres catégories mentionnées dans les notes et les transcriptions (NVivo 10, 2013). À chaque collecte de données, deux étudiantes et étudiants de second cycle encodaient les mêmes entretiens chacun de leur côté. Ils les comparaient ensuite et discutaient leur codification, s'efforçant de concilier leurs différences d'interprétation. On répétait la procédure jusqu'à obtenir une concordance entre codeurs. Les guides de codification étaient annotés d'explications sur les interprétations, ce qui contribuait à leur cohérence. Le défi consistait à élaborer et éprouver des hypothèses à partir de ces données tout en restant ouvert à des résultats imprévus.

Les transcriptions d'entretiens se prêtaient davantage à la concordance entre codeurs que d'autres types de données. En effet, bien que les entretiens soient menés par plusieurs membres de l'équipe, tous travaillaient à partir d'un même ensemble de questions à poser aux chercheurs et chercheuses du CENS. En revanche, sur le terrain, chaque membre prenait copieusement des notes sur ses observations. Ces notes devaient être codifiées par l'observateur ou l'observatrice. S'ils s'en chargeaient rapidement et régulièrement, les observateurs pouvaient combler les lacunes de leurs carnets ou déterminer si un élément manquant pouvait être acquis par d'autres moyens. On encourageait les observateurs à prendre beaucoup de notes, ainsi que des photos quand c'était possible. La prise de notes s'est améliorée avec l'expérience. Néanmoins, il y avait d'importantes variations dans l'attention au détail et dans la connaissance des phénomènes observés chez les membres de l'équipe. L'étudiante ayant une formation en biologie a noté de nombreux détails

sur les espèces étudiées, tandis que le diplômé en génie s'est davantage concentré sur l'instrumentation mise en place. Disposer sur le terrain d'observateurs multiples ayant des parcours complémentaires a élargi le spectre de l'information acquise sur les pratiques en matière de données du CENS.

Les données encodées avec NVivo pouvaient être agrégées par thématiques, événements et diverses autres catégories. On pouvait extraire des sous-ensembles de données et ainsi effectuer des comparaisons entre études. Plus il y avait de données recueillies, plus les comparaisons et les interprétations seraient riches. Les fichiers NVivo, les mémos, les photographies et d'autres traces sont devenus matière à discussion pour l'équipe sociotechnique et ont ainsi contribué à orienter les publications, les présentations et les posters scientifiques.

La publication des résultats

La première étape de partage des résultats de l'équipe sociotechnique se faisait souvent au sein même du CENS. Celui-ci mettait sur pied des présentations par posters et des démonstrations dans le cadre des événements publics qu'il organisait régulièrement, comme les visites annuelles de la National Science Foundation, les journées annuelles d'actualités ouvertes au public et les retraites scientifiques. L'équipe sociotechnique a réalisé des posters à partir de son propre travail et a contribué à ceux d'autres équipes du CENS. Dans ces manifestations, les visiteurs et visiteuses se rencontraient autour de trente à quatre-vingts posters, offrant autant d'occasions de discuter et de bénéficier de critiques. Le CENS organisait également des séminaires hebdomadaires à l'heure du déjeuner pour rapporter les travaux du moment. L'équipe sociotechnique assistait régulièrement à ces événements et y présentait de temps à autre ses résultats.

Les résultats de la recherche sur les pratiques en matière de données du CENS étaient publiés dans différents médias pour atteindre plusieurs publics. L'essentiel des fonds du centre provenant de l'informatique, de nombreuses communications ont été proposées aux conférences de l'ACM (Association for Computing Machinery). Il s'agit de lieux de publication prestigieux et très sélectifs. Intervenir lors de conférences ACM était donc une occasion importante de présenter des découvertes sociotechniques à un auditoire technique. D'autres publications étaient destinées aux chercheurs et chercheuses en sciences de l'information et en sociologie des sciences et des technologies. Quelques publications ciblaient le corps enseignant, notamment sur le thème de la réutilisation de données à des fins pédagogiques. Outre ces publications, cette littérature a donné lieu à de nombreuses présentations qui ont atteint des publics plus variés encore dans les sciences exactes, les sciences sociales, la technologie et les sciences humaines.

Ces publications dévoilent l'identité du CENS, mais les participantes et participants individuels ne sont pas nommés. Les personnes citées sont désignées par une catégorie (par exemple, scientifique, technologue) ou reçoivent un pseudonyme. Le nom du centre est révélé pour plusieurs raisons. La première est que les particularités de la recherche en réseaux de capteurs intégrés sont essentielles pour expliquer ses pratiques en matière de données ; le site est donc trop singulier pour être dissimulé aisément. Il aurait fallu, pour anonymiser le CENS, supprimer tant de contexte que les résultats en auraient été rendus vides de sens. Une seconde raison est de promouvoir le travail du CENS, puisque l'équipe sociotechnique appartenait à cette communauté. La National Science Foundation a reconnu la contribution du centre aux sciences sociales en plus de ses réussites scientifiques et techniques. En faisant du CENS un site exemplaire pour la recherche sur les pratiques en matière de données, elle a encouragé d'autres centres à accueillir des études en sciences sociales. Néanmoins, parce que le site de recherche était nommé, il a fallu prendre plus de précautions encore en agrégeant les résultats à un niveau qui dissimule l'identité des individus.

La conservation, le partage et la réutilisation des données

Les données de l'équipe sociotechnique, constituées d'enregistrements audio, de transcriptions, de notes de terrain, de fichiers NVivo et d'une variété de traces publiques et privées, sont stockées en sûreté sous le contrôle de l'équipe et de ses membres. Conformément aux recommandations de l'Institutional Review Board (IRB), les registres papier et les autres documents sont conservés dans des meubles de classements verrouillés dans des bureaux fermés. Les registres numériques sont stockés dans des serveurs sécurisés. Alors que l'IRB exige parfois que les données soient détruites en fin de projet, l'équipe renouvelle chaque année son autorisation pour continuer à analyser ses données et les comparer à celles recueillies ultérieurement. Selon les consignes actuelles, les données d'un projet ne peuvent plus être analysées une fois que l'autorisation de l'IRB a expiré.

En ce qui concerne les subventions impliquant plusieurs universités, chaque chercheur et chercheuse a dû obtenir une autorisation de l'IRB pour les données collectées par son équipe. Bien que les politiques universitaires soient globalement similaires, la possibilité de partager des données entre universités était limitée. De manière générale, les chercheurs pouvaient se transmettre des données encodées et anonymisées entre universités, mais pas les transcriptions et les autres traces comportant des informations permettant d'identifier les sujets humains. Plutôt que mettre en commun les données des collaborateurs et collaboratrices, chaque laboratoire universitaire participant gérait ses données isolément. Lorsque de nouveaux étudiants et étudiantes et chercheurs et chercheuses postdoctoraux rejoignaient une

équipe, ils devaient être certifiés et ajoutés au protocole IRB pour les jeux de données avec lesquels ils seraient amenés à travailler.

À l'heure où nous écrivons, les organismes de financement n'ont pas demandé que les données soient divulguées et aucun chercheur ou chercheuse hors de l'équipe n'a sollicité d'accès. Ces données représentent une mine de ressources pour l'équipe et sont utilisées en comparaison avec des études sur les pratiques en matière de données dans d'autres domaines. Lorsque le financement prendra fin et si les autorisations de l'IRB expirent sans être renouvelées, il sera difficile de les réutiliser, même par les spécialistes sociotechniques qui les ont produites. Les sujets de recherche ont signé des formulaires de consentement pour participer à ces études. Ces documents, qui font plusieurs pages et subissent un examen approfondi de la part de l'IRB, promettent, en échange de la participation à l'étude, la confidentialité des données permettant d'identifier les personnes dans l'analyse et la présentation des résultats.

En bref, ces données sont conservées pour l'usage de l'équipe de recherche. Elles ne peuvent être diffusées qu'aux individus figurant dans le protocole approuvé par l'IRB. Il est possible qu'une partie de ces données, comme les traces tirées de sites publics, soient diffusables un jour. Les documents obtenus dans le cadre des entretiens tombent sous le coup des accords de confidentialité des formulaires de consentement, sauf mention contraire expresse. Certaines notes de terrain pourraient demeurer utiles à d'autres si les noms et les autres informations identificatoires étaient retirés. En revanche, il est peu probable que les enregistrements audio, les transcriptions d'entretiens, les traces créées par les sujets de recherche et les autres informations étroitement liées à l'identité des personnes soient diffusés, pour des raisons éthiques et du fait des conditions de consentement qui ont permis d'obtenir ces données. Pris dans son ensemble, le jeu de données contient des informations pluridimensionnelles sur les activités du CENS lors de ses dix années de fonctionnement, ainsi que sur sa postérité. En revanche, s'il était étudié de manière parcellaire, le contexte et la provenance ne pourraient jamais être retracés.

Conclusion

La recherche sur Internet et la sociotechnique sont des domaines exemplaires des sciences sociales. Toutes deux appliquent des méthodes novatrices, interrogent le comportement humain et attirent des savantes et savants de diverses disciplines. Leurs infrastructures de la connaissance sont davantage caractérisées par un savoir partagé que par des infrastructures techniques ou des ressources informationnelles communes. Les accords sur les méthodes de recherche, telles qu'elles

s'incarnent dans les manuels et les cours universitaires, forment un substrat d'expertise partagée au sein des sciences sociales. Ces méthodes emploient des outillages techniques communs, comme les logiciels d'analyse de données et de modélisation informatique.

La recherche par sondage, qu'elle s'applique à l'usage d'Internet ou à n'importe quel autre comportement humain, s'appuie sur une longue tradition de savoir-faire méthodologique. Ces pratiques arbitrent la richesse des informations que l'on peut recueillir sur les personnes, le contrôle précis de la variance et la confidentialité des données. Les éléments probants sont rapportés avec une explication détaillée des plans de sondage, des populations, des questions posées, de la distribution des réponses et des conclusions au regard des résultats, le tout en garantissant l'anonymat des participantes et participants aux études. Les données issues de travaux présentant des niveaux adéquats d'assurance qualité peuvent être versées dans des archives de données ou diffusées directement par les chercheurs et chercheuses.

Les études sur les médias sociaux, comme les exemples sur le microblogage que nous avons présentés ici, s'appuient également sur de longues traditions méthodologiques, comme celles de l'analyse des réseaux sociaux. Les chercheurs et chercheuses venus d'autres champs apportent leurs propres traditions, comme les méthodes de topologie de réseau en informatique. Ces utilisations contrastées de ressources en données communes peuvent contribuer au transfert d'expertise entre disciplines ; mais ces démarches peuvent aussi avoir si peu en commun qu'elles deviennent opaques aux spécialistes d'autres champs. Le plus frappant dans cette comparaison est que chaque scientifique considérera comme des données des choses très différentes. Par exemple, on peut estimer qu'un lien entre des comptes Twitter met en évidence une influence, une structure de graphe, une communication illicite, une relation personnelle ou d'autres phénomènes. Le contenu d'un tweet ou d'un autre message sera traité, selon la question de recherche et la méthode, comme preuve d'une multitude de phénomènes. Déterminer quels messages représentent des conversations intentionnelles entre êtres humains, lesquels ont été échangés entre des êtres humains et des robots sociaux et lesquels constituent des échanges robotiques – et déterminer quand ces catégories ont une importance – requiert une expertise sociologique et technologique sur l'environnement étudié.

La sociotechnique, telle qu'elle s'est pratiquée dans les dix ans de travaux au sein du Center for Embedded Networked Sensing, applique à un problème un éventail de méthodes de recherche complémentaires. Dans notre cas, la problématique était de comprendre les pratiques en matière de données des nombreux participants et participantes du CENS et concevoir des outils et des services pour les

renforcer. Les questions et les méthodes de recherche ont évolué à mesure que la compréhension des pratiques s'améliorait, que les technologies étaient conçues et testées et que la recherche du CENS gagnait en maturité. L'infrastructure de la connaissance du centre incorporait une expertise partagée dans la science et la technologie des réseaux de capteurs intégrés, un espace collaboratif commun et d'autres formes d'assistance administrative et collégiale. L'investissement dans des ressources informationnelles a cependant été minime, puisque les technologies formaient, plus que les données, le terrain d'entente. L'équipe sociotechnique gérait son propre référentiel de données en interne. Ces données restent utiles pour des comparaisons avec d'autres environnements de recherche et pour poursuivre l'étude de la postérité du CENS. Leur richesse et la diversité de leurs genres les rendent cependant difficiles à partager. En effet, plus les informations sur des individus et des groupes sont complètes, moins elles sont anonymisables. Quand bien même on surmonterait ces difficultés, il est peu probable que l'on trouvera une demeure commune pour les enregistrements audio, les transcriptions, les photographies, les notes de terrain, les interviews codifiées, les publications, les posters scientifiques, les innombrables documents papier et numériques acquis auprès des sujets de recherche et les divers objets technologiques ; or, les disperser selon leur forme ou leur matière les rendra moins intéressants pour la recherche.

Au regard des provocations du chapitre 1, ces études de cas en sciences sociales montrent les diverses façons dont une même entité peut être traitée en donnée, la variété des méthodes de recherche applicables à une source de données commune et la diversité des infrastructures de la connaissance qui émergent pour faciliter ces types de travaux scientifiques. En ce qui concerne la première provocation, la propriété et le contrôle des données ne s'exercent pas de la même façon entre ces différents cas. L'Oxford Internet Institute recueillait les données des enquêtes OxIS et en conservait le contrôle. Le sondage ne saurait être reproduit, puisqu'il s'agit d'observations liées à des moments et des lieux spécifiques, mais il peut être répété plus tard avec de nouveaux échantillons. Les données d'OxIS sont diffusées après deux ans d'embargo, mais elles ne sont pas versées à des référentiels de sciences sociales. Les études sur le microblogage se prêtent moins à la reproductibilité ou au partage. Les travaux sociotechniques sur le CENS ne sont pas répliquables, mais les protocoles en ont été appliqués ultérieurement à d'autres sites de recherches pour effectuer des comparaisons. Ces données restent sous le contrôle des chercheurs et chercheuses et le partage en est limité par la réglementation sur les sujets humains.

La transférabilité, objet de la deuxième provocation, se joue aussi différemment selon les études de cas. Si les accords sur les méthodes de recherche fournissent un terrain d'entente à l'ensemble des sciences sociales, les techniques spécifiques

de manipulation des données – nettoyage, interpolation de valeurs manquantes, suppression des valeurs aberrantes, etc. – varient néanmoins suffisamment pour que l'analyse et l'interprétation des jeux de données divergent. L'évolution des fonctions de la communication savante, telle qu'évoquée dans la troisième et la quatrième provocation, semble plus graduelle que radicale dans les exemples présentés. Nos scientifiques s'inquiètent bien plus de publier leurs résultats que de diffuser leurs données. Elles et ils puisent à des sources de données plus volumineuses et variées que par le passé et recoupent des questions de recherche traitées dans les sciences exactes, la recherche en technologie et les sciences humaines. Malgré ces sources de données communes à plusieurs domaines, les différences de méthodes, de questions et de représentations peuvent limiter la découvrabilité et la réutilisation des jeux de données créés.

Ces études de cas montrent aussi la réorientation de l'expertise nécessaire à la recherche en sciences sociales, exposée dans la cinquième provocation. La combinaison de savoir-faire qualitatifs et quantitatifs dont il est fait preuve varie selon les études. Un certain degré de compétence informatique est nécessaire dans chacun de ces domaines, que ce soit pour développer de nouveaux outils, rédiger des scripts dans des outils existants ou exploiter des routines statistiques complexes. La capacité à réutiliser ces données est contrainte par l'accès à ces outils, ces scripts et ces routines.

Enfin, dans les domaines que nous abordons ici, les infrastructures de la connaissance et le partage de données en particulier attirent relativement peu d'investissements. La découvrabilité et la conservation des données de sondage se taillent la part du lion et leur infrastructure est encore loin d'être complète. Les médias sociaux représentent le front de recherche qui évolue le plus vite. Les objets d'étude et les outils pour l'examiner se transforment trop vite pour les partager fructueusement. Combiner des données issues d'études multiples ou effectuer des méta-analyses implique un investissement substantiel dans l'intégration des données. Dans tous les exemples cités, les jeux de données tendent à rester sous le contrôle des personnels de recherche, que ce soit du fait d'une préférence, d'un manque de solutions alternatives, de problèmes de confidentialité insolubles ou d'un mélange de ces raisons. De ce fait, ces spécialistes des sciences sociales ont peu de motivations pour investir dans des schémas de méta-données ou de classification qui faciliteraient le transfert de leurs données vers d'autres domaines. Une vision à plus long terme des besoins des infrastructures de la connaissance dans les sciences sociales doit se préoccuper de l'acquisition et de la gestion des ressources en données au sein de chacune de ces disciplines, entre elles et avec d'autres domaines scientifiques, ainsi qu'entre ressources

publiques et privées. Compte tenu de la dilution des sources de données, le flou des frontières disciplinaires, le caractère politique et sensible des sujets et la diversité des parties prenantes, les investissements dans les infrastructures de la connaissance des sciences sociales s'annoncent litigieux dans les années à venir.

7. Le travail scientifique des données dans les sciences humaines

Introduction

Les sciences humaines, ou humanités, ont en commun l'étude de la culture et du passé humains en s'appuyant sur toutes les sources de preuves imaginables. Pour le reste, elles sont très divisées, y compris en ce qui concerne les limites du champ. Tout agglomérat de disciplines ou de domaines censé représenter « les sciences humaines » est problématique. L'histoire et l'archéologie, par exemple, sont tantôt classées dans les humanités et tantôt dans les sciences sociales. Les lettres peuvent faire partie des sciences humaines, être classées à part ou être intégrées à des cursus de théâtre, d'architecture ou de *design*. Les programmes académiques en sciences humaines peuvent comprendre des départements de toute taille, de l'histoire antique aux langues étrangères et de la littérature aux cultures du Proche-Orient. N'importe lequel de ces départements peut abriter des chercheurs et chercheuses dotés de diplômes d'autres domaines. L'université de Californie à Los Angeles (UCLA), par exemple, regroupe plus de 150 départements et cursus de premier et second cycle et les affiliations de son corps professoral y sont multiples. Ainsi, le cursus d'humanités numériques y liste environ trente-cinq professeures et professeurs, qui relèvent de départements en sciences humaines, en sciences sociales, en sciences exactes et en formation professionnelle.

Pour les besoins du présent ouvrage, les sciences humaines recouvrent ici les champs de recherche qui étudient la culture et le passé humains. Notre étude inclut aussi les chercheurs et chercheuses qui se reconnaissent dans cette catégorie, quelle que soit leur affiliation officielle. Les chercheurs en sciences humaines font preuve d'ingéniosité pour trouver de nouvelles sources de preuves, qu'elles soient matérielles ou numériques, statiques ou dynamiques, sur terre ou sous terre, très anciennes ou toutes récentes. À mesure que le monde s'est ouvert, ils ont repoussé les limites géographiques de leurs recherches. Avec la numérisation des objets matériels – textes, images, son, etc. –, les types de preuves accessibles se sont diversifiés. Les sciences humaines sont tout autant confrontées à un déluge de données que les autres disciplines, mais elles sont moins promptes à le désigner ainsi. Pour gérer le volume de matière qui s'offre à eux, les chercheurs empruntent

des technologies et des méthodes de champs voisins et élaborent parfois les leurs. Ce faisant, ils commencent à penser en termes de données, de métadonnées, de standards, d'interopérabilité et de pérennité. Emprunter des techniques implique généralement d'emprunter des concepts. Ils ont aussi besoin de représenter des informations sous des formes maniables informatiquement.

Le terme « données » (*data*), tel que défini au chapitre 2, désigne des entités qui servent à mettre en évidence des phénomènes à des fins de recherche. Le « travail scientifique des données » (*data scholarship*), tel que défini au chapitre 3, correspond à l'agencement des rapports entre données et recherche. Cet agencement est particulièrement complexe dans les sciences humaines et ne constitue pas toujours un cadre confortable pour les chercheurs et chercheuses concernés. En se penchant sur les manières dont ils identifient les sources de preuves qu'ils pourront utiliser comme données et comment ils gèrent ces dernières, on fait la lumière sur les défis qu'ils rencontrent et les composantes des infrastructures de la connaissance dont ils pourraient avoir besoin.

Les méthodes de recherche et les pratiques en matière de données

Les spécialistes des humanités observent l'activité humaine par les traces qu'elle laisse derrière elle, qu'il s'agisse d'artefacts physiques, d'images, de textes ou d'objets numériques. Leurs méthodologies tendent vers l'explication idiographique, étudiant de près textes, objets, communautés et cultures. En travaillant avec de plus grandes quantités de données, l'explication nomothétique pourrait devenir réalisable, permettant aux mêmes questions d'être étudiées dans de nouveaux contextes.

On peut faire remonter les humanités numériques à 1949, année où IBM s'est alliée à Roberto Busa, un prêtre jésuite, pour créer une concordance des œuvres complètes de Thomas d'Aquin. Ce projet de numérisation textuelle sur trente ans est désormais en ligne. Les outils, les services et l'infrastructure pour la recherche en sciences humaines continuent de s'étendre parallèlement au volume et à la variété des contenus numériques. Les technologies sont aujourd'hui si intégrées dans les pratiques que certains chercheurs et chercheuses considèrent l'expression « humanités numériques » comme une tautologie. Pour d'autres, ce terme reste utile pour signifier la variété grandissante des méthodes de recherche et des sources de preuve (Alarcón, 2000 ; Borgman, 2007, 2009 ; Burdick *et al.*, 2012 ; Wouters *et al.*, 2012).

Les spécialistes des sciences humaines s'appuient largement sur des pièces uniques – lettres, mémorandums, traités, photographies, etc. – qui n'existent que

sous forme physique. Beaucoup de ces sources sont stockées dans des archives d'universités, de musées, d'administrations et d'institutions publiques et privées. Les archivistes organisent ces pièces de manière à fournir leur contexte, s'efforçant généralement d'en maintenir autant que possible l'ordre original. Les instruments de recherche, qui peuvent exister en version imprimée ou en ligne, offrent un accès hiérarchique à des collections, des boîtes de documents et des éléments individuels. Les chercheurs et chercheuses doivent se rendre aux archives, repérer les boîtes pertinentes et en demander un petit nombre. À l'intérieur des boîtes, les documents peuvent être organisés en dossiers. Les usagères et usagers n'ont souvent le droit d'ouvrir qu'un dossier à la fois pour que les pièces restent dans l'ordre où elles ont été placées. Les archives adoptent des politiques très variées quant à l'usage de technologies personnelles. Certaines permettent aux usagers d'apporter leurs propres scanners. Beaucoup les autorisent à prendre des photographies à l'aide d'appareils numériques et à saisir des notes sur ordinateur. Néanmoins, dans un grand nombre d'archives, les seules technologies autorisées sont le papier et le crayon.

Comparer des objets de recherche est essentiel pour de nombreuses formes d'études en sciences humaines. Ceux-ci peuvent être éparpillés au sein d'une archive ou à travers le monde ; c'est pourquoi les comparaisons se font souvent à partir des représentations de ces objets. Les scientifiques les représentent de nombreuses manières, par exemple en notant leurs caractéristiques, en prenant des photos ou en réalisant des croquis. Elles et ils s'appuient aussi sur des représentations d'autres personnes. Si l'original n'existe plus ou ne peut être observé directement, les chercheurs et chercheuses peuvent dépendre entièrement des représentations en tant que sources de preuve. La numérisation des objets et des représentations a transformé les méthodes de recherche reposant sur ces comparaisons. Les textes, les images, les sons et les autres entités – ou des fractions de celles-ci – peuvent être rapprochés comme jamais auparavant, soit côte à côte sur un écran, soit par modélisation informatique.

Certains chercheurs et chercheuses en sciences humaines effectuent des observations, par exemple en enregistrant une langue orale ou en fouillant des sites archéologiques. Dans ce cas, les représentations des informations peuvent devenir des preuves irremplaçables. Les enregistrements de langues rares ou disparues constituent des ressources pour l'étude de la linguistique et des cultures. Les fouilles archéologiques perturbent le sol au point que les sites ne peuvent jamais retrouver leur état d'origine. Les artefacts sont déterrés et parfois emportés pour leur conservation et leur sécurité. De nombreuses formes de documentation sont saisies, par exemple des photographies, des scans 3D, des pesées, des notes détaillées et des emplacements précis. Des méthodes moins destructives – par

exemple, des techniques d'imagerie comme la détection et télémétrie par ondes lumineuses, plus connue sous l'acronyme LIDAR – peuvent aussi être employées pour effectuer des observations sur les sites.

Les spécialistes des sciences humaines, souvent en collaboration avec des scientifiques d'autres domaines, combinent de nombreuses sources de données et méthodes pour poser de nouvelles questions. Certaines et certains utilisent des systèmes d'information géographique pour cartographier des sites patrimoniaux. D'autres modélisent les activités sociales au sein de villes et d'autres lieux sur des dizaines ou des centaines d'années. D'autres encore font des expériences sur des formes numériques de texte, de poésie et d'édition (Burdick *et al.*, 2012 ; Fitzpatrick, 2011 ; Frischer *et al.*, 2002 ; Presner, 2010). Les méthodes de recherche des sciences humaines sont aussi fluides que leurs sources de données. Bien que les chercheurs et chercheuses aient été formés à des méthodologies propres à leurs domaines, ils ne disposent pas, contrairement aux spécialistes des sciences sociales, d'un substrat commun de pratiques et de cours de méthodologie. Ils acquièrent souvent leur savoir-faire grâce à un long mentorat et à la recherche indépendante. La durée moyenne d'un doctorat en sciences humaines est d'ailleurs bien plus longue que dans la plupart des sciences exactes et sociales.

Études de cas en sciences humaines

Les études de cas de travail scientifique des données dans les sciences humaines s'intéressent à la création et à l'utilisation de collections numériques et à leur rapport à l'étude d'artefacts physiques. Elles ont été choisies pour illustrer la transition des objets matériels vers des objets numériques comme sources de preuve, la transformation du travail scientifique des données et les modalités, la temporalité et les raisons du choix d'utiliser certaines entités comme données. La première étude de cas recouvre deux exemples en histoire de l'art et archéologie antiques. Le premier est le projet CLAROS à l'université d'Oxford, une collaboration pluridisciplinaire pour créer un banc d'essai technologique pour les sites patrimoniaux, les monuments et les objets. Le deuxième exemple s'intéresse à l'analyse archéométallurgique du griffon de Pise, une sculpture de bronze ancienne. Dans ce dernier cas, des chercheurs et chercheuses aux expertises complémentaires ont collaboré pour rédiger des « biographies d'artefacts » pour des objets comme le griffon.

Le deuxième cas concerne les études bouddhiques et se penche sur la manière dont un spécialiste de la philologie bouddhique chinoise utilise des corpus numériques en sus des textes et des artefacts physiques. La numérisation des textes, dont certains datent du III^e siècle, a transformé la façon dont il mène ses recherches.

L'histoire de l'art et l'archéologie antiques

L'histoire de l'art antique et l'archéologie sont des domaines de recherche qui se recoupent. Tous deux s'appuient sur des méthodes, des théories et des savoir-faire tirés des sciences humaines, sociales et exactes. Les artefacts mis au jour lors de fouilles architecturales sont étudiés pour leur valeur artistique et culturelle (Wiseman, 1964 ; Yoffee et Fowles, 2011). Les spécialistes des sciences sociales s'intéressent également à ces objets pour ce qu'ils révèlent des sociétés qui les ont créés et utilisés (Smith *et al.*, 2012). Les archéométallurgistes et les spécialistes des matériaux les examinent pour identifier des métaux, des minerais, des céramiques, des teintures et d'autres caractéristiques physiques pouvant servir à les dater et à retrouver leurs origines (Northover et Northover, 2012). Les zooarchéologues étudient les restes animaux et le recours aux bêtes dans les sociétés, incorporant ainsi les sciences de la vie à l'étude des objets anciens (Crabtree, 1990).

Les fouilles archéologiques représentent une source majeure de preuves scientifiques. Les types de données issues de fouilles résultent de différentes formes de recherche, pratiques professionnelles et époques de collecte. Au ^{xix}^e siècle, lors de l'âge d'or de l'archéologie où le forum romain et d'autres sites majeurs ont été mis au jour, les objets étaient souvent extraits du site et entassés pour être examinés. Seules quelques pièces étaient cataloguées ; il est probable que beaucoup étaient pillées. On ne conduisait pas d'inventaire exhaustif, pas plus qu'on ne notait soigneusement les positions d'origines des artefacts. Au ^{xx}^e siècle encore, on ne consignait les informations sur l'emplacement des objets qu'*a minima*. Les archivistes continuaient de recevoir des boîtes de photographies, de diapositives et d'autres formes de preuves sans légende sur des sites aussi importants que l'île de Pâques. Dans les bonnes pratiques actuelles, l'emplacement de chaque objet est noté méticuleusement, de même que de nombreux détails sur les circonstances des fouilles. Consigner les positions absolues et relatives de chaque pièce trouvée est aujourd'hui jugé essentiel pour comprendre leur contexte. Lorsque des vestiges sont découverts par hasard, il est moins probable qu'une documentation détaillée de l'endroit soit constituée. Par exemple, les agriculteurs et agricultrices continuent de trouver des épées de l'âge de bronze en labourant leurs champs ou en exploitant des tourbières (Faoláin et Northover, 1998).

Beaucoup des collections des grands musées du monde ont été rassemblées sans vraiment se soucier du droit des pays à maîtriser leur patrimoine culturel. Jusqu'à la fin du ^{xx}^e siècle, les archéologues, aventurières et aventuriers pouvaient rentrer chez eux avec des cargaisons d'artefacts. Lorsque ces pièces patrimoniales ont pris de la valeur en tant qu'œuvres d'art et objets de collection, le trafic s'est intensifié.

En 1970, l'UNESCO a adopté une convention pour contrôler le commerce des biens culturels. Les dispositions juridiques précises varient selon les pays et s'appliquent différemment aux objets d'art, aux manuscrits, aux spécimens, aux restes humains et ainsi de suite (United Nations Educational, Scientific and Cultural Organization, 1970, 2013). Les scientifiques reviennent désormais chez eux chargés de documents numériques, de photographies, d'images laser et d'autres représentations de leurs observations. Elles et ils constituent les données à partir desquelles ils rédigeront leurs publications de recherche et planifieront leurs prochaines investigations.

La taille compte

Bien que les sciences humaines soient des disciplines mûres représentant des siècles d'érudition, la plupart de leurs projets de recherche n'impliquent qu'un ou deux chercheurs et chercheuses. Leurs sources de données peuvent présenter un gros volume et une très grande variété, mais leur vélocité est généralement faible. De nombreux chercheurs travaillent avec de petits sous-ensembles d'objets et de documents, passant des mois, voire des années, à inspecter et à rapprocher des sources de preuves disparates. Une seule fouille peut laisser des dizaines ou des milliers de traces. Les collections numériques en contiennent des millions. Le scannage en trois dimensions et en haute résolution d'artefacts peut produire des fichiers numériques de plusieurs gigaoctets, voire téraoctets.

Quand est-ce une donnée ?

Quand il s'agit de déterminer ce qui peut devenir une donnée et quand les sciences humaines se trouvent dans une position similaire à celle des sciences sociales. Des domaines longtemps pauvres en données, pour reprendre l'expression de Sawyer (2008), croulent maintenant sous elles. Loin de convoiter un unique manuscrit rare, les chercheurs et chercheuses ont désormais accès à de vastes corpus d'écrits numérisés. Choisir ce qu'on utilisera comme preuve de quel phénomène est aujourd'hui plus difficile. Identifier des entités pouvant servir de données confère aux chercheurs un avantage concurrentiel. Ces entités peuvent être obscures ou se trouver sous nos yeux. Les objets servant de données dans d'autres domaines peuvent trouver un nouvel usage dans les sciences humaines et devenir de nouvelles formes de données. Ainsi, pour les astronomes, les observations des lunes de Jupiter par Galilée mettent en évidence des objets célestes, mais pour les historiennes et historiens des sciences, des religions et de la politique, elles constituent des preuves de la culture de l'époque.

Pour ce qui est du travail scientifique des données, les sciences humaines se distinguent des sciences exactes et sociales sur plusieurs plans. Le premier est la place

primordiale qu'elles accordent à l'interprétation et à la réinterprétation : le boson de Higgs ne peut être découvert qu'une fois, mais le *Hamlet* de Shakespeare peut être réinterprété à de nombreuses reprises. Le deuxième est la vaste variété des sources de données et l'expertise nécessaire à leur interprétation. Pour de nombreux types de travaux, les spécialistes des sciences humaines doivent parler ou lire plusieurs langues, se familiariser avec différents systèmes d'écriture et caractères et maîtriser la culture de nombreuses parties du monde. En effet, ce n'est qu'en connaissant le contexte de chaque objet étudié que l'on peut les comparer. La conséquence de ces deux caractéristiques est que les sources de données peuvent être réutilisées à maintes reprises et gagnent souvent en valeur à mesure qu'elles s'accumulent. Le revers de la médaille est qu'il est difficile d'appliquer des classifications qui transcendent tous les usages possibles des collections, puisqu'un même objet peut être représenté et interprété d'une myriade de façons différentes.

Sources et ressources

Parce que n'importe quoi – ou presque – peut servir à mettre en évidence l'activité humaine, il est extrêmement difficile de délimiter ce qui constitue ou non une possible source de données pour la recherche en sciences humaines. Les musées, les bibliothèques et les archives opèrent ces délimitations en décidant quoi recueillir. Les rôles de ces trois « institutions de mémoire » se recoupent en grande partie, notamment en ce qui concerne les objets numériques. Chaque institution présente et ordonne ses pièces en fonction de ses missions. Sur ces plans, ce sont les bibliothèques qui sont les plus cohérentes, car elles recueillent majoritairement des documents publiés existants en de nombreux exemplaires. Sans être identiques, leurs mécanismes de catalogage et de classification sont suffisamment harmonisés de par le monde pour qu'un même livre soit décrit de manière similaire dans toute bibliothèque de recherche qui en possède un exemplaire. Il n'en va pas ainsi pour les musées et les archives. Deux objets semblables, comme les sculptures d'une danseuse par une même artiste, seront traités différemment selon s'ils sont accueillis par un musée des beaux-arts, un musée de la danse, une collection archivistique de pièces de cet artiste ou un musée culturel sur le pays ou la région dont l'artiste est originaire. De plus, des images numériques des sculptures peuvent être présentées et ordonnées de nombreuses manières dans d'autres collections. Le fait de devenir une source ou une ressource pour les sciences humaines dépendra de la forme de l'entité, de son genre, de sa provenance ou de son degré de transformation par rapport à son état originel.

Objets physiques ou numériques

Les scientifiques qui étudient des objets matériels – papiers, papyrus, reliures, notations, encres, illustrations, technique, structure – ont besoin d'observer, manipuler

ou éprouver les artefacts. Pourtant, peu de bibliothèques, d'archives et de musées leur permettent de toucher des objets rares, précieux ou fragiles. Si le contact est autorisé, ce sera souvent de manière brève et dans des conditions précises. Ainsi, les ouvrages rares sont souvent laissés ouverts à une page donnée pour limiter leur exposition à la lumière. Les tests sur le papier, les métaux et d'autres matériaux nécessitent la coopération des propriétaires de l'objet.

En revanche, pour les savantes et savants qui étudient la teneur des documents, les représentations numériques valent bien les originaux, voire leur sont préférables. En effet, les images numériques peuvent être comparées côte à côte et examinées de près, feuilletées, imprimées et annotées. Les illustrations et des portions de pages peuvent être grossies, offrant des possibilités aux techniques analytiques microscopiques. Les textes et les images peuvent aussi être rendus plus nets, révélant des détails invisibles à l'œil nu.

Numérique ou numérisé

Les ensembles numériques peuvent contenir des contenus qui étaient numériques dès l'origine, comme des livres et des articles de revue rédigés sur ordinateur, et des traces numérisées, qui sont des représentations de documents, d'images, d'enregistrements sonores ou d'autres objets créés sous forme physique ou analogique. La majeure partie des collections en sciences humaines sont constituées de représentations numérisées de ressources physiques et analogiques. On s'attend cependant à ce que cette proportion s'inverse maintenant que davantage de contenus naissent sous forme numérique.

Substituts ou contenu intégral

Jusqu'aux années 1980 et aux progrès réalisés dans l'automatisation, les chercheurs et chercheuses dépendaient des catalogues papier et du courrier postal pour se renseigner sur l'emplacement d'ouvrages, de documents et d'artefacts. Lorsque des substituts numériques, comme des notices bibliographiques, des vignettes, des extraits audio ou vidéo ou d'autres représentations d'objets, sont apparus en ligne, ils ont pu identifier facilement ce qui existait et à quel endroit, ce qui leur a fait gagner beaucoup de temps de trajet et de moyens humains. Toutefois, si les catalogues en ligne et les autres instruments de recherche apportent vitesse, envergure et commodité à la recherche d'informations, les collections de substituts amènent rarement des transformations fondamentales dans la pratique scientifique.

La mutation suivante de la recherche en sciences humaines a consisté à numériser intégralement les contenus, leur permettant d'être diffusés en ligne. Selon les méthodes de numérisation utilisées, ils peuvent être fouillés, combinés, modélisés et

exploités comme des ressources en données. Un nombre énorme de livres ont été scannés et publiés en ligne, de même que des documents historiques, des images et des enregistrements audio et vidéo. Le nombre croissant de ressources anciennes ou obscures disponibles est une aubaine pour les lectorats de pays ou d'institutions n'ayant pas accès aux vastes collections des bibliothèques de recherche et a donc démocratisé la recherche de façon inattendue.

Les substituts demeurent souvent nécessaires à la description, la localisation, la découvrabilité et l'authentification du contenu numérique. C'est surtout les ressources non textuelles qui dépendent de ces données substitutives. En effet, une image ou un extrait audio présente rarement un intérêt pour la recherche si on en ignore l'artiste, la date, la provenance, la propriété et d'autres informations contextuelles.

Images statiques ou représentations interrogeables

La transformation suivante des pratiques de recherche en sciences humaines est survenue avec la diffusion des représentations interrogeables de textes, d'images, d'audio et de vidéo. L'exploitation du contenu numérique est une question de degré, qui va des images statiques aux objets enrichis d'informations sémantiques. Pour comprendre ces distinctions, l'exemple le plus simple est celui de la conversion de texte. Les premières méthodes de représentation mécanique de documents consistaient à saisir une reproduction visuelle de chaque page aussi parfaitement que la technologie de l'époque le permettait. Les photographies et les images numériques de pages ainsi obtenues sont plus faciles à diffuser que les exemplaires papier, mais ne présentent guère de fonctionnalités supplémentaires par rapport à un livre imprimé. Les utilisateurs et utilisatrices ne peuvent repérer les pages qui les intéressent qu'en consultant la table des matières – quand elle existe – ou en feuilletant le document. Une large part des ressources scientifiques ne sont encore disponibles que sous forme d'images de pages, scannées et enregistrées dans les premières versions de PDF (Portable Document Format) ou d'autres formats, dont beaucoup sont propriétaires.

Pour rendre un texte interrogeable, les lettres et les mots individuels doivent être identifiables. Le texte peut être converti en le tapant à la main ou grâce à des technologies de reconnaissance optique de caractère (ROC ou OCR pour Optical Character Recognition). Les techniques de ROC remontent aux systèmes télégraphiques du début du ^{xx}e siècle. Bien que ces technologies aient accompli de grands progrès, elles sont surtout efficaces sur du texte tapé dans des polices modernes sur un fond blanc et propre. Or, les ouvrages et les documents historiques ne remplissent que

rarement ces critères. Convertir du texte imprimé dans les siècles passés reste un défi technique et un domaine de recherche pour la lexicographie et l'informatique.

L'adoption en 1992 d'Unicode, un standard désormais intégré dans tous les principaux systèmes d'exploitation, les écrans d'affichage et les technologies d'impression, rend interrogeable tout texte créé numériquement dans une langue et un jeu de caractères modernes. Pour la conversion de texte, les technologies de ROC contemporaines peuvent identifier des caractères et les convertir en Unicode à condition qu'une correspondance existe (Unicode, Inc., 2013 ; ISO 10646, 2013). La numérisation massive de textes dans le cadre du Million Books Project ou de Google Books est devenue réalisable grâce à Unicode et à la ROC, même si le caractère suffisamment exact de la conversion pour la recherche reste sujet à débat (Duguid, 2007 ; Nunberg, 2009, 2010).

Un texte scanné difficile à convertir par ROC sera interprété grâce au *crowdsourcing* (production participative). Le CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart, « test de Turing complètement automatisé pour distinguer ordinateurs et êtres humains ») est une technique astucieuse utilisée en sécurité web. Lorsque des internautes demandent un compte, un mot de passe ou d'autres privilèges, on leur enjoint de transcrire une série de lettres présentées comme un casse-tête. Environ deux cents millions de CAPTCHA sont résolus chaque jour, ce qui représente 150 000 heures de travail humain. ReCAPTCHA crée les casse-tête à partir de textes pour lesquels la conversion par ROC a échoué. Plus de quarante mille sites web utilisent reCAPTCHA pour transcrire des textes anciens. Ces agaçants casse-tête contribuent donc à l'exactitude de la transcription de texte, exploitant utilement le labeur humain (von Ahn *et al.*, 2008 ; reCAPTCHA, 2013).

Les textes bien antérieurs à l'époque de Gutenberg sont plus difficiles encore à interpréter et nécessitent une intervention humaine. Les scientifiques ont recours au *crowdsourcing* pour convertir des réserves de ressources rares. Ancient Lives, par exemple, est un projet Zooniverse pour transcrire des textes écrits sur des papyrus en Égypte il y a plus de mille ans. On présente aux bénévoles des images de papyrus et on leur demande d'y identifier des signes en marquant leur centre et en les associant à une table de caractères connus. Identifier ne serait-ce que quelques signes ou d'autres caractéristiques d'un papyrus, par exemple ses dimensions, aide le projet scientifique à déterminer la teneur du document (Ancient Lives, 2013).

Chaînes de caractères interrogeables ou contenu enrichi

Une transformation plus radicale encore des méthodes de recherche devient possible lorsque des ressources sont enrichies pour mieux les exploiter. Les sciences

humaines ont été pionnières dans la représentation structurée de documents numériques à travers des entreprises comme la Text Encoding Initiative (TEI) dans les années 1980. Les projets de numérisation de texte suivent encore aujourd'hui les normes de la TEI et les outils qui lui sont associés, avec des variations pour le latin et le grec, la musique et d'autres types de contenus (Text Encoding Initiative, 2013).

Dans les sciences humaines comme dans les sciences exactes, la capacité à traiter tout un corpus de texte comme des objets numériques ouvre de nouvelles possibilités pour la recherche. Les chaînes de lettres, d'espaces et de signes de ponctuation interrogeables sont déjà plus utiles que les images de pages. Lorsque des informations sémantiques sur des mots, des phrases, des titres, des chapitres, des sauts de page, des noms de personnes, des lieux, des citations et d'autres éléments sont représentées, l'analyse s'en trouve considérablement enrichie. Par exemple, en texte brut, des mots comme « Paris » ou « brun » peuvent être repérés, alors que dans du texte encodé de manière sémantique, on peut distinguer la ville de Paris, le personnage de la mythologie grecque et le plâtre qui porte ce nom, ou encore le nom de famille Brun et la couleur. De même, des images, des sons et des vidéos peuvent être enrichis grâce aux tags et à l'encodage.

Une fois que les caractères du texte, les notes de la musique et les autres séquences peuvent être traités comme des objets numériques, de nouvelles possibilités informatiques s'offrent aux scientifiques. BLAST, l'outil de recherche d'alignement mentionné au chapitre 5 pour apparier des séquences d'ADN, s'avère précieux pour les sciences humaines : il y est en effet employé pour repérer des textes historiques similaires, qu'ils soient en français du XVIII^e siècle ou sur des papyrus, des séquences sonores semblables dans des collections musicales et de nombreuses autres formes de contenu (Altschul *et al.*, 1990 ; Ancient Lives, 2013 ; Olsen *et al.*, 2011).

Les images peuvent également être enrichies grâce à des techniques d'imagerie au laser ou aux rayons X. Les palimpsestes, c'est-à-dire des documents sur lesquels on a écrit un nouveau texte, sont notoirement difficiles à déchiffrer. L'imagerie permet de séparer les couches d'écriture et d'accentuer et distinguer les gravures dans la pierre et le bois. Les matériaux anciens sont devenus un front de recherche dans des domaines comme l'optique (Knox, 2008 ; Miller, 2007 ; Salerno *et al.*, 2007).

Les infrastructures de la connaissance

L'éventail des collections, corpus et ensembles de données accessibles aux spécialistes des sciences humaines est vaste, étant donné les champs de compétence des musées, des archives, des bibliothèques, des monuments, des sites patrimoniaux et des ressources non institutionnelles de par le monde. Par ailleurs, l'expertise

méthodologique de ces savantes et savants possède une longue et riche histoire. Rassembler toutes ces ressources et ces savoir-faire dans une infrastructure de la connaissance pour soutenir la recherche représente cependant une gageure. La diversité des ressources, leur éparpillement et la multiplicité des modalités d'accès constituent à la fois les forces et les faiblesses des sciences humaines. Même en restreignant la question à la recherche en art et en archéologie antiques, les limites des collections et des expertises restent floues.

Les enjeux des infrastructures de la connaissance dans les sciences humaines en général et dans l'histoire de l'art et l'archéologie antiques en particulier diffèrent de façon importante de ceux des sciences exactes et sociales. Le premier est la distinction entre collections de recherche et collections grand public. Le financement des collections de recherche est minime comparé à celui d'ensembles similaires dans les sciences exactes. En effet, ces derniers, quand ils réussissent à attirer une large communauté, peuvent devenir des ensembles de données ressources, voire de référence, comme la Protein DataBank ou le Sloan Digital Sky Survey (National Science Board, 2005). Il est rare que les collections de recherche des sciences humaines accomplissent cette transition. Les collections grand public, comme celles des musées, des archives et des bibliothèques, présentent un grand intérêt pour la recherche. Néanmoins, chaque institution est investie de sa mission propre, sert sa propre communauté et dispose de son propre organe de gouvernance. La plupart s'adressent à un large public, dont les scientifiques ne représentent qu'une partie. Il est peu probable qu'on parvienne à unifier les collections et les services de ces organismes disparates en une infrastructure commune. On peut toutefois atteindre une certaine interopérabilité grâce à des accords sur les standards et les mécanismes de classification, mais des recoupements multiples restent nécessaires pour répondre aux besoins disparates des usagères et usagers potentiels.

La deuxième différence, qui découle largement de la première, est le faible investissement direct réalisé dans les infrastructures des sciences humaines, comparé aux sciences exactes et sociales. Elles disposent de bien moins de fonds et les subventions qu'elles octroient sont donc plus modestes et de plus courte durée. La diversité des collections, des questions et des méthodes participe aux difficultés dans l'élaboration d'outils communs et fait de la pérennité une gageure perpétuelle (Borgman, 2009 ; Burdick *et al.*, 2012 ; Zorich, 2008). Les départements universitaires en sciences humaines tendent aussi à bénéficier de moins de support technique et leurs personnels de recherche à disposer de moins de savoir-faire technique. Les exceptions sont bien sûr nombreuses : des poches d'expertise technique poussée existent dans certaines universités et un nombre croissant de

pôles d'humanités numériques fournissent un soutien infrastructurel, au moins à l'échelle locale (Parry, 2014).

Alors que, dans les sciences exactes, le partage de connaissances va souvent de pair avec des ensembles de données communs, dans les sciences humaines les communautés et les collections en ligne sont souvent plus indépendantes. Les chercheurs et chercheuses partagent néanmoins leur expertise et bâtissent des communautés au travers de réseaux sociaux comme H-Net (Humanities and Social Sciences Online). Ces sites permettent la discussion, l'examen, les offres d'emploi, les annonces de colloques et d'autres échanges d'informations.

Métadonnées

Les métadonnées fournissent du contexte. Mais le contexte de qui ? Cette question est particulièrement litigieuse dans les sciences humaines. Pour les mêmes raisons qui ont conduit les approches éducatives fondées sur le « canon occidental » de la littérature et les « grands auteurs » (Bloom, 1994) à s'effacer au profit d'un éventail de ressources plus diverses, les méthodes conventionnelles d'organisation des connaissances se voient complétées par des démarches « post-coloniales ». Les méthodes classiques visent à établir un terrain d'entente entre les collections et les pratiques professionnelles. Parmi les exemples les plus connus, on citera les mécanismes de classification élaborés et perpétués par le Getty Trust et des partenaires tels que l'Art and Architecture Thesaurus, l'Union List of Artist Names, le Getty Thesaurus of Geographic Names et la Cultural Objects Name Authority. Ces mécanismes fournissent des métadonnées pour les noms, les lieux, les objets, les styles et leurs relations. L'Art and Architecture Thesaurus est largement utilisé dans les communautés d'histoire de l'art et d'archéologie antiques, mais l'emploi de la terminologie et les relations varient selon les applications (Baca, 1998, 2002 ; Borgman et Siegfried, 1992 ; Getty Research Institute, 2013 ; Harpring, 2010). Le modèle de référence conceptuelle CIDOC, développé depuis plusieurs décennies, est aujourd'hui mis en œuvre au sein des communautés muséales pour l'échange et l'intégration d'informations afin d'encoder les descriptions d'objets. Les projets en histoire de l'art et en archéologie antiques ont été les premiers à l'adopter (Binding *et al.*, 2008 ; International Council of Museums, 2013).

Ce type de mécanismes formels promeuvent la normalisation et l'interopérabilité des systèmes. Ils fournissent la structure rigoureuse nécessaire aux bases de données de conservation des musées, des archives et des bibliothèques. Cependant, la terminologie employée par les spécialistes pour représenter précisément des objets uniques est souvent étrangère aux visiteurs et visiteuses, ainsi qu'aux scientifiques qui posent des questions depuis d'autres points de vue. Les démarches alternatives

cherchent à représenter le patrimoine culturel en utilisant le langage de ceux et celles qui le créent et l'utilisent. Parmi elles, on trouve les descriptions par *crowdsourcing* grâce à l'étiquetage d'images et l'élaboration « ascendante » de mécanismes de classification. Ces efforts pour « décoloniser » la langue du patrimoine culturel risquent, malgré l'intérêt qu'ils présentent en matière de participation du public et de valeur ajoutée pour les collections, de se faire au détriment de la normalisation (Marshall, 2008a, 2008b, 2009 ; Srinivasan *et al.*, 2009a., 2009b ; Srinivasan, 2013 ; Steve: The Museum Social Tagging Project, 2013).

Provenance

La provenance est un concept central des sciences humaines ; mais savoir de quelle provenance on parle pourrait être une question de recherche à part entière. Les scientifiques consignent les origines, la propriété, l'emplacement et les usages des objets, des documents, des textes ou des autres ressources patrimoniales dans le cadre de leur recherche. En histoire de l'art et en archéologie antiques, on distingue l'endroit où quelque chose a été fabriqué, celui où il a été découvert, ses usages et leur évolution. Déterminer la provenance peut impliquer de retracer la composition d'un objet, comme l'origine de la peinture, de la pierre, du métal, du bois ou des outils utilisés pour le fabriquer. D'autres démarches s'efforcent de reconstituer la fabrication d'un objet grâce aux méthodes et aux matériaux connus à l'époque envisagée. En histoire de l'art et en archéologie antiques, les objets de recherche ne sont pas toujours associés à des traces documentaires, voire peuvent être antérieurs à l'invention de l'écriture.

La provenance, les métadonnées et les ontologies sont particulièrement difficiles à établir en histoire de l'art et en archéologie antiques, car le même élément peut être compris de nombreuses manières. Par exemple, une ancienne théière chinoise peut revêtir une importance en raison de sa céramique, de sa conception, de sa fabrication, de son âge ou de son rôle dans l'histoire du thé. Elle peut être remarquable parce qu'elle a été achetée sur la route de la soie ou qu'elle appartenait à une personne particulière, à cause du lieu de sa découverte ou des autres artefacts auxquels elle est associée. Ce ne sont là que quelques exemples de manières dont la provenance peut être consignée ou étudiée pour un objet donné.

Les registres de conservation des collections de musées fournissent des renseignements de base sur les dates d'acquisition ou sur l'origine – quand elle est connue – des objets, mais les informations sensibles sur la propriété, les transferts et les dates et prix correspondants sont rarement incluses dans les documents publics. Les catalogues de vente aux enchères de grandes maisons telles que Christie's et Sotheby's comportent des descriptions détaillées et d'autres éléments de

documentation qui peuvent aider à établir la provenance. Celle-ci peut représenter une question juridique complexe, puisque toutes les parties doivent se soumettre aux conventions de l'UNESCO et aux autres lois en vigueur (UNESCO, 1970). Comme en astronomie, la provenance des objets antiques peut s'avérer une régression à l'infini.

Lorsque la propriété d'un artefact est litigieuse, la provenance passe souvent au premier plan. Sur fond de décennies de batailles juridiques avec la Grèce, le British Museum affirme être le propriétaire légitime des statues du marbre du Parthénon qu'il expose. Près de ces sculptures, le musée affiche des documents étayant l'autorité légale par laquelle il a acquis ces objets au ^{xix}^e siècle. Il juxtapose les statues grecques à celles d'autres civilisations antiques d'Égypte, d'Assyrie et de Perse (British Museum, 2013). Le musée de l'Acropole à Athènes situe des sections de la frise du Parthénon dans le contexte de l'histoire antique de la cité. En 2009, le musée s'est agrandi d'un nouveau bâtiment, comportant un espace d'exposition dix fois plus grand que celui de l'Acropole, dans l'espoir du retour des artefacts retenus depuis bien longtemps à Londres (Acropolis Museum, 2013).

Collections

Les collections, ou ensembles de données, qu'elles soient matérielles ou digitales, sont une composante essentielle des infrastructures de la connaissance de toute discipline. L'histoire de l'art et l'archéologie antiques illustrent le degré de développement de ces ensembles de données pour la recherche, le grand public ou comme éléments d'une infrastructure.

Les organismes de financement soutiennent le développement de certains ensembles de recherche, soit pour eux-mêmes, soit en tant que sous-produits de la recherche. Pour rester compétitifs, les chercheurs et chercheuses soumettent des projets qui prennent délibérément des risques avec de nouveaux médiums et formes d'arguments. Cependant, il est rare que les financeurs puissent faire des choix stratégiques sur les contenus à numériser pour le bien commun. Leur mode de décision est plutôt réactif : ils accordent des subventions aux meilleurs dossiers reçus chaque cycle. Par conséquent, la constitution de collections numériques est davantage orientée par l'offre que par la demande. Les communautés n'expriment que rarement d'une même voix leurs besoins en matière de collections et d'infrastructure, contrairement à l'astronomie.

Les collections de recherche qui bénéficient d'une forte mobilisation disciplinaire deviennent parfois des ensembles de référence. La Cuneiform Digital Library a ainsi survécu, dans les années 1990, à sa transition de projet de bibliothèque numérique subventionnée à un système opérationnel bénéficiant d'un soutien institutionnel

continu. Des tablettes cunéiformes autrefois contiguës sont désormais éparpillées dans des collections du monde entier. Elles sont cataloguées par type, genre, emplacement, période et autres caractéristiques. Les scientifiques peuvent étudier individuellement les objets ou en reconstituer d'autres numériquement grâce aux outils et aux ressources offerts par cette bibliothèque numérique (Cuneiform Digital Library Initiative, 2013).

CLARIN (Common Language Resources and Technology Infrastructure), un projet européen de recherche en linguistique de corpus, travaille aussi sur la problématique globale des collections, de l'infrastructure et de la constitution de communautés. Il vise un « accès facile et pérenne » aux « données numériques sur la langue sous forme écrite, orale, vidéo ou multimodale ». Son action comprend le développement technologique, l'implication des utilisateurs et utilisatrices et des services interopérables à partir d'une fédération de référentiels de données en Europe (CLARIN European Research Infrastructure Consortium, 2013).

Les organismes de financement en lettres et sciences humaines se sont intéressés plus directement aux questions d'infrastructure et de pérennité après plusieurs échecs flagrants. Ainsi, l'Arts and Humanities Research Council (AHRC) au Royaume-Uni a créé l'Arts and Humanities Data Service (AHDS) pour archiver des données issues de projets de recherche sur fonds publics et rendre leur réutilisation possible. Le système permettait la création et le dépôt de ressources numériques et fournissait plusieurs guides de bonnes pratiques. L'AHDS se préoccupait de la préservation des ressources en données plutôt que de la pérennisation d'environnements en ligne permettant de les interroger. Le financement a été interrompu en 2008 ; aujourd'hui, le site et une bonne partie de la documentation sont toujours en ligne, mais les ensembles de données ne sont plus interrogeables. La fermeture de l'AHDS a jeté le doute sur la volonté des organismes de financement et des communautés scientifiques de maintenir de telles ressources. Elle a soulevé des questions sur la sauvegarde de contenus d'arrière-plan dans des archives fermées, la pérennisation de sites web et de services actifs, la pérennité académique opposée à la pérennité technique, la sauvegarde opposée à la conservation et les utilisations et la demande pour ces collections (Arts and Humanities Data Service, 2008 ; Robey, 2007, 2011).

Une autre initiative prometteuse en matière d'infrastructure qui n'a pas survécu est le projet Bamboo, financé de 2008 à 2012 par la fondation Andrew W. Mellon. Il s'agissait d'un investissement significatif dans la constitution de communautés, d'outils et de services pour les sciences humaines. À l'heure où nous écrivons, le projet est officiellement à l'arrêt : le site web et le wiki sont archivés et le code source est diffusé grâce à un référentiel public.

Le souci de l'infrastructure et de la pérennité des organismes de financement dans les sciences humaines se manifeste dans les nouvelles exigences pour les plans de gestion des données ou les plans techniques. Les chercheurs et chercheuses doivent préciser comment l'accès aux données et aux technologies associées, ainsi que leur réutilisation, sera maintenu et pour combien de temps. Bien que toutes les collections et technologies n'aient pas vocation à garder indéfiniment une valeur, planifier leur continuité ou leur fin est devenue une étape des demandes de subvention (Arts and Humanities Research Council, 2012 ; McLaughlin *et al.*, 2010). Des initiatives telles que ConservationSpace, qui est financé par la fondation Mellon, apportent leur aide dans la gestion de documentation pour la conservation de ressources (ConservationSpace, 2013).

Une brève rétrospective des premiers ensembles numériques illustre quelques enseignements tirés en matière d'infrastructure. Le projet Gutenberg, les archives Beazley, le projet Perseus et Valley of the Shadows ont été lancés dans les années 1970 et 1980. Tous ont démarré comme des collections de recherche menées par un ou plusieurs scientifiques. Michael Stern créa les premiers « *e-books* » en 1971 en lançant le projet Gutenberg afin de donner accès à des livres en ligne et gratuitement, grâce aux ordinateurs de l'université de l'Illinois et à des bénévoles qui tapent les textes. Bien qu'académique, le projet ne pouvait numériser que des contenus tombés dans le domaine public. Or, il s'agissait rarement des versions de référence utilisées par les scientifiques. Le projet Gutenberg se veut aujourd'hui avant tout une source de livres électroniques gratuits et est financé par des dons, des subventions et d'autres partenariats (Hart, 1992, 2013).

Les archives Beazley, créées en 1979 sur des ordinateurs centraux et dirigées par Donna Kurtz et d'autres salariés de l'université d'Oxford, se concentrent sur les représentations numériques d'objets d'art antiques. Conçu pour répondre aux besoins des écoles et du public, il a fait partie des collections récompensées par le Queen's Anniversary Prize en 2009. En croissance et progrès perpétuels, les archives Beazley sont devenues le cœur du projet CLAROS, qui s'intéresse à l'infrastructure autant qu'aux collections (Burnard, 1987 ; Hockey, 1999 ; Kurtz et Lipinski, 1995 ; Moffet, 1992 ; Robertson, 1976).

Le projet Perseus, mené par Gregory Crane de Tufts University depuis 1985, a démarré comme une base de données Hypercard de textes grecs et romains. En 1994, le système et les services avaient été soumis à des évaluations sociotechniques (Marchionini et Crane, 1994) et, en 1995, ils ont migré sur le Web naissant. Bien des générations technologiques plus tard, on le connaît désormais sous le nom de Perseus Digital Library, laquelle inclut des ressources artistiques, archéologiques et historiques de

nombreuses époques et dans bien des langues. Crane et ses collègues poursuivent, grâce aux ressources de Perseus, leurs recherches sur l'Antiquité, ainsi que sur les technologies des bibliothèques numériques (Crane et Friedlander, 2008 ; Smith *et al.*, 2002 ; Perseus Digital Library, 2009).

Valley of the Shadow a été développé par Edward Ayers entre 1993 et 2007 pour accompagner ses recherches sur la guerre de Sécession et servir de ressource scientifique et pédagogique publique. Le système fournit un accès numérique à des ressources archivistiques, que de nouvelles formes de présentation et de consultation enrichissent. Ayers, ses collègues et ses élèves ont utilisé ces ressources pour rédiger des publications, dont un ouvrage récompensé par la bourse Bancroft, mais d'autres scientifiques y ont aussi puisé pour leurs travaux (Ayers et Grisham, 2003 ; Ayers, 2003, 2007). Bien qu'Ayers ait changé d'établissement depuis bien longtemps, l'université de Virginie a investi deux ans et environ 100 000 dollars pour conserver ces ressources, afin qu'elles restent accessibles et continuent de servir la recherche (Howard, 2014).

Europeana et la Digital Public Library of America (DPLA) représentent deux exemples d'efforts stratégiques pour construire une infrastructure par l'agrégation d'ensembles numériques disparates, qui concernent tous deux les ressources patrimoniales (Europeana, 2013 ; Digital Public Library of America, 2013). Elles comprennent du contenu tiré de bibliothèques, de musées et d'archives, bénéficient de financements multiples et couvrent un large champ. Europeana rassemble des ressources de pays européens et la DPLA des collections basées aux États-Unis ; toutes deux ont noué des partenariats internationaux. L'Union européenne est l'un des principaux financeurs d'Europeana, tandis que la DPLA est une association à but non lucratif soutenue par des dons de fondations publiques et privées. Elles comportent des collections importantes qui peuvent servir à la recherche, mais aucune des deux n'aspire à être un ensemble de recherche pour une communauté particulière.

Les influences extérieures

Dans les sciences humaines, l'accès aux ressources en données est particulièrement contraint par des facteurs externes. L'histoire de l'art et l'archéologie antiques peuvent être un terrain miné pour les personnels de recherche. Les artefacts sont souvent anciens, précieux et font l'objet de conflits de propriété. Leur gestion est régie par quantité de lois et de politiques, qui se contredisent souvent entre juridictions. Les parties prenantes sont nombreuses et en concurrence : scientifiques, universités, écoles, bibliothèques, musées, archives, gouvernements, propriétaires privés, marchands et grand public. Certains de ces problèmes économiques, éthiques et de propriété demeurent insolubles, quand d'autres font l'objet de négociations.

Économie et valeur

Comme dans d'autres domaines, la valeur d'une ressource donnée dans les sciences humaines dépend souvent de son « emballage ». Certains objets antiques possèdent une grande valeur scientifique, d'autres une grande valeur financière, certains les deux et d'autres aucune. Les spécialistes de l'histoire de l'art et de l'archéologie antiques travaillent dans des économies multiples et en concurrence. Les ensembles de données développés à des fins scientifiques et ouverts à tous, comme le projet Perseus, les archives Beazley, Valley of the Shadow, l'Arts and Humanities Data Service, l'Archaeological Data Service et les ressources linguistiques de CLARIN ERIC, peuvent être considérés comme des réservoirs communs de ressources. Ils sont plus ou moins administrés par des membres de la communauté ; la pérennité et les passagers clandestins y posent problème. Il en va de même pour les collections rassemblées pour le grand public, comme la DPLA et Europeana. Une part importante de ces ensembles sont constitués de substituts numériques d'objets matériels et de textes dont les droits sont réservés. Ces objets et textes peuvent appartenir à des institutions publiques ou privées ou à des particuliers ; ce sont eux qui décident qui peut y accéder et à quelles conditions. Les scientifiques ont de plus en plus accès aux représentations d'objets, mais pas nécessairement aux objets eux-mêmes. Selon leurs méthodes et leurs questions de recherche, cette situation peut s'avérer plus ou moins satisfaisante.

L'archéologie se dirige plus résolument vers l'accès ouvert aux données que la plupart des autres disciplines des sciences humaines. Ainsi, l'Archaeological Data Service, qui faisait originellement partie de l'AHDS, n'a pas fermé en 2008 avec lui, ce qui montre la demande pour ces collections. Certains organismes de financement britanniques exigent que les données soient versées dans ces ensembles et d'autres le recommandent. La communauté archéologique internationale a élaboré des principes directeurs pour la gestion et le dépôt des données et les scientifiques versent leurs matériaux de recherche dans cette collection, ainsi que dans d'autres (Archaeology Data Service, 2013 ; The Digital Archaeological Record, 2013 ; *Journal of Open Archaeology Data*, 2013 ; Open Context, 2013 ; Kansa, 2012 ; Kansa *et al.*, 2007).

L'accès aux données, aux collections matérielles et aux objets numériques est cependant inégal. Bien des ressources qui seraient utiles aux spécialistes de l'histoire de l'art et de l'archéologie antiques peuvent être considérées comme des biens de club et ne sont accessibles qu'à des individus répondant à certains critères. Ainsi, les États réservent souvent l'accès aux fouilles, aux sites, aux artefacts et aux documents à des scientifiques, personnels de recherche et fonctionnaires nationaux ou à des équipes menées par leurs ressortissants. Les embargos – explicites ou non – peuvent durer indéfiniment, jusqu'à ce que les autochtones soient

prêts à publier leurs résultats ou aient terminé leurs travaux. Obtenir des rapports de recherche sur des sites archéologiques peut être très difficile ; des artefacts, encore plus.

Comme pour tout champ de recherche, accéder aux logiciels nécessaires à l'interprétation, à l'analyse ou la réutilisation de données peut s'avérer problématique. Le caractère généralisable des outils est en effet limité par la diversité des ressources et les nombreuses manières dont on peut les employer. Chaque collection et chaque projet tend à développer des logiciels maison. Même si les scientifiques divulguent leur code, maintenir le logiciel peut être plus difficile encore que pérenniser les collections correspondantes. Les difficultés de conservation de données disparates, des logiciels associés et des mécanismes d'interopérabilité semblent avoir contribué à la disparition de l'Arts and Humanities Data Service, du projet Bamboo et d'autres entreprises louables (Borgman, 2009 ; Robey, 2007, 2011, 2013 ; Unsworth *et al.*, 2006 ; Zorich, 2008).

Droits de propriété

La recherche en sciences humaines en général et en histoire de l'art et en archéologie antiques en particulier est profondément empêtrée dans les questions de droits de propriété. Dans ces domaines, les droits sur les données sont liés à la propriété des objets matériels et, parfois, de leurs représentations. Par exemple, les scientifiques peuvent contrôler leurs observations sur un objet ou un phénomène, mais ne pas avoir le droit d'en reproduire des images dans une publication. Des droits de propriété multiples peuvent s'appliquer à un même objet. Ainsi, les droits sur une photographie concerneront la ou le propriétaire de l'objet, la ou le titulaire des droits de reproduction, la ou le photographe et la maison d'édition qui publie la photographie. La réticence de certains musées à autoriser la reproduction d'images par les chercheurs et chercheuses entrave la numérisation de l'édition dans le monde de l'art (Whalen, 2009). Cependant, des initiatives récentes de versement d'images au domaine public pourraient grandement aider les scientifiques (Getty Trust, 2013). Maintenant que la notion de donnée arrive sur le devant de la scène des sciences humaines, les droits sur les données et les représentations pourraient devenir le prochain front des conflits de propriété.

Éthique

Il existe plusieurs sortes de considérations éthiques en histoire de l'art et en archéologie antiques ; l'une d'elles est juridique. Les chercheurs et chercheuses sont en effet censés connaître les réglementations en vigueur et négocier de manière déontologique avec les musées, les conservateurs et conservatrices, les propriétaires, les autorités locales et les autres parties prenantes. Une considération connexe est celle du respect

des objets. Les scientifiques se doivent de limiter les dommages occasionnés aux sites de recherche et aux artefacts. Par exemple, tester des métaux implique un échantillonnage méticuleux dans des recoins de l'objet.

Il n'est pas rare d'observer des « périodes d'utilisation exclusive » de fait dans les domaines des sciences humaines qui étudient des textes rares et des produits culturels. Une fois que les chercheurs et chercheuses ont obtenu un accès aux ressources, ils souhaitent parfois les explorer en privé jusqu'à être prêts à publier. Les États peuvent encourager ce comportement, préférant que leurs propres scientifiques aient la primeur du patrimoine local au détriment de la communauté internationale. L'affaire des manuscrits de la mer Morte est un cas bien connu de contrôle local, souvent assimilé à de la rétention. Découverts dans le désert de Judée dans les années 1940, ils ont été conservés par des scientifiques israéliens jusqu'en 1991 et divulgués seulement sous la pression, à la suite de la diffusion d'un microfilm réalisé à des fins de conservation (Schiffman, 2002). Les manuscrits sont aujourd'hui plus largement accessibles, mais leur interprétation fait toujours l'objet de débats acrimonieux (Leland, 2013).

Mener des recherches en histoire de l'art et en archéologie antiques

Ces études de cas mettent en perspective, d'une part, la création d'une infrastructure pour l'histoire de l'art et l'archéologie antiques et, d'autre part, la recherche dans un domaine où très peu de ressources informationnelles existent sous forme numérique. CLAROS, « le monde de l'art sur le Web sémantique », construit une infrastructure à l'aide d'outils, de technologies et d'ensembles numériques conçus pour la recherche, l'éducation et l'usage par le public. Le projet – dont le nom était l'acronyme de Classical Art Research Online Services, « services en ligne de recherche en art antique » – est une émanation des archives Beazley, qui avaient commencé à numériser leurs collections d'antiquités à la fin des années 1970. CLAROS est basé à l'université d'Oxford, qui accueille quatre des plus belles collections universitaires au monde – l'Ashmolean Museum, le musée d'histoire naturelle, le musée d'histoire des sciences et le Pitt Rivers Museum –, et constitue un banc d'essai pour la recherche sur les bases de données multimédias. Conçu sur un logiciel *open source*, il comprend des outils d'intégration de données développés pour la zoologie, de reconnaissance d'image pour le génie et d'intelligence artificielle pour les services web. Les collections sont tirées de musées, de monuments et d'autres institutions patrimoniales d'Europe et d'Asie. Par ailleurs, CLAROS expérimente actuellement le modèle de référence conceptuel CIDOC de la communauté muséale afin de permettre l'interopérabilité entre ces ensembles

(University of Oxford, 2013b ; Kurtz *et al.*, 2009 ; University of Oxford Podcasting Service, 2011 ; Rahtz *et al.*, 2011).

Le griffon de Pise est un polar scientifique qui illustre les obstacles à la découverte d'information sur un objet matériel. Le griffon est une grande statue de bronze d'un animal mythique qui est restée perchée au sommet de la cathédrale de Pise, en Italie, pendant au moins trois siècles (voir figure 7.1). En retraçant sa provenance, une historienne de l'art en a trouvé une référence visuelle datant de la fin du ^{xv}^e siècle. D'autres éléments indiquent que le griffon a été retiré de son piédestal en 1828 avant d'être remplacé par une copie moderne. L'original se trouve aujourd'hui dans un musée proche de la cathédrale de Pise. Malgré plusieurs siècles de recherche en art, en histoire, en archéologie et dans d'autres disciplines, les origines du griffon, sa provenance et sa fonction font toujours débat. Les Pisans auraient pu l'acquérir auprès d'artistes locaux, dans un butin de guerre ou par d'autres moyens (Contadini *et al.*, 2002 ; Contadini, 2010 ; The Pisa Griffin Project, 2013).

Figure 7.1. Le griffon de Pise

Crédit : Gianluca de Felice, du Pisa Griffin Project.

Les questions de recherche

Les chercheurs et chercheuses en sciences humaines ont davantage tendance à problématiser un objet ou un phénomène qu'à poser des questions de recherche ou des hypothèses à éprouver. Néanmoins, leurs activités de recherche comportent des phases itératives comparables à celles des sciences exactes et sociales.

Si la mission de CLAROS est vaste, son objectif de recherche central est de développer une plateforme technologique qui puisse être utile à un éventail de collections patrimoniales. Il s'agit d'un système fédéré et donc facilement extensible par d'autres partenaires. La plupart des traces répertoriées dans CLAROS comportent des images et de brefs descriptifs textuels, dont l'identifiant unique attribué par l'institution. Pour l'heure, chaque sous-ensemble est indépendant, mais il est prévu d'ajouter une fonction pour lier des objets entre les ensembles de CLAROS et des collections externes.

La recherche sur le griffon de Pise vise à comprendre ses origines, sa provenance, sa valeur culturelle et ses diverses fonctions au cours de son existence. En somme, qu'est-il, d'où vient-il et comment est-il arrivé là ? Des artefacts antiques tels que le griffon de Pise n'ont pas une forme, une fonction ou un rôle statiques. Il pourrait avoir des origines islamiques ou chrétiennes et peut-être avoir été un objet sacré. L'hypothèse la plus probable – et la plus intrigante – sur sa fonction serait l'émission « de sons étranges et fascinants » lorsque le vent soufflait dans sa cavité depuis le sommet de la cathédrale (Contadini *et al.*, 2002, p. 69). Afin d'éprouver les hypothèses et d'en trouver d'autres, on a procédé à des analyses métallurgiques sur le griffon. Il semblerait qu'il ait servi de cible de tir, car il porte des impacts de balle ; aussi des analyses balistiques sont-elles prévues. Grâce à la datation au carbone, l'année de création a pu être estimée entre l'an 1020 et 1250 de notre ère. Le cuivre vient certainement de Chypre, mais l'objet a été sculpté dans le sud de l'Italie, en Égypte ou en Espagne. Le griffon de Pise a connu un regain d'intérêt lorsqu'un lion de style semblable est apparu sur le marché de l'art new-yorkais au début des années 1990 (Contadini *et al.*, 2002 ; Contadini, 2010). L'une des raisons qui rendent son étude complexe est qu'il est difficile à classer. Des objets similaires sont éparpillés dans des musées des beaux-arts, d'arts appliqués, d'archéologie, d'histoire naturelle, d'histoire locale et d'anthropologie.

La collecte de données

En histoire de l'art et en archéologie antiques, de nombreuses activités peuvent être considérées comme de la collecte de données. Une première dimension de la collecte pour CLAROS est l'acquisition de contenus pour les bases de données, une seconde est le recueil d'informations sur les communautés d'utilisateurs des collections actuelles et potentielles et une troisième est la veille sur les technologies

qui pourraient être appliquées à ces collections pour servir ces communautés. Ces dimensions interagissent de plusieurs façons.

CLAROS ne peut recueillir que les traces que ses partenaires sont prêts à verser. Leur forme, leur teneur et leur intégrité varient grandement. Certaines institutions participantes ont accumulé leurs collections sur plusieurs siècles : la conversion de ces traces dure depuis trente ans, au cours desquels les normes, les pratiques et les technologies ont considérablement évolué. Elles comportent donc différents types et quantités d'informations sur l'origine, l'acquisition et la conservation des objets. La structure des traces et les mécanismes de classification varient aussi considérablement. CLAROS a commencé par intégrer deux millions de ces traces sur CIDOC, l'ontologie et la structure de données élaborée par la communauté muséale internationale. Ce projet est l'un des premiers à appliquer CIDOC à grande échelle (International Council of Museums, 2013 ; Kurtz *et al.*, 2009).

Les études sur les utilisateurs et les évaluations de système telles qu'elles se pratiquent en sociotechnique sont inhabituelles dans les sciences humaines, bien qu'un tel travail soit envisagé dans l'avenir de CLAROS. Faute d'études, l'interface utilisateur du banc d'essai a été créée en partant de ce que l'on savait de l'usage actuel des collections. Outre la recherche usuelle par mots-clés, l'ajout d'une fonction pour filtrer les vases grecs par forme et par couleur a ajouté de la valeur à la collection Beazley. Les partenaires issus de la zoologie, de l'informatique et du génie ont adapté leurs technologies aux nouveaux ensembles et aux publics envisagés. L'objectif était de garder une conception suffisamment ouverte pour pouvoir l'adapter si nécessaire à d'autres contenus et publics.

La recherche sur le griffon de Pise combine des méthodes et des savoir-faire de l'histoire de l'art et de l'archéométallurgie : la collecte de données dans un domaine informe l'autre et vice versa. L'historienne de l'art a rassemblé des éléments sur la sculpture à partir de traces historiques et de travaux antérieurs. Ses sources comprennent des dessins, des sculptures sur bois, des gravures, des traductions d'inscriptions et des analyses stylistiques (Contadini *et al.*, 2002 ; Contadini, 1993). L'archéométallurgiste a prélevé des échantillons à différents endroits sur la sculpture afin d'identifier les métaux de chaque partie de l'objet et les méthodes qui ont permis de les joindre. Pour déterminer quels échantillons prélever et comment les tester, il s'appuie sur les analyses historiques.

Une grande partie des informations recherchées par l'archéométallurgiste concernent l'utilisation des métaux à ces époques et dans ces régions. Son travail est contraint par le manque de ressources numériques en ligne et par l'exclusivité de nombreuses

traces pertinentes. Les analyses métallurgiques ne sont pas des tâches routinières, mais elles peuvent avoir été réalisées dans le cadre de la conservation, de l'évaluation, de la recherche de provenance ou de la contestation de l'authenticité. Lorsque des rapports d'analyses antérieures existent, ce n'est parfois que sur papier ou microfilm. Des traces concernant d'autres objets susceptibles d'éclairer les recherches de l'archéométallurgiste sont retenues dans les collections personnelles de propriétaires privés, de marchands et de scientifiques, y compris des archéométallurgistes ayant prélevé des échantillons. Le partage d'information repose largement sur les échanges interpersonnels. L'archéométallurgiste qui étudie le griffon de Pise a contacté des spécialistes par e-mail, téléphone et autres moyens afin de repérer des analyses métallurgiques pertinentes sur des objets similaires. Il a également beaucoup voyagé pour consulter des documents sur papier et microfilm.

L'analyse des données

C'est au stade de l'analyse que les différences de méthode pour la représentation, la classification et l'organisation de l'information sont les plus manifestes. Le projet CLAROS a intégré une panoplie de traces issues des différentes institutions participantes dans une structure commune, élaboré des ontologies pour organiser les descriptions sémantiques d'objets et conçu des technologies pour exploiter les données résultantes. L'avantage d'encoder les descriptions d'objets avec le modèle de référence conceptuelle CIDOC est que celui-ci peut intégrer toutes les métadonnées possibles plutôt que de réduire les traces à leur plus petit dénominateur commun. L'architecture en « Web de données » utilisée en bio-informatique a été mise en place par des zoologues spécialisés en taxonomie. Elles et ils s'étaient en effet rendu compte que les problèmes sémantiques en biologie et en art antique étaient similaires. Ainsi, en anglais, les musées utilisent les termes *artist* et *creator* de façons différentes ; les collections de biologie ont quant à elles tendance à privilégier le terme *insect* à son synonyme *bug*. D'autres systèmes de classification distinguent clairement entre la ville de Paris et le plâtre du même nom, alors que les biologistes veillent à différencier les mots « mouche » et « voler », qui se disent tous deux *fly* (University of Oxford, 2013b ; Kurtz *et al.*, 2009 ; Rahtz *et al.*, 2011). L'interopérabilité de CLAROS s'est heurtée à des barrières technologiques dues à la disparité des formats anciens à intégrer. Les traces d'ores et déjà converties dans des formats de base de données XML et SQL étaient les plus simples à absorber. L'opération a été particulièrement difficile pour les traces destinées à des usages spécifiques, comme les bases de données relationnelles utilisées pour la composition d'ouvrages imprimés.

À la suite de l'analyse des données sur le griffon de Pise, il est apparu qu'il valait mieux appréhender cet artefact comme un amalgame d'objets différents. En effet, des ornements et d'autres éléments semblent avoir été ajoutés et supprimés au

cours des siècles depuis la réalisation originelle de la sculpture. Ainsi, des métaux de nombreuses provenances ont pu être incorporés au griffon que nous connaissons. L'historienne de l'art a retracé dans le temps les mentions du griffon de Pise et d'objets comparables pour tenter de le situer dans des contextes multiples et mouvants (Contadini, 2010). Les analyses métallurgiques sur la microstructure des échantillons ont révélé les différents minerais et la composition des alliages : cela aide à déterminer la mine d'où ils proviennent, les dispositions d'impuretés – qui révèlent la quantité de ferraille ajoutée –, l'époque approximative de création de l'objet et la méthode de fabrication probable. D'autres analyses du griffon comprennent des visualisations en deux et trois dimensions grâce au scannage numérique.

Une fois au laboratoire de métallurgie, les échantillons sont préparés, polis et soumis à des tests de dureté et de composition. Ces tests et analyses microscopiques sont choisis en fonction d'hypothèses sur la composition des métaux et des limites de détection de chaque méthode. Certaines sont destructives, tandis que d'autres permettent de conserver l'échantillon. Les méthodes les moins destructrices sont préférables, puisque les échantillons peuvent être réanalysés avec de nouvelles technologies ou pour éprouver de nouvelles hypothèses. Les techniques actuelles permettent déjà une résolution en deçà du nanomètre, certaines analyses atteignent même le picomètre ou l'angström.

Afin d'établir les origines des matériaux le plus précisément possible, l'analyse est approfondie grâce à l'examen d'archives, de brevets, de documents muséaux, de registres de conservation et de catalogues de ventes aux enchères. Savoir de quelle mine le métal a été extrait est nécessaire, mais insuffisant ; tout aussi importantes sont les connaissances sur le commerce autour des mines à l'époque présumée, sur l'utilisation du métal par les artisans et artisans de la région où l'objet aurait été fabriqué, sur les influences sociales et politiques sur l'échange d'objets – telles que les routes commerciales et le pillage de guerre – et toute autre information contextuelle sur l'objet. Chaque progrès technologique, chaque avancée en matière d'accès aux ressources informationnelles permet d'obtenir de meilleures données. Ainsi, des artefacts culturels tels que le griffon de Pise peuvent être réinterprétés indéfiniment à mesure que l'on en apprend sur eux.

La publication des résultats

Les pratiques de publication de l'histoire de l'art et de l'archéologie antiques sont caractéristiques. Les collaborateurs et collaboratrices rapportent bien leurs découvertes à leurs communautés respectives, mais la structure des publications à plusieurs mains est inhabituelle. En effet, dans la plupart des disciplines, les travaux cosignés présentent un récit unique. Or, l'article de synthèse sur CLAROS, qui est une

communication de colloque, est divisé en sections personnelles. Chacune est signée par l'un des auteurs et autrices ; seul l'un d'eux n'a pas sa propre section. Les autres collaborateurs et collaboratrices sont mentionnés nommément dans la première section, plutôt que dans les remerciements en fin de communication ou dans les notes de bas de page (Kurtz *et al.*, 2009).

Les recherches sur le griffon de Pise sont rapportées dans plusieurs publications et signées par différentes combinaisons d'auteurs et autrices. L'étude initiale sur le griffon de Pise est l'œuvre d'Anna Contadini, la principale historienne de l'art du projet, et a été publiée en italien dans un catalogue d'exposition. Ses différents articles sur cette statue ont été rédigés en anglais et en italien (Contadini, 1993, 2010). Un article de synthèse comparant le griffon de Pise et le lion de New York obéissent au même format compartimenté que celui sur CLAROS (Contadini *et al.*, 2002). Les trois premières sections traitent du contexte historique et artistique : deux ont un seul auteur et une en a plusieurs. La dernière section, signée par Peter Northover, rend compte des analyses métallurgiques de manière détaillée, listant la composition des échantillons des différentes parties des deux objets, ainsi que d'un troisième qui serait lié au lion et au griffon. Les remerciements sont listés dans les notes de chaque section de l'article et la bibliographie est commune. On trouve également des images des objets étudiés. Dans les deux cas, les mérites de chacun dans la collaboration sont soigneusement répartis, ce qui est peut-être un héritage de la tradition de l'auteur solitaire dans les sciences humaines.

La conservation, le partage et la réutilisation des données

Les difficultés de conservation, de partage et de réutilisation des données en histoire de l'art et en archéologie antiques proviennent de la diversité des entités pouvant être considérées comme des données dans ces situations, ainsi que des intérêts concurrents des parties prenantes.

CLAROS se rapproche davantage d'un banc d'essai que d'une collection. En tant que banc d'essai, il est conçu pour être ouvert et interopérable avec des technologies de Web sémantique qui permettent à des tiers d'ajouter des mécanismes de classification, des interfaces utilisateur et d'autres fonctionnalités. En tant que collection, son ouverture permet de trouver un nouvel usage aux traces. La pérennité de CLAROS dépend de la continuité du financement, de l'engagement renouvelé des partenaires actuels et du développement de nouveaux partenariats ; or, tous ces éléments se sont avérés difficiles à garantir. Parce qu'il est peu probable qu'une institution remplace sa structure de données interne pour rejoindre le projet, CLAROS a choisi d'exporter les données de chaque système partenaire dans le modèle de référence conceptuelle CIDOC. Cette démarche se prête bien à un banc d'essai d'infrastructure, mais a

peu de chances de motiver les institutions partenaires à harmoniser leurs systèmes. Les structures maison doivent généralement s'intégrer à des systèmes internes de gestion des finances, du personnel et des collections des archives et bibliothèques, ainsi qu'avec des systèmes tournés vers l'extérieur, comme les sites web publics, les visites virtuelles de musées, les collections fédérées telles que CLAROS et de grands services fédérés comme Europeana et la Digital Public Library of America.

Les informations au sujet du griffon de Pise sont diffusées bien plus largement que ce que l'on observe habituellement dans d'autres domaines de recherche. Étant un objet imposant, unique et doté d'une grande importance historique, le griffon dispose de son propre site web, où figure la documentation du projet. Les recherches à son sujet peuvent éclairer des travaux dans de nombreux domaines des arts, de l'archéologie, de l'histoire, de la religion, de la culture, du commerce et de la métallurgie. Les publications constituent les traces durables de ces recherches. Les documents historiques sur lesquels elles sont fondées demeurent dans les bibliothèques, les archives, les services de conservation et les autres endroits où ils se trouvaient. Le projet crée des traces numériques volumineuses : scanographie, imagerie laser et photographie en haute résolution peuvent produire des fichiers de plusieurs gigaoctets, voire téraoctets. Ceux-ci ne sont pas diffusables tels quels et ces pratiques de partage ne sont pas courantes dans cette communauté (Contadini *et al.*, 2002 ; Contadini, 2010 ; The Pisa Griffin Project, 2013).

Dans les publications, les données issues de l'échantillonnage des métaux du griffon sont rapportées sous forme de tableaux, comme il est d'usage en métallurgie scientifique. Ces tableaux sont ouverts au sens où le lectorat y accède librement, mais pas au sens où ils consisteraient en des données structurées découvrables ou exploitables informatiquement (Murray-Rust et Rzepa, 2004). La propriété et la diffusion des échantillons physiques sont un terrain peu exploré. En effet, les pratiques disciplinaires veulent que les minuscules échantillons résiduels restent en possession de l'archéométallurgiste. Bien que celui-ci considère que ses échantillons appartiennent toujours aux musées ou aux autres propriétaires des objets qu'il a étudiés, les documenter de manière à ce qu'ils puissent servir à d'autres représente un vaste chantier, dans lequel il compte se lancer à sa retraite.

Études bouddhiques

Le bouddhisme, religion née il y a 2 500 ans, est pratiqué dans une grande partie de l'Asie et compte des disciples, un clergé et des érudites et érudits dans le monde entier. La doctrine bouddhiste est solidement documentée, des premiers manuscrits

aux corpus numériques contemporains. On pense qu'en l'an 705, l'impératrice Wu de la dynastie Tang a fait imprimer et distribuer cent mille copies d'un texte bouddhique, accomplissant ainsi un grand pas dans la diffusion du papier et de l'imprimerie (Barrett, 2008). Les progrès de ces deux domaines sont étroitement liés en Asie de l'Est. L'impression sur bloc de bois y a été perfectionnée des siècles avant que Gutenberg ne mette au point son système d'encre et de caractères mobiles vers 1450.

Les premiers textes bouddhiques remontent à plus de deux mille ans. En tant que textes sacrés, ils étaient transmis oralement en prose et en vers. Certains textes constituent des objets sacrés ; d'autres jouent un rôle liturgique ou savant. Le papier a été précédé de manuscrits sur écorces de bouleau, bambou, briques, tablettes d'argile et autres. Les textes ont ensuite été consignés dans différents systèmes d'écriture et langues, puis traduits de l'un à l'autre. Le bouddhisme s'est ainsi rapidement répandu par le truchement de la traduction, alors que les premiers pratiquants du christianisme, de l'islam et d'autres religions interdisaient souvent la diffusion de textes sacrés dans une langue différente de l'originale. La décision radicale de Martin Luther de traduire la Bible en allemand, par exemple, a fait connaître le protestantisme dans l'ensemble du monde germanophone du ^{xvi}^e siècle.

Cette éthique de large diffusion des textes bouddhiques se poursuit aujourd'hui avec des projets de numérisation massive selon les normes scientifiques et technologiques les plus exigeantes. Les pratiques savantes se transforment à mesure que les textes deviennent accessibles et que les outils pour les étudier gagnent en sophistication. Parce que ces documents peuvent être étudiés dans beaucoup de domaines pour répondre à de nombreuses questions de recherche, les chercheurs et chercheuses emploient une multitude de moyens pour sélectionner les entités qui leur permettront de mettre en évidence des phénomènes. Cette étude de cas sur la philologie bouddhique chinoise traite de l'analyse textuelle, de la création de grands corpus numériques de textes sacrés et savants, des comparaisons de l'usage des artefacts physiques et des objets numériques et du rôle de ces corpus dans les nouvelles méthodes de recherche. La philologie est un domaine qui combine la linguistique, la littérature et l'histoire pour étudier des sources historiques textuelles.

Les premiers textes bouddhiques peuvent être étudiés pour leur valeur historique, culturelle, artistique, linguistique et religieuse. La Chine, le Japon, l'Inde, le Sri Lanka, la Thaïlande, la Mongolie, le Tibet et la Birmanie/le Myanmar comptent parmi les pays où, historiquement, le bouddhisme a prospéré. Il est encore la religion majoritaire au Japon et dans de nombreuses autres régions d'Asie. En Occident, le bouddhisme tend à être un objet d'étude plus qu'une véritable discipline

universitaire : les spécialistes en sont répartis dans des départements de religion, d'histoire, de langues, de civilisation et d'études régionales. Bien que la recherche sur le bouddhisme soit centrée sur les canons de textes sacrés, les questions, les méthodes et les médias de publication de chaque scientifique peuvent correspondre aux disciplines auxquelles elle ou il a été formé ou qu'elle ou il enseigne.

De nombreux textes bouddhiques importants semblent avoir d'abord été rédigés en sanscrit et en d'autres langues indo-aryennes. Le sanscrit est une langue culturelle indo-européenne utilisée depuis longtemps en Inde, qui s'écrit dans plusieurs systèmes d'écriture modernes et dans de nombreux systèmes anciens. Les textes sanscrits et indo-aryens ont ensuite été traduits en chinois et en d'autres langues pour être diffusés et étudiés. Le système d'écriture chinois, qui est utilisé dans plusieurs langues, a été normalisé au III^e siècle avant notre ère. Le chinois représentait la langue culturelle commune de l'Asie de l'Est, remplissant un rôle comparable à celui du latin dans l'Europe antique. Parce que ces textes ont été transmis sur des siècles, une expertise linguistique, historique et culturelle considérable est nécessaire pour les étudier. Par exemple, pour lire des textes bouddhiques traduits en chinois entre le II^e et le V^e siècle de notre ère, un chercheur ou une chercheuse doit connaître le chinois, le sanscrit et les autres langues de la période. Idéalement, il doit aussi maîtriser les écritures dans lesquelles ces langues étaient alors rédigées.

La taille compte

Les spécialistes du bouddhisme constituent une petite communauté éparpillée. Ses grands pôles se trouvent dans les régions du monde où cette religion est la plus pratiquée, principalement en Asie. Leurs ressources sont très variées et recouvrent de nombreux systèmes d'écriture, langues, régions et époques. Les chercheurs et chercheuses travaillent généralement seuls ou en petits groupes. Leurs méthodes reposent notamment sur une lecture méticuleuse des textes, ce qui engendre un petit nombre d'observations ; cependant, avec la croissance des corpus numériques bouddhiques, l'analyse automatisée devient possible. Les philologues bouddhiques commencent ainsi à expérimenter les technologies d'analyse de textes. Les outils de CLARIN pour la linguistique de corpus, que nous avons mentionné plus haut, sont parfois utilisables. De même, BLAST, un outil d'appariement ADN déjà adapté pour les textes historiques en français, pourrait contribuer à certaines questions de recherche de cette communauté.

Quand est-ce une donnée ?

À l'origine, les textes bouddhiques sont religieux. Le choix des éléments probants y dépend de s'ils sont étudiés pour cette raison ou pour d'autres motifs savants. La diffusion des textes sacrés est un principe fondamental du bouddhisme : ses adeptes sont donc motivés pour les traduire, les éditer et les rendre accessibles au plus grand nombre. Ce processus a eu des effets exponentiels au cours des siècles. Chaque texte copié ou traduit peut être exploité pour créer encore plus d'exemplaires dans d'autres langues. Les travaux sur les textes et leur commentaire présentent aussi un intérêt pour la recherche, car ils ont aussi été copiés et traduits. Cet effet multiplicateur s'accélère avec le recours croissant à l'imagerie et à la numérisation, puisque chaque texte enregistré sous forme numérique peut servir à créer encore d'autres produits. La numérisation aide aussi les scientifiques à traiter les différents éléments comme des entités distinctes. En effet, une fois numérisés, les textes peuvent être étudiés dans leur ensemble, mais on peut également en examiner les phrases, les caractères, les frappes, la ponctuation ou d'autres unités. Ces entités peuvent être fouillées et combinées de nombreuses manières pour étudier différents phénomènes.

Sources et ressources

Distinguer sources et ressources dans les études bouddhiques est particulièrement problématique en raison de la très longue histoire de ces œuvres. Les corpus numérisés de textes anciens constituent clairement des ressources. Pour être une source, un texte doit être considéré comme original – autre concept sujet à controverse dans les sciences humaines. Quelques distinctions nous permettront d'illustrer les difficultés à déterminer quand et comment quelque chose peut être traité en donnée.

Sources primaires ou secondaires

La distinction entre sources primaires et secondaires est utile, mais floue. En philologie bouddhique, un texte original en sanscrit ou une traduction chinoise ancienne sont considérés comme des sources primaires, alors que les traductions et les travaux de recherche contemporains sur des textes anciens sont généralement tenus pour des sources secondaires. Dans les livres et les articles de revue, les bibliographies dressent parfois des listes séparées pour les sources primaires et secondaires. Les éditions critiques sont généralement considérées comme des sources primaires parce qu'elles présentent les textes sous une forme qui fait consensus. Elles rassemblent parfois plusieurs sources, produisant ainsi un format lisible au plus près des originaux.

Ces distinctions reposent souvent sur une expertise pointue. Les textes bouddhiques anciens sont en effet bien plus difficiles à interpréter que les documents occidentaux.

En philologie grecque et romaine, par exemple, on peut admettre certains postulats sur les pratiques à l'époque de l'empire unifié. C'est moins vrai dans le cas du bouddhisme : les pratiques culturelles associées à ces documents varient selon le pouvoir en place, la langue, la région, l'époque, le système d'écriture et bien d'autres facteurs. Par ailleurs, le chercheur ou la chercheuse doit aussi prendre en compte le rôle de la personne qui traduit ou transmet le texte. À certaines périodes de l'histoire de l'Inde et, parfois, de la Chine, le traducteur ou la traductrice pouvait légitimement modifier le texte ou s'abstenir de le faire. Même un texte sacré revêtait des sens différents selon les traductions, bien loin du littéralisme qui a cours dans la traduction des textes chrétiens. Les éditions critiques ont également une longue histoire et peuvent comprendre des corrections apportées aux textes. Certaines de ces éditions du ^{xiii}^e siècle sont aujourd'hui considérées comme des sources primaires.

Images statiques ou contenu enrichi

Les études bouddhiques en général et la philologie en particulier reposent sur la capacité à comparer des textes, intégralement ou en partie. Les spécialistes étudient comment ils étaient compris dans différents contextes temporels, géographiques, linguistiques et culturels. Comparer les objets matériels originaux côte à côte est pratiquement impossible pour celles et ceux qui étudient des textes du ^v^e siècle ou antérieurs. En effet, rares sont ceux qui ont survécu sous forme physique : ils ne sont connus que grâce à des traductions ou des mentions ultérieures. Ceux qui restent sont précieux, fragiles et dispersés à travers le monde. La diffusion d'images statiques sur microfilms, de fac-similés sur papier, puis sur CD-ROM a permis pour la première fois aux spécialistes de confronter de nombreux textes. L'étape suivante a été la diffusion d'images statiques en ligne, les rendant plus largement accessibles encore.

Au cours de la dernière décennie, la conversion de ces textes en représentations interrogeables et contenus enrichis s'est accélérée, conduisant à des avancées radicales dans l'exploitation de ces ressources historiques. Au lieu de visualiser une page, un dossier ou une boîte d'archives à la fois, les chercheurs et chercheuses peuvent maintenant comparer, combiner et extraire des informations de ces textes numériques. Leur relation aux textes s'en est trouvée changée de manière subtile et complexe, comme l'ont montré des études sur l'adoption du *Thesaurus Linguae Graecae* par les spécialistes de l'Antiquité (Ruhleder, 1994).

Les infrastructures de la connaissance

La communauté des spécialistes du bouddhisme est unie par son intérêt pour ces textes. Ces chercheurs et chercheuses viennent de différentes disciplines, appliquent des méthodes et théories variées, lisent de nombreuses écritures

et langues et travaillent dans diverses parties du monde. L'infrastructure de la connaissance commune dont ils disposent, pour autant qu'elle existe, est centrée sur les moyens d'accès aux textes bouddhiques. Il est rare qu'ils obtiennent des subventions pour autre chose que des frais de déplacement. Ils ne disposent pas d'un accès immédiat aux organismes de financement public qui pourraient permettre la numérisation des corpus, la création de référentiels, l'acquisition de technologies et d'outils, l'instauration de normes ou l'embauche d'assistantes et assistants. Dans la mesure du possible, ils s'appuient sur des sources trouvées en bibliothèques ou accessibles gratuitement en ligne. L'infrastructure de la connaissance de la communauté savante bouddhique dispose en revanche d'un soutien institutionnel dans la communauté religieuse qui fait vivre ces textes. Les monastères et autres centres religieux œuvrent activement à leur numérisation et à leur diffusion, souvent de concert avec la communauté scientifique.

Les spécialistes en études bouddhiques partagent leur expertise grâce aux colloques, aux médias sociaux et aux technologies internet. H-Buddhism, par exemple, fait partie de H-Net, qui répond aux besoins de spécialistes des sciences humaines et sociales du monde entier. Ce site web actif comprend des listes de discussions, des recensions, des actualités, des offres d'emplois et d'autres annonces (H-NET 2013a, 2013b). Les chercheurs et chercheuses de domaines tels que le bouddhisme chinois peuvent organiser des réunions en ligne où ils lisent des textes bouddhiques, les interprètent, les traduisent et en discutent grâce à des technologies vidéo. Leurs publications incluent de longs remerciements aux collaborateurs et collaboratrices qui leur ont fourni des ressources, ont commenté leurs manuscrits et les ont aidés d'une manière ou d'une autre.

Métadonnées

La question des métadonnées est aussi importante dans les études bouddhiques que dans les autres domaines des sciences humaines. La communauté de recherche n'est pas d'accord sur les versions de référence des textes, ni sur qui a écrit quel texte, ni même sur la notion de paternité. Il n'existe pour les études bouddhiques aucun mécanisme commun de classification comme l'Art and Architecture Thesaurus, ni de modèles de référence tels que CIDOC pour la version numérique des textes.

Les mécanismes d'organisation, pour autant qu'ils existent, semblent être associés à des corpus spécifiques. La version numérique de référence du canon bouddhique chinois est établie d'après l'édition Taisho, élaborée dans les années 1920 et 1930. Cette version papier de référence du canon a été créée à des fins religieuses et savantes. C'est un texte gigantesque, constitué de cinquante-cinq volumes principaux et de quatre-vingts volumes annexes. Bien que coûteuse, elle est devenue une

acquisition nécessaire pour les collections d'études asiatiques et bouddhiques des bibliothèques de recherche. La plupart des citations académiques ultérieures du canon bouddhique chinois renvoient explicitement à cette source, souvent à la page près (Takakusu et Watanabe, 1932). En 1998, la Chinese Buddhist Electronic Text Association (CBETA) a publié la première version numérique de l'édition Taisho, qui suit de près la version papier. Elle a conservé la même pagination afin de garantir la continuité historique de la recherche. Les éditeurs et éditrices de la CBETA ont amélioré le contenu en corrigeant la ponctuation, ce qui constitue une importante contribution scientifique. En effet, les marques de ponctuation permettent de déterminer, par exemple, si un mot chinois est un nom ou un verbe, ce qui influe profondément sur l'interprétation du texte. Le système comporte un balisage syntaxique et sémantique, que des outils peuvent exploiter pour sélectionner et comparer des portions du texte, mais n'instaure pas un vocabulaire de métadonnées ou un quelconque système de classification (CBETA Chinese Electronic Tripitaka Collection, 2013).

En un sens, les méthodes de ces philologues produisent des métadonnées sur les textes qu'elles et ils étudient. En effet, les descriptions philologiques sont des interprétations, lesquelles constituent des métadonnées dans la mesure où elles aident à trouver et interpréter les textes. La limite entre données et métadonnées est floue, car une bonne part de l'activité de recherche consiste à produire des descriptions denses de phénomènes minces, comme le fait l'ethnographie (Geertz, 1973).

Provenance

Plus les sources sont anciennes et plus les questions de provenance peuvent devenir complexes. On ne sait pas précisément quels sont les textes bouddhiques initiaux dont émaneraient les autres parce que les scientifiques n'ont pas réussi à dégager un consensus sur les dates de naissance et de mort du Bouddha. On estime aujourd'hui qu'il aurait vécu au ^{iv}^e ou ^v^e siècle avant notre ère. Les textes se transmettaient par oral et les premières versions écrites sont apparues vers le ⁱ^{er} siècle avant Jésus-Christ. Un manuscrit bouddhique indo-aryen du ⁱ^{er} siècle de notre ère a récemment été découvert : l'époque a pu en être déterminée grâce à des analyses paléographiques et à la datation au carbone (Falk, 2011). Les manuscrits indo-aryens sont antérieurs à tous les écrits chinois connus : ces derniers remontent au ⁱⁱⁱ^e siècle après Jésus-Christ.

L'ambiguïté de la provenance d'anciens manuscrits fait partie de la recherche en études bouddhiques. Les chercheurs et chercheuses comparent des textes en différentes langues, de la même époque ou non, pour tenter d'en identifier les auteurs et autrices, éditeurs et éditrices, traducteurs et traductrices et autres parties impliquées dans la création et la communication de ces textes. Certains chercheurs

s'intéressent à la transmission des idées, d'autres à la langue et d'autres encore aux objets matériels. Les textes bouddhiques peuvent en effet apparaître non seulement sur des supports familiers tels que le bambou et le papier, mais aussi sur des objets du quotidien, comme les briques de Gopalpur qui se trouvent à l'Ashmolean Museum. La provenance même de ces briques est contestée (Johnston, 1938).

Corpus

Les corpus tels que l'édition CBETA et les collections des bibliothèques, archives et musées sont des composantes de l'infrastructure de la connaissance des études bouddhiques. Les ressources sont éparpillées à travers le monde, mais leurs représentations peuvent être rassemblées grâce aux technologies numériques. Lors des dix ans de la CBETA, les spécialistes du bouddhisme ont applaudi l'influence transformatrice qu'elle a eue sur leur travail (Goenka *et al.*, 2008). La CBETA n'est pas un système statique : elle continue d'ajouter du contenu et d'intégrer de nouvelles technologies à son projet. La version actuelle est enrichie d'encodage textuel, dont le XML et l'Unicode. Les dispositifs de recherche et de fouille de textes permettent aux utilisateurs et utilisatrices d'explorer les corpus, d'exporter les résultats dans d'autres outils analytiques et de transférer des extraits dans des logiciels de traitement de texte en conservant intacts les caractères, la ponctuation et les métadonnées. Le système n'est pas seulement utilisable sur ordinateur et dispose d'applications pour la majorité des appareils mobiles, des systèmes d'exploitation, des liseuses de livres électroniques et des médias sociaux.

L'édition CBETA du canon bouddhique chinois continue de croître au-delà de la version papier grâce à la numérisation d'autres textes importants liés au bouddhisme. À l'heure où nous écrivons, la CBETA contient plus de 2 370 sutras et soixante-dix millions de mots chinois. Ce corpus est une ressource d'une ampleur inégalée pour les chercheurs et chercheuses qui effectuent des analyses linguistiques de textes bouddhiques chinois. L'édition CBETA est à son tour exploitée pour créer de nouveaux recueils en ligne. Par exemple, les Vietnamiens ont combiné une traduction automatique du corpus de la CBETA et des dictionnaires de vietnamien électroniques pour servir de base à leur propre édition en langue vernaculaire. Les savantes et savants vietnamiens corrigent et éditent les traductions automatiques (Nguyễn et Nguyễn, 2006).

L'International Dunhuang Project (IDP), lancé en 1994 par le British Museum et qui regroupe aujourd'hui des partenaires du monde entier, est un autre exemple de contribution des collections à l'infrastructure des études bouddhiques. Dunhuang, en Chine, était un important comptoir commercial sur la route de la soie. Au début du ^{xx}e siècle, une large cache de documents bouddhiques scellée depuis près de mille

ans a été découverte non loin de la ville dans les grottes de Mogao. Les cinquante mille documents environ ont été dispersés par divers moyens dans des bibliothèques, des archives et des collections privées du monde entier, pour un résultat controversé. Parmi les documents figurait le fameux *Sutra du Diamant*, aujourd'hui en possession de la British Library. Cet important texte sacré est le premier livre imprimé à être clairement daté : 11 mai 868 (The British Library Board, 2013). L'IDP est un projet de catalogage et de numérisation de manuscrits, des peintures, des textiles et d'autres artefacts de Dunhuang et des sites archéologiques de la route de la soie orientale. À l'heure où nous écrivons, plus de quatre cent mille images sont disponibles sur le site (International Dunhuang Project, 2013). Elles sont librement accessibles à des fins éducatives et de recherche.

Les influences extérieures

Les facteurs externes qui influent sur l'accès aux sources dans les études bouddhiques sont très spécifiques. Si son terrain n'est peut-être pas aussi miné que celui de l'histoire de l'art et de l'archéologie antiques, cette communauté est confrontée à ses propres contraintes économiques, éthiques et de propriété en ce qui concerne ce qu'elle peut traiter en donnée.

Économie et valeur

En raison des principes de leur religion, les textes bouddhiques sont plus largement diffusés que la plupart des autres ressources historiques. Ils ne sont pas pour autant ouverts ou libres. L'énorme édition papier du canon bouddhique chinois est coûteuse : ces volumes peuvent être considérés comme des biens privés, disponibles à la vente. Néanmoins, une fois acquis par les bibliothèques, ils peuvent être vus comme des réservoirs communs de ressources dont l'usage est régi par les bibliothécaires et leurs communautés. Chaque bibliothèque, archive, musée et site historique possède ses propres règles sur qui peut étudier ses ressources et à quelles conditions.

La CBETA, l'équivalent numérique de l'édition Taisho du canon bouddhique chinois, est un système en ligne ouvert et gratuit. Il en existe également des versions sur CD-ROM. La CBETA est un réservoir commun de ressources possédé et régi par l'association qui l'a élaborée et la maintient. La communauté bouddhiste investit dans ces ressources à la fois pour ses croyants et pour les scientifiques. La pérennité et les passagers clandestins sont moins problématiques que dans les domaines qui dépendent de subventions.

Beaucoup de ressources importantes demeurent entre les mains de collectionneurs et collectionneuses privés, dont certains mettent leurs recueils à la disposition des chercheurs et chercheuses. Ces derniers peuvent d'ailleurs rassembler eux-mêmes

des collections considérables au cours de leur carrière. Leurs publications et leurs bibliothèques privées deviennent des ressources précieuses qui seront léguées ou vendues à leur retraite ou à leur mort.

Droits de propriété

Les problèmes de propriété sont généralement liés à la possession des artefacts matériels et à la diffusion des représentations des textes. Les questions peuvent être éthiques autant que juridiques. Des textes bouddhiques ont fait surface au ^{XXI}^e siècle lorsque des régions isolées d'Asie – en particulier l'Afghanistan, le Pakistan et le Tibet – se sont ouvertes. Certains constituent des découvertes exceptionnelles, en particulier les textes que l'on pensait disparus, car connus seulement par des mentions intertextuelles ou des traductions dans d'autres langues. Des documents précieux tombent parfois entre les mains de collectionneurs et collectionneuses privés, qui peuvent en refuser l'accès aux scientifiques. Des feuillets tirés de certains ouvrages sont vendus au marché noir : les textes concernés ont donc peu de chance d'être reconstitués. Des portions substantielles de textes récemment découverts ont disparu. D'autres sont détenus en secret pendant de longues années, connus uniquement d'un petit groupe de chercheurs et chercheuses et dévoilés au grand jour seulement quand ceux-ci sont prêts à publier leurs travaux sur ces textes (Allon, 2007, 2009).

Éthique

Dans les études bouddhiques, les considérations éthiques sont subtiles et complexes. La religion, la recherche, la culture et la politique sont inextricablement mêlées. Par exemple, la diffusion des textes a été soutenue ou au contraire restreinte par les autorités selon les époques. La CBETA est un projet taïwanais, accessible au lectorat sinophone du monde entier. Les relations entre Taïwan et la Chine sont délicates et le bouddhisme représente une source de tensions au sein de la république populaire et vis-à-vis des États étrangers. Le dalaï-lama, lauréat du prix Nobel de la paix en 1989, vit exilé de son Tibet natal, une province indépendantiste de la Chine. Les scientifiques sont très conscients des tensions passées, présentes et potentiellement futures entre les institutions religieuses et politiques, ainsi que de leur influence sur l'accès aux ressources.

La communauté bouddhiste affirme la valeur de ses textes en tant qu'artefacts culturels, historiques et linguistiques et encourage la recherche. À leur tour, les scientifiques respectent les traditions religieuses qui fondent les textes sacrés. Certains objets matériels appellent la vénération, mais la plupart conviennent à l'étude scientifique. Les textes constituent un substitut à la présence du Bouddha. Les artefacts qui l'évoquent sont enchâssés dans un endroit dédié du temple. Les inscriptions sur des

briques peuvent être intégrées au mur d'un bâtiment pour le bénir. Les technologies numériques soulèvent des questions intéressantes, comme ce que pensent les moines bouddhistes de textes enchâssés dans un ordinateur ou un autre appareil.

Mener des recherches en études bouddhiques

La présente étude de cas décrit les travaux récents d'un chercheur en philologie bouddhique chinoise. Pour l'un de ses projets, il a comparé des textes en différentes langues pour étayer une possible origine commune. Il en a tiré un article de revue, puis a repris l'étude de ces textes (Zacchetti, 2002). Un autre de ses projets consiste dans une édition critique synoptique d'importants textes sacrés chinois (Zacchetti, 2005). Depuis, ses travaux reposent largement sur la CBETA, ce qui a apporté de nouvelles questions, méthodes et sources de preuve à ses recherches.

Le degré et l'ampleur de l'expertise nécessaire à l'étude de la philologie bouddhique chinoise donnent le vertige. Notre chercheur a effectué un parcours universitaire en études chinoises et en philologie, qui n'est pas sans rappeler la double formation en archéologie et en métallurgie du scientifique de l'étude de cas précédente. On notera en particulier le nombre et la variété des langues avec lesquelles il travaille, tout en enseignant et en publiant la plupart de ses travaux en anglais. La philologie classique suppose de connaître le latin et le grec, qui sont des langues proches de l'anglais ; notre chercheur les a apprises dans le cadre de sa formation. Sur cette base, il a assimilé le chinois, le sanscrit, le tibétain, le pali et le japonais. L'étude de ces langues et de leurs systèmes d'écriture n'est pas courante chez les scientifiques en Occident. La langue maternelle de notre chercheur est l'italien, dans laquelle il enseigne et publie également. Lorsqu'il donne un cours avancé sur l'histoire et la culture chinoises, il s'exprime en mandarin. Ses vastes compétences linguistiques sont un moyen en vue d'une fin : l'étude de l'histoire, de la culture et de la langue du bouddhisme telles que les anciennes civilisations asiatiques les ont consignées et appréhendées.

Les corpus numériques et les dispositifs associés ont révolutionné la recherche linguistique en études bouddhiques chinoises. Notre chercheur appelle la CBETA son télescope, car il peut l'utiliser pour voir plus loin que jamais auparavant. Pour expliquer cette métaphore, il évoque son compatriote Galilée, qui a décrit comment une technologie nouvelle – un télescope amélioré par ses soins – lui permit de voir des phénomènes jusqu'alors invisibles. Le philologue compare aussi la CBETA et les recueils similaires à un microscope, car il peut examiner chaque texte de bien plus près. Il est en mesure de produire ses propres concordances sur des portions du texte, de les fouiller et de réaliser des comparaisons d'un nouveau genre.

Les questions de recherche

Notre chercheur examine les liens intertextuels qui révèlent comment ces textes ont été communiqués et compris par différentes cultures à des époques et dans des lieux particuliers. Les questions qu'il traite dans l'article de revue ont émergé alors qu'il étudiait une version papier d'un ancien texte bouddhique chinois, où il a repéré une ressemblance avec un texte qu'il avait lu en pali, une langue ancienne de l'Inde. Établir un lien géographique ou culturel entre ces textes représenterait une importante découverte historique : en effet, on supposait jusque-là qu'ils étaient indépendants. Le texte chinois est le plus court des deux ; le pali est un chapitre d'un écrit plus long. L'hypothèse du chercheur est que les deux textes sont la traduction d'un texte indo-aryen antérieur, ce qui illustre le cheminement complexe des idées entre les cultures (Zacchetti, 2002).

Les éditions critiques tentent d'établir l'esprit original du texte le plus complètement possible. Pour cela, elles en consignent la provenance, comparent leurs traductions et leurs variations et expliquent ce qu'on sait de leur contexte d'écriture, de circulation et d'utilisation. Les éditions critiques deviennent des sources primaires si elles sont acceptées comme des travaux de référence dans leur domaine. L'édition critique réalisée par notre chercheur est une analyse complète de textes similaires dans deux langues différentes. Il a consacré des années de travail à comparer trois chapitres d'un important texte bouddhique chinois avec leurs pendants en sanscrit. La section concernée du canon chinois est un texte unique, tandis que les équivalents sanscrits sont des tablettes issues de quatre textes différents, dont une partie jusque-là non publiée d'un manuscrit. En comparant les textes sanscrits avec la traduction chinoise, il en a déduit comment ils avaient été traduits des siècles auparavant et comment les concepts qu'ils contiennent étaient alors appréhendés par ces deux cultures (Zacchetti, 2005).

La collecte de données

Le terme « donnée » ne faisait pas partie du vocabulaire usuel de notre chercheur avant que nous nous parlions pour les besoins de la présente étude de cas. Ses méthodes impliquent une sélection méticuleuse de termes, d'expressions, de concepts et de caractères dans les textes, que l'on assimilera aisément à une collecte de données. Parce qu'il effectue des comparaisons binaires, il dresse des tableaux à deux colonnes dans un traitement de texte. Lorsqu'il travaille sur des artefacts matériels, il effectue des recherches dans le texte artisanalement, puis saisit la description de ses observations dans le tableau. Lorsque les entités sélectionnées proviennent de la CBETA, il les saisit sous forme numérique en les copiant et collant dans son document. Le dispositif de la CBETA intègre les métadonnées du texte sélectionné dans les copier-coller, ce qui permet une sélection des données plus rapide et plus fiable. Certaines comparaisons

sont effectuées ligne à ligne et d'autres, proposition à proposition. Chaque entrée est copieusement annotée d'informations sur sa source et sur les similarités et différences des éléments appariés. Notre chercheur gère ses tableaux à l'aide de conventions simples de nommage des fichiers, comme le numéro de chapitre du texte chinois.

Dans l'article de revue et l'édition critique monographique, son but était d'interpréter les textes et plus particulièrement l'usage de la langue dans chacun d'eux. Ses recherches dépendent d'une panoplie d'outils et de ressources numériques, dont certains sont courants et d'autres conçus sur mesure par d'autres scientifiques. Un outil obscur qui lui est particulièrement utile est une source lexicographique en ligne sur les variantes historiques des caractères chinois (ministère de l'Éducation, République de Chine, 2000).

Le texte chinois sur lequel notre chercheur a travaillé est un exemplaire papier, tandis que son équivalent en pali se trouvait dans un corpus numérique. Il étaye son argumentation à l'aide de l'histoire de ces documents et d'autres travaux scientifiques, tout en reconnaissant qu'il est difficile d'établir leurs origines de manière certaine. Il note par exemple les « originaux putatifs » de termes importants (Zacchetti, 2002, p. 79). Afin de rassembler des preuves, il recherche des sources numériques et papier, dont certaines se trouvent en bibliothèque, d'autres en ligne ou sur CD-ROM et d'autres encore dans la considérable collection de reproductions de textes qu'il a accumulés au cours de sa carrière. Il repère les traductions antérieures, ainsi que les bibliographies et catalogues anciens où ces textes sont mentionnés pour la première fois (Zacchetti, 2002, p. 75).

L'analyse des données

Dans la démarche de notre chercheur en philologie bouddhique chinoise, l'analyse consiste à interpréter les éléments figurant dans ses notes et ses tableaux pour construire un argumentaire. L'article de revue présente ses analyses sous la forme d'une série de tableaux, avec le texte chinois à gauche et le pali à droite. Bien que l'article soit écrit en anglais, le commentaire comprend beaucoup de termes en chinois et en pali pour expliquer les comparaisons. Certains commentaires expliquent des expressions obscures ; d'autres comparent des interprétations littérales et métaphoriques des textes. Comme dans les revues juridiques, les notes de bas de page occupent parfois plus d'espace que le corps du texte ; comme dans les revues de sciences sociales, l'article commence par une revue de la littérature pour situer le propos et se termine par une discussion des résultats et une conclusion.

Nous considérons la monographie comme une édition critique synoptique en raison des comparaisons ligne à ligne et proposition à proposition qu'elle contient. L'essentiel de l'ouvrage consiste en une comparaison des textes dans un tableau à deux colonnes, dont nous montrons un extrait à la figure 7.2 (Zacchetti, 2005, p. 157). Comme dans l'article de revue, les comparaisons sont abondamment annotées de commentaires en anglais et de références en chinois et en sanscrit. Certaines de ces notes confrontent des définitions de termes ou expliquent l'énumération des concepts. Les tableaux comparent des sources, des variantes et des traductions, mettant en évidence l'évolution de la langue et l'origine possible des textes. Les sources de notre chercheur comportent souvent des variantes anciennes de caractères chinois et sanscrit. Celles-ci sont difficiles à rendre précisément sous forme numérique, car elles n'existent pas dans Unicode. Une partie de l'analyse du chercheur consiste à translittérer, s'il y a lieu, le texte en caractères modernes afin de rendre le contenu plus accessible au lectorat contemporain. L'édition critique comprend en outre une longue réflexion sur la provenance de ces textes (Zacchetti, 2005).

La publication des résultats

Les spécialistes du bouddhisme chinois publient leurs travaux dans des revues, des livres et des actes de colloques. Les publications philologiques reposent lourdement sur les citations textuelles. Celles-ci sont bien plus longues que les quelques mots autorisés par le droit de citation et représentent donc une pratique typiquement scientifique. Dans les deux publications que nous avons évoquées, le texte est présenté avec des jeux de caractères modernes, alors qu'une partie des caractères chinois sur lesquels notre chercheur a travaillé correspondent à d'anciennes écritures qui ne peuvent pas être représentées dans Unicode. Unicode étant aujourd'hui intégré à tous les logiciels et matériels grand public, ses écrits en anglais, en chinois et en sanscrits apparaissent correctement sur les écrans d'ordinateur et les imprimantes, du moins pour les écritures modernes. Certes, des problèmes peuvent survenir avec certains claviers et navigateurs, mais l'interopérabilité s'améliore. Bien que notre philologue travaille sur Microsoft Word, certains chercheurs et chercheuses en sciences humaines qui travaillent avec plusieurs systèmes d'écriture lui préfèrent le langage de composition LaTeX. La plupart des maisons d'édition acceptent l'un ou l'autre de ces formats ou des PDF pour les manuscrits.

SYNOPTIC EDITION OF THE *GUANG ZAN JING* – CHAPTER 1: 光讚品

		karmāntājivā virahitakuśalakāyavānmanas- karmāntājivās ¹⁰⁹ ca bhavanti sma
§ 1.66 (147c 2b-1 dha 1)	一切衆生得平等心，展轉相瞻如父、如母、如兄、如弟、如姊、如妹，各各同心，等無偏邪，皆行慈心。	PG 4r 4-5 (Ś 18, 22-19, 1; PD 10, 1-2; PSL <i>kā a</i> 4-5): sarvasatvās ca sarvasatveṣu samacittā abhūvan* yad uta ¹¹⁰ mātāpitṛbhrātṛbhaginīsamacittāḥ mitrajñātisahāyasamacittāḥ ¹¹¹
§ 1.67 (148a 1-2)	一切群萌悉修十善，清淨梵行，無有塵埃。	PD 10, 2-3 (PG 4r 5-6; Ś 19, 2-3; PSL <i>kā a</i> 5): daśakuśalakarmapathasevinās ca bhavanti sma ¹¹² / brahmacāriṇaḥ śucayo nirāmayagandhāḥ ¹¹³
§ 1.68 (148a 3-4)	一切黎庶悉獲安隱，所得安隱猶如比丘得第三禪。于時衆生而致智慧，而悉具足善快調定，離於卑劣，逮得和雅。	PG 4r 6-8 (PD 10, 3-8; PSL <i>kā a</i> 5-6; Ś 19, 3-8): sarvasatvās tasmin samaye sarvasukhasamarpitā abhūvan* evamrūpeṇa sukhena samanvāgatās ¹¹⁴ tadyathā {s} tṛtiyadhyānasamāpānnasya bhikṣoḥ sukhaṃ sarvasatvās ca tasmin samaye evamrūpayā prajñayā samanvāgatā abhūvan* yad evaṃ jānānti sma ¹¹⁵ • sādhu dānaṃ sādhu damaḥ sādhu saṃnyamaḥ ¹¹⁶ sādhu satyaṃ • sādhu apramādaḥ sādhu maitri sādhu karuṇā sādhu avihiṃsā prāṇibhūteṣu ¹¹⁷ •

¹¹⁰ sarvasatvās ... yad uta: not in PD & PSL.

¹¹¹ PG wrongly repeats verbatim this latter compound. PD 10, 2 and PSL have at this point a longer reading: mitrāmātyajñātisāhāyitasamacittā. Note that Ś has all the words construed as one compound.

¹¹² PG 4r 5-6 & Ś 19, 2: daśakuśalakarmapa(tha)samanvāgatā [Ś without daśa-] abhūvan.

¹¹³ PG 4r 6, Ś 19, 3 and PSL *kā a* 5: nirāmagandhāḥ, which seems to be the correct reading; after this word, PG & Ś + sarvakuśalavitarakavigatāḥ.

¹¹⁴ PD 10, 4 & PSL *kā a* 5: idṛśaṃ sukhaṃ pratilabhante sma.

¹¹⁵ yad ... sma: PD 10, 6 & PSL *kā a* 6: yad anyabuddhakṣetrasthā buddhā bhagavanta evaṃ [PSL + udānaṃ] udānayanti sma.

¹¹⁶ Ś 19, 7: saṃnyamaḥ.

¹¹⁷ sādhu dānaṃ ... prāṇibhūteṣu: PD 10, 7-8 & PSL *kā a* 6: sādhu damaḥ [PSL + sādhu śamaḥ] sādhu saṃnyamaḥ sādhu cīrṇa brahmacāryāvāsaḥ sādhu prāṇibhūteṣv avihiṃseti.

Figure 7.2.

Page 157 d'*In Praise of the Light: A Critical Synoptic Edition with an Annotated Translation of Chapters 1–3 of Dharmarakṣa's Guang Zan Jing, Being the Earliest Chinese Translation of the Larger Prajnaparamita* de Stefano Zacchetti (2005).

Crédit : reproduit avec l'autorisation de Stefano Zacchetti, titulaire de la chaire Numata d'études bouddhiques à l'université d'Oxford.

L'article de revue de notre philologue est long de 23 pages, dont 134 notes de bas de page et trois pages de références (Zacchetti, 2002). Son édition critique synoptique en fait 469, divisées en trois parties comptant chacune plusieurs chapitres, plus différents appendices et une longue bibliographie (Zacchetti, 2005). Ce dernier ouvrage a été publié sous forme de monographie au Japon par un important centre de recherche en bouddhologie. Bien que les droits en soient réservés, le livre a été mis à disposition gratuitement en ligne au format PDF et a intégré d'autres ensembles de ressources bouddhiques en ligne. Une édition papier est également disponible. Les deux publications incluent de longs remerciements aux autres chercheurs et chercheuses qui ont lu le manuscrit, proposé des commentaires ou des ressources ou ont apporté leur aide d'une quelconque manière.

La conservation, le partage et la réutilisation des données

Les méthodes de recherche et les pratiques professionnelles de notre chercheur ont beaucoup changé avec les progrès des technologies de l'information. Les possibilités dont il dispose pour conserver, partager et réutiliser ses données se sont étendues. Un bref historique de ses méthodes et technologies illustre cette transition. Il a soutenu son mémoire de premier cycle en études chinoises en 1994, deux ans seulement après l'adoption officielle d'Unicode. À l'époque, ce standard n'était pas encore largement installé dans les logiciels de traitement de texte, les imprimantes ou les affichages. Il existait en Asie d'autres logiciels pour enregistrer et représenter des jeux de caractères chinois, japonais et coréens, dont certains étaient déjà utilisés dans les bibliothèques de recherche en Occident.

N'étant alors qu'un étudiant ne disposant pas de ces premières technologies pour représenter des caractères chinois, notre chercheur en herbe a rédigé son mémoire – en italien – sur un traitement de texte courant en ménageant des espaces vierges pour les caractères chinois. Il a ajouté les idéogrammes à la main sur l'impression finale qu'il a envoyée pour valider son cursus. Il n'existe donc aucune copie numérique de ce mémoire comportant les caractères chinois.

Pour l'édition critique synoptique, il a eu recours à une ancienne génération de traitement de texte chinois. Le résultat, produit de plusieurs années de labeur, a été converti au format PDF pour être publié. Il a été largement diffusé en ligne, mais cette ancienne version de PDF ne pouvait représenter que des images de page statiques. Par conséquent, la publication n'est pas interrogeable et ne comporte aucune structure sémantique lisible par ordinateur. Le logiciel de traitement des textes des années 1990 qu'il utilisait a correctement rendu les jeux de caractères chinois et sanscrits, mais les fichiers ne sont plus pris en charge par les technologies d'aujourd'hui. En effet, ce logiciel n'a pas été largement adopté et les fichiers n'ont pas été migrés. La migration

peut être rendue impossible par un encodage dans un format autre qu'Unicode, un logiciel qui n'est plus pris en charge, un stockage dans des formats physiques qui ne sont plus lisibles ou une combinaison de ces facteurs. Quelle qu'en soit la raison, la conséquence est qu'il n'existe plus de version lisible par ordinateur qui conserve la structure syntaxique et sémantique de cette œuvre majeure. La monographie est ouverte au sens où elle est librement accessible, mais pas au sens où son contenu est découvrable ou explorable par des ordinateurs. Cette situation est fréquente dans les sciences humaines et particulièrement problématique quand il s'agit d'un ouvrage de référence. Corriger ou mettre à jour le texte supposerait de produire une édition en repartant de zéro.

Pour ses travaux actuels, notre chercheur veille à enregistrer et conserver ses données et métadonnées dans des formes lisibles par ordinateur. Ses données proviennent de sources diverses et variées. Il faut beaucoup de travail pour exploiter ces dernières et seuls quelques spécialistes travaillent au même moment sur un même texte. En organisant ses données au fur et à mesure, il sera d'autant plus capable de les réutiliser ultérieurement. Il peut aussi partager ses données et ses analyses avec d'autres. Bien que les échanges interpersonnels soient la forme de partage la plus courante dans cette communauté, il existe quelques référentiels numériques qui acquièrent des ressources bouddhologiques, comme celles qu'il a comparées dans sa monographie ouverte.

Conclusion

Les sciences humaines sont les disciplines les plus difficiles à caractériser, car leurs limites sont les moins bien définies. De plus, tout regroupement de champs ou de domaines en leur sein est aussitôt contesté. C'est la diversité des objets d'études, des questions et des méthodes de recherche qui cause la difficulté à caractériser les données des sciences humaines. Un même objet peut être étudié de nombreuses façons, pour répondre à de nombreuses questions, à de nombreux endroits et époques. Chaque chercheur ou chercheuse peut présenter ses résultats de façon différente. Ces représentations deviennent alors des entités qui mettent en évidence des phénomènes. Chaque objet – et même chaque représentation d'objet – peut être réinterprété plusieurs fois. La capacité à interpréter ou à réinterpréter – que ce soient des artefacts tirés de fouilles archéologiques, des vases grecs, des textes bouddhiques, la transmission de savoir dans la Chine du III^e siècle ou tout autre aspect de la culture humaine – à la lumière de nouvelles perspectives ou de nouveaux éléments constitue l'essence même de la recherche en sciences humaines. Cependant, cette capacité à réinterpréter dépend d'une forme d'infrastructure différente de ce qui existe dans

les disciplines où les connaissances sont plus cumulatives. Élaborer des normes pour décrire des objets ou leur provenance conduit à demander qui est légitime pour établir des mécanismes de classification ou la documentation de provenance. Ces caractéristiques de leur activité de recherche, auxquelles s'ajoutent les fonds limités et les parties prenantes en concurrence, ont limité la capacité des spécialistes et des organismes de financement à développer une infrastructure commune.

La première provocation – s'intéresser à qui possède les données de la recherche, les contrôle, y a accès et les pérennise permettra de déterminer comment leur valeur peut être exploitée et par qui – est particulièrement problématique dans le cas des sciences humaines. En effet, les ressources nécessaires aux spécialistes de l'histoire de l'art antique, de l'archéologie et des études bouddhiques sont éparpillées et chaque ensemble, collection ou corpus est régi et maintenu – ou non – par sa propre communauté. Les efforts pour construire une infrastructure, comme le banc d'essai de CLAROS, démontrent ce qu'il est possible d'accomplir en agrégeant des ensembles sur une plateforme commune et en les enrichissant grâce à de nouvelles fonctionnalités techniques. Des démarches similaires de fédération des ensembles sous-tendent les efforts infrastructurels tels que CLARIN, Europeana et la DPLA. Pour toutes ces entreprises, la pérennité représente un défi. La CBETA, élaborée par la communauté bouddhiste, semble disposer du soutien infrastructurel le plus durable.

Transmettre des connaissances au fil du temps et dans différents contextes est difficile dans les sciences humaines, comme affirmé dans la deuxième provocation. Cependant, on commence à appréhender ce qui est nécessaire pour y parvenir. Dans les cas présentés ici, la capacité à comparer des objets et à les examiner de plus près est essentielle. Dans beaucoup de projets, confronter des représentations d'objets répond aussi bien, voire mieux, aux besoins de la recherche que la comparaison d'objets matériels. CLAROS permet de comparer des images de vases grecs côte à côte, de les faire pivoter et de les agrandir. Grâce à la CBETA, il est pour la première fois possible d'interroger et de confronter des textes bouddhiques chinois à grande échelle. Les images de caractères issus d'anciennes écritures peuvent être agrandies. Les visualisations en trois dimensions du griffon de Pise offrent des détails invisibles à l'œil nu. Bien que ces technologies poursuivent des objectifs de recherche semblables, elles ne sont pas transférables d'un contexte scientifique à l'autre. En effet, CLAROS n'a pas grand-chose à offrir à l'archéométallurgiste, même s'il comporte des collections qui l'intéressent, car les données ne contiennent pas les informations métallurgiques dont il a besoin. Le dispositif élaboré pour la CBETA constitue une puissante trousse à outils pour les chercheurs et chercheuses capables de travailler en chinois et qui sont experts dans son écriture et sa culture. Le défi pour

cette infrastructure de la connaissance consiste à généraliser ses fonctionnalités de manière à pouvoir les transférer d'un domaine à l'autre.

Pour ce qui est des troisième et quatrième provocations, les formes et les genres des publications dans ces domaines des sciences humaines sont distincts de ce qui se pratique dans les autres types de sciences. Dans nos deux exemples d'histoire de l'art et d'archéologie antiques, les articles sont divisés en courtes sections signées d'un auteur ou d'une autrice. Les publications en philologie bouddhique présentées dans notre étude se situent entre les pratiques des sciences exactes, qui fournissent de longs tableaux de données, et celles du droit, avec des notes de bas de page prenant parfois plus de place que le corps du texte. Bien que la diffusion des données ne soit pas courante dans ces domaines de recherche, une partie des travaux sont ouverts. Ainsi, CLAROS fournit son code en *open source* et constitue des recueils de traces qui sont accessibles aux institutions partenaires. Les données métallurgiques sur le griffon de Pise sont présentées sous forme de tableaux dans les articles qui le concernent. Les publications en philologie bouddhique étayent en détail leur argumentaire en présentant des comparaisons d'entités dans des tableaux. Aucune de ces actions ne correspond à la diffusion des données au sens où on l'entend dans les sciences exactes, mais elles remplissent les fonctions de la communication savante dans ces communautés.

Enfin, il est important de rappeler l'histoire des infrastructures de la connaissance de ces domaines de recherche. Bien que le travail de réinterprétation soit continu, il s'appuie sur des ensembles et une expertise accumulés au cours des siècles. Les connaissances se transmettent au fil du temps et entre contextes, certes, mais pas toujours bien ni facilement. Les spécialistes de l'art antique, de l'archéologie, de l'archéométaballurgie et des études bouddhiques ont adapté leurs pratiques à beaucoup de nouvelles technologies – papyrus, imprimerie et papier, pour n'en citer que quelques-unes. Elles et ils avancent en terrain miné entre problématiques économiques, éthiques, de droits de propriété et de concurrence entre acteurs. Beaucoup utilisent les technologies de l'information de manière innovante, ce qui leur confère un avantage concurrentiel sur les autres chercheurs et chercheuses. Ils sont aussi confrontés aux risques posés par ces nouvelles technologies, en particulier le philologue incapable de récupérer ses données depuis les anciennes générations de logiciels dans lesquels il s'est investi. Une négligence bienveillante ne saurait suffire à préserver les traces culturelles qui ne sont enregistrées que sous forme numérique. À moins que des investissements substantiels soient réalisés dans l'infrastructure, ces traces disparaîtront, contrairement aux papyrus, aux tablettes cunéiformes, aux vases grecs et aux sculptures qui sont toujours étudiés aujourd'hui. Les infrastructures de la connaissance les plus résilientes dans ces

disciplines sont les réseaux personnels des chercheurs. Ces derniers échangent des connaissances grâce à des médias nouveaux comme anciens ; ils partagent des idées et construisent des arguments par de longues discussions régulières ; ils fournissent un point d'entrée pour les ressources informationnelles qu'on ne peut trouver sur Internet, dans les catalogues de bibliothèques ou dans d'autres sources. Alors que le *big data* vient généralement à votre ordinateur, souvent, les chercheurs doivent aller à leurs données. Leurs infrastructures de la connaissance doivent d'une manière ou d'une autre s'adapter à ce mélange complexe de ressources intellectuelles, matérielles et numériques et fournir des moyens pour les exploiter au mieux.

III. Politiques et pratiques en matière de données

8. Le partage, la diffusion et la réutilisation des données

« L'eau, l'eau était partout, et toutes les planches du bord se rétrécissaient.
L'eau, l'eau était partout, et nous n'avions pas une goutte d'eau à boire. »
– Samuel Taylor Coleridge, *La complainte du vieux marin*, partie II. Traduction
d'Auguste Barbier.

Introduction

Il est bien connu que Galilée (1610) au ^{xvii}^e siècle et Cavendish (1798) au ^{xviii}^e ont fourni l'intégralité de leurs données pour étayer leur argumentation. Les travaux de Galilée sur les lunes de Jupiter et ceux de Cavendish sur la densité de la Terre demeurent des jalons de l'histoire scientifique parce que leurs données, leurs méthodes et leurs argumentaires peuvent être examinés encore aujourd'hui (Goodman *et al.*, 2014). Lorsqu'on demande aux chercheurs et chercheuses du ^{xxi}^e siècle s'ils seraient prêts à partager leurs données, la plupart répondent que oui. Mais quand on leur demande s'ils ont entrepris de les diffuser, ils admettent généralement que non (Tenopir *et al.*, 2011 ; Wallis *et al.*, 2013). Bonne volonté ne veut pas dire action, c'est pourquoi « les données, les données sont partout, et nous n'avons pas une goutte à boire » est la norme dans la plupart des domaines de recherche. Galilée et Cavendish divulgueraient-ils leurs données de nos jours ?

Une autre manière d'envisager la question consiste à se demander si, effectivement, « la valeur des données réside dans leur usage » et si cela implique qu'il faudrait « faire de l'accès intégral et ouvert aux données scientifiques la norme internationale pour l'échange de données de la recherche » (National Research Council, 1997, p. 10). Ce genre de prises de position maintes fois entendues promeut l'idée que les données constituent une fin en elles-mêmes. Si elles possèdent une valeur intrinsèque, alors elles devraient être enregistrées, conservées comme des atouts et maintenues indéfiniment pour être réutilisées. Un point de vue différent, lui aussi répété à l'envi, est que les objets de recherche sont malléables, changeants et mobiles (Latour, 1987). Les données ne sont pas des « objets naturels », comme expliqué aux chapitres 1 et 2 ; elles sont plutôt des moyens en vue d'une fin, indissociables du processus de recherche. Dans cette perspective de vue, on peut se débarrasser des données à la fin d'un projet ou après publication d'un article. Elles représentent une charge autant qu'un atout et ne valent pas nécessairement l'investissement nécessaire à

leur conservation. La vérité se situe sans doute entre ces deux extrêmes. Certaines données valent la peine d'être partagées et beaucoup d'autres non.

Malgré plusieurs décennies d'avancées réglementaires en matière d'accès ouvert aux données, les rares statistiques existantes montrent de faibles taux de dépôt et de diffusion. Une enquête menée par la revue *Science* (2011) sur ses évaluateurs et évaluatrices (*peer reviewers*) a révélé que seulement 7,6 % d'entre eux versent leurs données dans des référentiels disciplinaires, alors que 88,7 % les conservent sur les serveurs de leurs universités ou les ordinateurs de leurs laboratoires, hors de portée des autres scientifiques (*Science Staff*, 2011). Seulement 1 % des données environnementales sont accessibles après publication des résultats (Reichman *et al.*, 2011). Même lorsque les revues appliquent des politiques de mise à disposition, celles-ci sont peu respectées. Une étude portant sur des articles de recherche dans cinquante-cinq revues à fort impact, dont la plupart exigeaient une forme de diffusion des données, a découvert que seulement 9 % des publications déposaient l'intégralité de leurs données en ligne. Lorsque des règles étaient applicables (ce qui était le cas pour environ 70 % des articles), 59 % suivaient partiellement les consignes de la revue. En revanche, les publications qui n'étaient pas soumises à ces exigences n'ont jamais divulgué l'intégralité de leurs données (Alsheikh-Ali *et al.*, 2011).

La capacité à diffuser, partager et réutiliser des données – ou leurs représentations – dépend de la présence d'infrastructures de la connaissance adéquates. Celles-ci reposent à leur tour sur des accords entre des acteurs nombreux et en concurrence. Qui doit investir dans l'infrastructure et qui doit en bénéficier ? Comme le montrent bien les études de cas, les données et leurs usages diffèrent selon les contextes et au fil du temps. Les types d'investissement dont ces communautés ont besoin dans leurs ensembles, leurs outils et leurs politiques varient en fonction, de même que leurs rapports avec les autres parties prenantes.

Ce chapitre offre trois analyses contrastées. La première évalue dans quelle mesure l'intérêt des parties prenantes à diffuser et réutiliser des données est une question d'offre et de demande. La deuxième porte sur les motivations des scientifiques à diffuser et réutiliser des données en replaçant les pratiques en la matière dans le contexte de la communication savante. La troisième analyse examine les implications de ces deux perspectives pour la conception et la pérennité des infrastructures de la connaissance. Ces analyses s'appuient sur les études de cas, sur des rapports – demandant qui partage les données, comment, quand, pourquoi et avec quels résultats, qui réutilise les données comment, quand, pourquoi et avec quels résultats – et sur des discussions avec des scientifiques, chercheurs et chercheuses de nombreuses disciplines.

L'offre et la demande de données de recherche

Le rapport *Bits of Power* du National Research Council (1997) n'est qu'un des nombreux documents d'orientation qui encouragent l'accès ouvert aux données de la recherche. Une décennie après sa parution, l'Organisation de coopération et de développement économiques (OCDE) a formulé des principes internationaux pour l'accès aux données de la recherche issues de fonds publics. Les treize principes, développés dans un bref document, sont l'ouverture, la flexibilité, la transparence, la conformité au droit, la protection de la propriété intellectuelle, la responsabilité formelle, le professionnalisme, l'interopérabilité, la qualité, la sécurité, l'efficacité, la responsabilité de rendre compte et la pérennité. Bien que largement applicables, ces principes ont une portée étroite. En effet, la définition des données de la recherche du rapport de l'OCDE n'inclut que les « enregistrements factuels [...] utilisés comme sources principales [...] généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche ». Elle exclut spécifiquement « carnets de laboratoire, analyses préliminaires [et] objets matériels » tels que des échantillons de laboratoires (Organisation de coopération et de développement économiques, 2007, p. 18). La plupart des politiques et rapports ultérieurs promeuvent le partage d'un plus large éventail de données (Australian National Data Service, 2014 ; Ball, 2012 ; Wood *et al.*, 2010). Le document de l'OCDE reste le plus précis concernant l'identification des nombreuses contraintes qui pèsent sur la diffusion et leurs variations selon le type de données, la juridiction et autres.

Le problème de l'offre et de la demande s'apparente à celui de l'œuf ou de la poule. Puisque la demande de réutilisation est faible, peu de données sont divulguées. Puisque peu de données sont divulguées, la demande ne suffit pas à bâtir l'infrastructure nécessaire à une réutilisation aisée. La plupart des politiques de partage de données essaient d'augmenter l'offre en obligeant ou en encourageant les chercheurs et chercheuses à diffuser leurs données. Les responsables politiques, les organismes de financement, les revues scientifiques et les autres parties prenantes manient surtout le bâton – en imposant des plans de gestion des données, le dépôt... – parce qu'ils ont peu de carottes à proposer. Curieusement, ces politiques mentionnent à peine la demande en données attendue ou l'infrastructure nécessaire à la diffusion et à la réutilisation.

Il ne suffit pas de déclarer qu'il faut diffuser les données. Les politiques de partage en appellent à la générosité des scientifiques par des arguments historiques en faveur de l'ouverture. Elles prennent rarement en compte le caractère concurrentiel de la recherche, la quête de la récompense et de l'attribution des mérites, la disproportion du travail fourni par rapport aux bénéfices obtenus, la diversité des pratiques

d'un domaine à l'autre et d'un scientifique à l'autre, la disparité des ressources entre communautés, les difficultés à interpréter les données partagées et la quantité de ressources nécessaires pour partager ou pérenniser des données. Les politiques comme celles de l'OCDE s'adressent aux États, appelant là encore à l'ouverture et à la transparence.

Les États aussi sont en concurrence et les politiques nationales diffèrent par le degré d'ouverture ou de fermeture par défaut des informations. De nombreux mécanismes existent pour augmenter l'offre de données, que nous considérons tous, pour les besoins du présent ouvrage, comme de la diffusion de données. Parmi eux, on citera les données divulguées par les chercheurs et chercheuses, les données de recherche disponibles dans des référentiels, des observatoires ou d'autres types d'ensembles et les données recueillies à des fins autres que la recherche.

L'offre de données de recherche

Les origines des politiques de partage des données sont multiples. La marchandisation croissante de l'information qui a suivi l'évolution des régimes de propriété intellectuelle dans les années 1970 et 1980 a constitué un facteur significatif (Schiller, 2007). Les tentatives de brevetage et de commercialisation du génome humain aux débuts du séquençage ont provoqué un tollé autour du contrôle de l'information scientifique. Ces tentatives d'« anticommons » ont contribué au développement de réservoirs communs de ressources pour la recherche. L'instauration de politiques pour l'accès ouvert aux données génétiques a constitué un tournant dans les pratiques scientifiques de partage de données (Hamilton, 1990 ; Hess et Ostrom, 2007b ; Koshland, 1987 ; *Science Staff*, 1987 ; Roberts, 1987, 1990 ; Watson, 1990).

De nombreux acteurs prônent le partage des données : les organismes de financement – publics comme privés –, les instances politiques, telles que les académies nationales et les organes de financement des universités, les maisons d'édition de revues, le corps enseignant, le grand public et les chercheurs et chercheuses eux-mêmes. Nous reformulons ici quatre raisonnements, développés dans un ouvrage antérieur (Borgman, 2012a), qui sont avancés pour justifier le partage des données : 1) pour reproduire la recherche, 2) pour mettre des ressources publiques à la disposition de la population, 3) pour optimiser les investissements dans la recherche et 4) pour faire progresser la recherche et l'innovation. Cette liste n'est pas exhaustive, mais elle fournit un cadre pour examiner les interactions de l'offre et de la demande dans les politiques, les pratiques et les parties prenantes du partage des données.

Les termes « raisons », « raisonnement », « arguments », « motivations », « incitations » et « bénéfices » sont souvent confondus. Un raisonnement (*rationale*) est une explication des principes d'une opinion, d'une croyance ou d'une pratique. Un argument, en revanche, vise à convaincre : c'est l'ensemble des raisons données par un individu ou une instance pour entreprendre quelque chose. Les raisonnements sont sous-tendus par des motivations intrinsèques et extrinsèques, qui peuvent être explicites ou implicites. Une motivation intrinsèque (*motivation*) est quelque chose qui pousse à agir, alors qu'une motivation extrinsèque ou incitation (*incentive*) est une influence externe qui suscite l'action. Un bénéficiaire est ici un individu, un organisme, une communauté, un secteur économique ou une autre partie prenante qui tire un bénéfice du partage des données, par exemple leur utilisation dans un but particulier.

Pour reproduire la recherche

On dit souvent de la reproductibilité qu'elle est la référence absolue de la science (Jasny *et al.*, 2011). Cette justification du partage des données est percutante, mais problématique. Bien qu'elle soit fondamentalement axée sur la recherche, on peut aussi considérer qu'elle sert le bien commun. En effet, reproduire une étude confirme les travaux scientifiques et donc le fait que les deniers publics ont été dépensés à bon escient. Cependant, cet argument ne s'applique qu'à certains types de travaux de recherche et repose sur des postulats discutables.

Définir la reproductibilité

Le problème fondamental de ce raisonnement est l'absence de consensus sur ce qu'est la « reproductibilité ». Un numéro spécial de *Science* consacré à la réplication et à la reproductibilité souligne les différences conceptuelles entre domaines. La question est particulièrement compliquée en métabiologie (qui regroupe la génomique, la transcriptomique, la protéomique et la métabolomique) où « des découvertes significatives au point de vue clinique sont cachées parmi des millions d'analyses » (Ioannidis et Khoury, 2011, p. 1230). Chaque domaine de ce champ opère des distinctions fines entre reproductibilité, validation, utilité, réplication et répétabilité, qui n'ont pas la même définition d'une discipline à l'autre. Cette variabilité est encore plus grande en élargissant la question aux sciences exactes et sociales (Ioannidis et Khoury, 2011 ; Jasny *et al.*, 2011 ; Peng, 2011 ; Ryan, 2011 ; Santer *et al.*, 2011 ; Tomasello et Call, 2011).

Une nouveauté est apparue avec la possibilité pour les scientifiques de payer pour que leur recherche et les données associées soient validées par un tiers, soit avant, soit après publication. Les débats autour de la Reproducibility Initiative portent sur les problèmes qu'elle pourrait résoudre – s'il y en a –, si elle tente de réinventer la science et si elle détourne des ressources qui pourraient être consacrées à de

nouveaux travaux. Les maisons d'édition se sont montrées peu enthousiastes au motif que les études avec appariement et les répliques ne feraient qu'encombrer la littérature scientifique (Bamford, 2012 ; Science Exchange Network, 2012).

Déterminer quoi reproduire

La difficulté à distinguer les notions de reproductibilité, de répétabilité, de réplique, de validation et de vérification se manifeste lorsqu'on décide quels éléments seront reproduits et à quelles fins. Certaines démarches tentent de réaliser une réplique exacte avec les mêmes observations, ressources, conditions, instruments, logiciels, etc., tandis que d'autres s'efforcent d'obtenir des résultats comparables avec des données d'entrée et des méthodes similaires. La première méthodologie permet de valider un article spécifique, tandis que la dernière peut produire des travaux plus valides en confirmant l'hypothèse de départ.

Il y a plusieurs niveaux de reproductibilité, par exemple la duplication précise d'observations ou d'expérience, la réplique exacte d'un processus logiciel, différents niveaux d'efforts à fournir et la nécessité éventuelle d'outils propriétaires. Les sciences numériques représentent le champ le plus prometteur pour la reproductibilité, car on peut y répéter des procédés avec exactitude pour peu qu'on dispose d'un accès suffisant aux données et aux programmes (Stodden, 2009a, 2009b, 2010 ; Vandewalle *et al.*, 2009). Les logiciels de flux de travaux scientifiques, qui maintiennent des traces détaillées des tâches de recherche, visent souvent la reproductibilité ou du moins l'établissement de registres de provenance (Bechhofer *et al.*, 2010 ; De Roure *et al.*, 2009 ; Goble *et al.*, 2013 ; Hettne *et al.*, 2012).

Les essais cliniques de médicaments et de procédures médicales sont préoccupés par la réplique et la vérification, mais reproduire intégralement représente un coût prohibitif. Un essai réussi sur une population donnée peut néanmoins être répliqué sur une autre. La médecine fondée sur des données probantes repose, plus que sur la reproductibilité proprement dite, sur l'examen systématique d'essais multiples, qui valident les résultats par la comparaison de populations, de méthodes et de résultats (Chalmers, 2011 ; Cochrane Collaboration, 2013 ; Goldacre, 2012 ; Thomas, 2013). Pour les laboratoires pharmaceutiques, essayer de répliquer des résultats rapportés dans un article de revue est souvent une première étape pour déterminer si une piste de recherche a des chances d'aboutir. Ils peuvent dépenser des dizaines, voire des millions de dollars pour reproduire une étude publiée, souvent en vain. Les avis des parties prenantes divergent : certaines estiment qu'il s'agit d'argent bien dépensé afin d'éviter des investissements erratiques, tandis que d'autres pensent que la situation révèle les failles de la méthode scientifique ; c'est notamment l'avis du *Wall Street Journal* (Naik, 2011). Toutefois, les articles journalistiques simplifient souvent

les méthodes et résultats scientifiques d'une manière qui induit en erreur, voire qui contredit les résultats effectifs (Goldacre, 2008).

Détecter la fraude

Dès que des articles publiés dans des revues majeures sont rétractés, on s'interroge sur ce que les évaluateurs et évaluatrices connaissent – ou auraient dû connaître – des données et des procédures (Brumfiel, 2002 ; Couzin et Unger, 2006 ; Couzin-Frankel, 2010 ; Normile *et al.*, 2006). À la suite d'un cas de fraude particulièrement médiatisé en psychologie, un groupe de psychologues a lancé un projet de réplication d'études récentes parues dans les principales revues du domaine. Celui-ci a provoqué des tensions, d'aucuns se demandant si ces efforts renforceraient vraiment la recherche en psychologie et si l'innovation méthodologique serait pénalisée (Carpenter, 2012 ; Doorn *et al.*, 2013 ; Enserink, 2012a).

Ce n'est qu'aujourd'hui qu'on pose la question de ce qui constitue un examen de données valide, adéquat ou raisonnable, avec peu de réponses (Borgman, 2007 ; Lawrence *et al.*, 2011 ; Parsons et Fox, 2013). Il n'est pas toujours nécessaire d'avoir accès aux données pour repérer des résultats erronés. Dans bien des cas, la fraude n'est détectable qu'après publication, par exemple lorsqu'on trouve les mêmes illustrations ou chiffres dans plusieurs articles (Couzin et Unger, 2006 ; Wohn et Normile, 2006). Distinguer entre fraude, faute et erreur est plus difficile. L'évaluation par les pairs (*peer review*) est un mécanisme de validation imparfait, mais il n'en existe pas encore de meilleur. Les évaluateurs et évaluatrices, qu'on appelle aussi arbitres scientifiques, sont tenus d'apprécier la fiabilité et la validité d'un rapport de recherche en se fondant sur les informations fournies. L'évaluation par les pairs repose moins sur la reproductibilité que sur un jugement d'expert (Arms, 2002 ; Fischman, 2012 ; Fitzpatrick, 2011 ; Harnad, 1998 ; King *et al.*, 2006 ; Shatz, 2004 ; Weller, 2001).

Clore des controverses

Les perspectives les plus ambitieuses voient dans la reproductibilité un moyen de clore des controverses scientifiques. Si un résultat peut être confirmé ou infirmé, la réponse devrait être claire, semble-t-il. Cependant, les tentatives de reproduction de résultats révèlent souvent les désaccords épistémologiques qui sous-tendent ces controverses, dévoilant ainsi le cœur de la recherche scientifique. Par exemple, Albert Wegener a avancé des théories sur la tectonique des plaques au début du ^{xx}e siècle, mais celles-ci n'ont pas été largement acceptées avant les années 1960, quand les progrès technologiques en cartographie des fonds marins ont confirmé son modèle. Wegener s'appuyait sur des faits tirés de différentes disciplines et n'avait pas une connaissance suffisamment poussée de chaque spécialité pour convaincre les scientifiques des avantages de sa théorie (Frankel, 1976). De même, en physique,

les tentatives pour résoudre la question de l'existence des ondes gravitationnelles ont échoué en raison de désaccords sur les méthodes de validation. Certains scientifiques estimaient que seules les expériences qui avaient détecté les ondes avaient été menées correctement, alors que d'autres ne faisaient confiance qu'aux travaux qui n'étaient pas parvenus à les déceler (Collins, 1975, 1998).

Pour mettre des ressources publiques à la disposition de la population

Un autre raisonnement courant en faveur de la diffusion des données est que les produits de fonds publics devraient être accessibles aux citoyennes et aux citoyens. On retrouve ce raisonnement dans les arguments en faveur du gouvernement ouvert, de l'accès ouvert aux publications et de l'accès ouvert aux données (Boulton, 2012 ; Lynch, 2013 ; Wood *et al.*, 2010). Par exemple, les politiques britanniques combinent l'accès ouvert aux publications et aux données (Research Councils UK, 2011, 2012c ; Suber, 2012b). L'Australie a incorporé la gestion des données dans son code déontologique pour la recherche (National Health and Medical Research Council, 2007 ; Australian National Data Service, 2014). Enfin, les politiques états-uniennes varient d'un organisme de financement à l'autre. Les National Institutes of Health ont instauré l'accès ouvert aux publications à travers PubMed Central avant de traiter la question des plans de gestion des données. À l'inverse, la National Science Foundation a établi des politiques de partage des données et de plans de gestion en laissant de côté l'accès ouvert aux publications. Les réglementations états-uniennes pourraient converger sous l'influence de nouvelles politiques fédérales en matière d'information de recherche, mais la route est encore longue (Burwell *et al.*, 2013 ; Holdren, 2013a, 2013b).

Le fait de confondre ces deux types de politiques a pour conséquence d'assimiler accès ouvert aux publications et accès ouvert aux données, ce qui soulève les problèmes que nous avons vus au chapitre 3. L'accès ouvert aux publications bénéficie à tous les lectorats, que l'on parle de scientifiques, de praticiennes et praticiens, de la population étudiante ou du grand public. L'accès ouvert aux données offre, quant à lui, des bénéfices directs à bien moins de gens, sans compter qu'ils varient en fonction des acteurs. Les rapports préliminaires sur l'accès ouvert aux données d'essais cliniques, par exemple, indiquent que la majorité des demandes proviennent de compagnies pharmaceutiques, de juristes et de cabinets de conseil ; quelques-unes seulement émanent de chercheurs et chercheuses universitaires (Bhattacharjee, 2012 ; Cochrane Collaboration, 2013 ; Fisher, 2006 ; Goldacre, 2012 ; Marshall, 2011 ; Rabesandratana, 2013 ; Vogel et Couzin-Frankel, 2013).

Pour optimiser les investissements dans la recherche

Un troisième raisonnement pour justifier la diffusion des données est que d'autres devraient avoir la possibilité de les exploiter. Là encore, la différence d'intérêts des parties prenantes est manifeste. Les motivations des scientifiques à diffuser leurs données sont influencées par les bénéficiaires supposés. Cette optique défend le besoin d'un plus grand nombre de référentiels qui acceptent et conservent les données de la recherche, de meilleurs outils et services pour les exploiter et d'autres investissements dans l'infrastructure de la connaissance. Par ailleurs, elle corrobore les arguments des acteurs du secteur privé qui veulent « libérer la valeur » des données produites par les universités (Biemiller, 2013 ; Thomson Reuters, 2013). Cependant, elle mélange investissements dans la recherche et investissements dans les données, qui jouent un rôle différent dans la communication savante.

On avance souvent des exemples d'usages imprévus de données anciennes pour étayer l'idée qu'il faudrait en conserver autant que possible. Par exemple, des résidus d'ADN de Louis XVI, guillotiné en 1793 avec son épouse Marie-Antoinette, ont été comparés à l'ADN d'un de ses ancêtres directs. Les analyses ont révélé que Louis était concerné par un risque génétique de diabète, d'obésité et de troubles bipolaires. Ces découvertes ont apporté des éléments aux débats sur les sources de l'indécision du roi (*Science Staff*, 2013). Les progrès réalisés dans le décodage d'ADN ancien ont fait partie des « découvertes de l'année » citées par la revue *Science* (Gibbons, 2012 ; *Science Staff*, 2012). L'anticipation d'applications futures constitue un problème classique des bibliothèques, des musées et des archives, mais même eux ne peuvent pas tout garder.

Pour faire progresser la recherche et l'innovation

Le dernier raisonnement veut que le partage des données fasse progresser la recherche et l'innovation. À la différence des deux précédents, il s'intéresse aux bénéfices pour la science. Cette logique est implicite dans des phrases telles que « la science repose sur des données de qualité » (Whitlock *et al.*, 2010, p. 145) et « la conservation scientifique des données est un moyen de les organiser, de les valider et de les préserver pour que les scientifiques trouvent de nouveaux moyens de répondre aux grandes questions de recherche auxquelles notre société est confrontée » (Data Conservancy, 2010). Elle constitue un argument pour investir dans l'infrastructure de la connaissance pour pérenniser les données de la recherche, qui seraient organisées selon des normes professionnelles exigeantes. Toutes les données n'ont pas besoin d'être gardées, mais celles qui en valent la peine doivent être conservées au mieux et devenir des atouts pour la communauté scientifique.

La demande de données de recherche

Les quatre raisonnements justifiant le partage des données partent du principe qu'elles seront réutilisées : si les données sont partagées, les utilisateurs et utilisatrices viendront. Ce postulat sous-tend les discours selon lesquels partager les données permettrait de « libérer leur valeur » (Thomson Reuters, 2013). Or, les données de la recherche sont des objets sociotechniques complexes qui existent au sein de communautés et non de simples marchandises que l'on peut échanger sur un marché public. Si les scientifiques cherchaient activement des données à réutiliser, elles seraient davantage partagées. C'est dans les domaines où la dépendance mutuelle est la plus importante que la demande en données réutilisables est la plus forte, comme expliqué au chapitre 4. En effet, ces champs ont lourdement investi dans les infrastructures de la connaissance pour partager l'instrumentation, les ensembles de données et d'autres ressources. Comme nous l'avons vu dans les études de cas, de nombreux aspects du travail scientifique des données semblent influencer sur le moment, la manière et le fait de diffuser ou réutiliser des données.

Dans certains domaines, on dissuade activement les chercheurs et chercheuses de réutiliser des données. Pour paraphraser Jamie Callan et Alistair Moffat (2012), la réutilisation revient à chercher ses clés sous un lampadaire parce que c'est là qu'il y a de la lumière : les données et les questions intéressantes se trouvent ailleurs. On bâtit sa carrière en explorant des territoires vierges. Or, poser de nouvelles questions à partir de nouvelles données est la manière la plus sûre d'innover. Certes, poser de nouvelles questions à partir de données anciennes peut aussi conduire à des découvertes, mais il est difficile de convaincre les comités de rédaction et les arbitres scientifiques que les réanalyses sont des contributions de valeur. Poser les mêmes questions en répliquant des études n'amène guère de reconnaissance et il est particulièrement difficile de se faire publier dans ces conditions.

Les études de cas en sciences exactes et en sciences sociales laissent à penser que l'échange privé de données serait plus courant que l'échange public à travers des référentiels. Les chercheurs et chercheuses indiquent qu'ils cherchent rarement à obtenir des données d'autres scientifiques et qu'on ne leur demande guère les leurs (Faniel et Jacobsen, 2010 ; Pienta *et al.*, 2010 ; Wallis *et al.*, 2013). Lorsqu'on les sollicite en ce sens, les chercheurs peuvent répondre favorablement ou non en fonction de divers facteurs, tels que les ressources nécessaires pour partager les données, leurs rapports avec le demandeur ou la demandeuse et le temps écoulé depuis la création des données (Campbell *et al.*, 2002 ; Hanson *et al.*, 2012 ; Hilgartner et Brandt-Rauf, 1994 ; Hilgartner, 1997). Mayernik (2011) a découvert que certaines

données n'étaient pas diffusées parce que les chercheurs ne voyaient pas qui en voudrait ou à quoi elles pourraient servir.

Le plus difficile lorsqu'on évalue les pratiques de réutilisation des données est l'absence de consensus sur la notion de « réutilisation ». Cette dernière dépend elle-même de ce que signifie l'« utilisation » des données ou d'autres formes d'informations. La recherche d'information, les besoins et les usages sont des problèmes anciens et épineux en sciences de l'information. Puisqu'aucune définition d'« utilisation d'information » ne s'applique de manière satisfaisante à l'ensemble des disciplines et des contextes, l'absence d'unanimité sur ce qui constitue l'« utilisation » ou la « réutilisation » des données n'est pas surprenante. Après de nombreuses années d'analyses, les archivistes en astronomie sont parvenus à un consensus international sur les bibliographies de télescope, mais sans définir clairement ce que signifie « utiliser les données » d'un télescope (IAU Working Group Libraries, 2013). Ce problème de définition se répercute : les archives diffèrent quant à ce qu'elles considèrent comme des publications « évaluées » et à leur manière de mesurer des concepts clés, comme « observation ». Bien que les mesures de l'usage des données soient plus cohérentes en astronomie que dans la plupart des champs, elles ne sont pas transférables à des domaines proches, ni comparables d'une discipline à l'autre.

En somme, le modèle de l'offre et de la demande ne décrit pas adéquatement les défis de la divulgation, du partage et de la réutilisation des données. Une analyse autrement plus profonde est nécessaire pour éclairer les motivations intrinsèques et extrinsèques, ainsi que les parties prenantes de ces activités.

Les motivations scientifiques

Le partage des données est difficile à réaliser et le retour sur investissement ne suffit pas à le justifier. Peu de scientifiques travaillent en ayant la réutilisation des données en tête. Elles sont un moyen en vue d'une fin, laquelle consiste généralement à chercher des résultats qui seront rapportés dans des publications, et non une fin en soi. Traiter les données comme des produits à diffuser ou partager suppose des changements dans les méthodes et les pratiques de recherche. Dans certains cas, les pratiques peuvent être adaptées, mais les problèmes insolubles de propriété, d'éthique, de ressources et de normes communautaires sont légion. Dans la plupart des domaines, les scientifiques disposent de ressources minimales pour investir dans le partage de leurs données, et même quand ils s'en occupent, il n'existe pas de référentiels pour les prendre en charge. Le succès des politiques encourageant la diffusion des données dépendra sans doute d'investissements bien plus radicaux

dans les infrastructures de la connaissance que les organismes de financement, les revues, les bibliothèques, les archives ou les scientifiques ne veulent l'admettre.

Publications et données

Dans la recherche, les données jouent un rôle très différent des publications, comme nous l'avons évoqué au chapitre 3. L'amalgame entre l'accès ouvert à la littérature et l'accès ouvert aux données embrouille les relations complexes entre ces deux réalités. Ces rapports évoluent, quoique lentement, comme nous le verrons au chapitre 10.

Communiquer la recherche

La communication de la recherche a commencé bien avant les livres et les revues. Les discussions « souvent précieusement préservées dans des manuscrits maintes fois recopiés » (Meadows, 1998, p. 3) des Grecs contemporains d'Aristote influencent la pensée scientifique encore aujourd'hui. La communication savante s'est progressivement formalisée et professionnalisée. Les livres ont été les premiers à émerger comme forme longue. Les revues ont été créées au ^{xvii}^e siècle pour formaliser la correspondance entre savants. Ces genres littéraires ont évolué jusqu'à trouver leurs formes actuelles. Les articles présentent une structure formelle qui comprend les mentions d'auteur, les titres, les résumés, les références à d'autres publications, les figures, les tableaux, etc. Cette structure a divergé dans les différents champs, mais tout article traitant d'un domaine de recherche quelconque conserve ces éléments clés. De plus, les genres diffèrent peu entre les formats papier et en ligne.

Les livres, les articles de revue, les actes de colloque et les autres publications permettent à la communauté d'examiner les résultats de recherche. Ils comportent suffisamment de détails sur les arguments, les méthodes, les sources de preuve et les procédures pour en étayer les conclusions, mais rarement assez pour pouvoir reproduire les résultats. Dans certains domaines, les revues publient des tableaux et des figures détaillés. Dans d'autres, les articles consistent principalement en un argumentaire où l'explication des sources de preuve est réduite à portion congrue.

Les publications simplifient nécessairement les activités de recherche, omettant des détails jugés inutiles à la compréhension de l'argumentaire et des preuves par le lectorat cible. Les éléments passés sous silence peuvent être, dans les sciences exactes, les réglages des appareils, les étalonnages, les filtres et les protocoles ; dans les sciences sociales, les méthodes d'analyse statistique, la réduction des données, la conception de l'entretien et les protocoles de terrain ; dans les sciences humaines, les méthodes de recherche dans les archives, la traduction et l'accès au terrain. S'ils

ne sont pas forcément utiles à la compréhension de l'article, ces détails peuvent s'avérer nécessaires pour comprendre les données.

Les publications ne sont donc pas de simples contenants de données. Les chercheurs et chercheuses formulent leur discours de manière à convaincre autrui de la validité et de l'importance de leur travail. Les publications sont des argumentaires étayés par des preuves. Elles sont les produits finis au terme d'un long processus de négociations entre partenaires, financeurs, arbitres, comités de rédaction, maisons d'édition, présidentes et présidents de programmes et collègues. Tout résultat d'étude est présenté de manière stratégique pour plaire au lectorat de la revue à laquelle un article est proposé, au comité de rédaction d'une maison d'édition à laquelle un manuscrit est envoyé ou à l'assistance d'un colloque. Un même résultat peut être adapté à de multiples publics. Les auteurs et autrices choisissent leur terminologie, leurs questions de recherche, leurs données, leurs tableaux et figures, la longueur et la structure de leur manuscrit et leurs arguments pour chaque publication. Même le format des références diffère d'un article à l'autre (Bowker, 2005 ; Kwa, 2011 ; Latour et Woolgar, 1979, 1986 ; Lynch et Woolgar, 1988a ; Merton, 1963b, 1970, 1973 ; Star, 1995 ; Storer, 1973).

Identifier « les données » associées à un article, une communication de colloque, un livre ou une autre publication est souvent difficile. Les chercheurs et chercheuses recueillent des données en continu, sur de longues périodes et à l'aide de multiples méthodes pour étudier un ensemble de problèmes interdépendants. Ils s'appuient sur des jeux de données cumulatifs, plus ou moins documentés, pour avancer différents arguments dans chacune de leurs publications. Ils extraient ou non un jeu de données distinct pour un article particulier. Expliquer comment un jeu de données a été créé peut nécessiter d'expliquer les méthodes employées dans toute une série d'articles antérieurs et de phases préliminaires de réduction des données. La règle générale en matière de publications est que celles-ci doivent être suffisamment distinctes pour être compréhensibles isolément, impliquant qu'une interprétation plus complète peut nécessiter la lecture des références citées.

Une fois publiés, ces produits finis prennent une vie propre. Chaque lecteur ou lectrice apporte ses propres connaissances au document et en tire sa propre signification. Une personne lira l'article pour ses résultats, une autre pour ses méthodes, une autre encore pour une figure spécifique et encore une autre pour la revue de la littérature. Un lecteur travaillant dans le même domaine de spécialité lira la prépublication pour se tenir au courant des nouvelles découvertes. Une lectrice dans un domaine connexe lira l'article un an après sa parution pour comparer ses résultats aux derniers développements de sa discipline. Un lecteur d'un champ différent le lira

en quête de méthodes ou de résultats transférables. Cinq ou dix ans plus tard, une doctorante évaluera l'article pour sa thèse : entre-temps, il aura pu être jugé comme une œuvre majeure, une fraude ou une découverte fallacieuse qui aura entraîné la discipline dans une impasse (Brown et Duguid, 1996, 2000 ; Latour et Woolgar, 1986).

Publier la recherche

La diffusion des jeux de données avec chaque publication n'est jamais devenue la norme dans la communication savante. Du point de vue des maisons d'édition, les arguments en faveur de leur incorporation dans le processus éditorial ne sont pas les mêmes pour l'imprimé et le numérique. Dans l'univers du papier, le coût de reproduction de vastes jeux de données est prohibitif. Les auteurs et autrices sont tenus de se limiter à un nombre de pages fixes, lesquelles sont généralement dévolues au déroulement de l'argumentaire avec le minimum de données probantes. Les sections sur la méthodologie sont souvent portion congrue pour des raisons d'espace et parce qu'on suppose le lectorat suffisamment expert dans le domaine pour ne pas avoir besoin de beaucoup de détails.

Dans l'univers du numérique, l'économie n'est plus la même. Les contraintes d'impression et de pagination ne sont plus une préoccupation majeure. Cependant, les revues qui publient à la fois sur papier et en numérique, comme *Science* et *Nature*, tendent à conserver un périmètre identique pour chaque article. Ainsi, une limite à quatre pages dans la version papier se traduit par un article numérique de même longueur. Ces contraintes sont en partie contournées en demandant des « informations complémentaires » (*supplemental information*) qui ne sont publiées qu'en ligne ; celles-ci peuvent comprendre des éléments essentiels sur la méthodologie et les données. Les informations complémentaires sont restées largement cantonnées à l'édition en sciences exactes et ne se sont jamais démocratisées dans les sciences humaines et sociales. Les évaluateurs et évaluatrices peuvent demander davantage d'éléments sur le déroulement d'une expérience, par exemple, et les auteurs doivent satisfaire à ces demandes dans le cadre de la politique éditoriale de la revue. Certaines revues imposent des règles strictes sur les éléments à faire figurer dans le corps du texte, à inclure dans les informations complémentaires, à publier ailleurs et à lier, ainsi que sur la présentation des rapports entre ces éléments. À mesure que les jeux de données et les autres formes de preuve gagnent en volume, les maisons d'édition s'inquiètent de devoir assurer à long terme les coûts de stockage et de conservation. Au point de vue de la réutilisation, le plus problématique reste les jeux de données derrière des *pay walls* (accès payant) ou postés sur les sites web des auteurs, car ils ne sont pas facilement découvrables.

Des atouts et des charges

Exiger la divulgation des données au moment de la publication suppose que les auteurs et autrices « en aient fini » avec elles. La diffusion des données est très courante dans des champs comme la génomique, où la correspondance entre la publication et le jeu de données est plutôt claire et où « les jeux de données sont relativement simples, homogènes et bien définis » (Shotton, 2011). Lorsque la carrière du chercheur ou de la chercheuse est concentrée sur l'étude à long terme d'une espèce, d'un lieu ou d'un ensemble d'artefacts particulier, les données gagnent en valeur à mesure de leur accumulation. Les chercheurs peuvent ne jamais en avoir fini avec leurs données. Ils sont réticents à diffuser les jeux associés à une publication parce que cela peut impliquer de divulguer des années de données. De même, reproduire des données associées à une publication quelconque est problématique, car l'ensemble des observations rapportées peut dépendre largement d'études antérieures et de l'interprétation de données bien plus anciennes.

Les jeux de données représentent aussi des atouts qui peuvent être échangés avec d'autres scientifiques, utilisés comme leviers dans des collaborations et offerts en dot. Les chercheurs et chercheuses en sciences humaines peuvent ainsi apporter des mines de ressources à un nouveau poste ou à de nouvelles collaborations. Si les données sont diffusées, elles perdent leur valeur en tant que monnaie d'échange. Il est difficile de persuader les chercheurs de renoncer à ces atouts sans qu'ils obtiennent quelque chose en retour, surtout si leur collecte a nécessité d'importants efforts et dépenses (Borgman, 2007 ; Edwards *et al.*, 2011 ; Hilgartner et Brandt-Rauf, 1994). Le moment est un autre facteur dans la diffusion des données. Les personnels de recherche protègent leurs collaborations et leur investissement dans leur travail en ne divulguant pas leurs données plus tôt qu'il n'est nécessaire. Si elles sont diffusées trop tôt, les partenaires peuvent renâcler à engager les sommes et le travail salarié nécessaires à la collaboration. L'objectif des périodes d'embargo est de trouver un compromis entre la prérogative des chercheurs et une diffusion rapide.

Les données représentent aussi une charge. Selon leur type, les stocker suppose un espace physique et des ressources informatiques. Les gérer exige des ressources humaines. Le temps consacré à les préparer pour leur diffusion est perdu pour les autres activités scientifiques. De même, les fonds des subventions dévolus à la gestion des données ne sont pas utilisables pour mener des recherches. Les droits de diffusion et le devoir de conservation peuvent s'avérer inextricables ; les questions juridiques et économiques imprègnent donc la problématique du partage des données. Souvent, les données ne sont pas diffusées en l'absence de certitude sur qui a le droit ou la responsabilité de le faire. Les chercheurs et chercheuses peuvent ignorer si les données

leur appartiennent ou si elles sont possédées par leur université, leur laboratoire ou département, leur organisme de financement, leur maison d'édition ou une autre instance. La réponse peut ne pas avoir d'importance tant que la question de la diffusion des données n'est pas soulevée. Plus il y a de partenaires et de juridictions impliqués, plus il est difficile de déterminer les droits de diffusion des données (Arzberger *et al.*, 2004 ; Berman *et al.*, 2010 ; Birnholtz et Bietz, 2003 ; Hirtle, 2011 ; Reichman *et al.*, 2009 ; Stanley et Stanley, 1988).

Les scientifiques s'inquiètent aussi de leur responsabilité légale et des risques pour leur réputation si les données sont mal utilisées ou mal interprétées. Le mésusage peut sembler un problème mineur puisque les scientifiques ont malheureusement l'habitude de voir leur travail compris de travers dans les médias de masse (Goldacre, 2008). Pourtant, l'extraction sélective de données, la réidentification de sujets humains ou d'animaux de laboratoire et d'autres formes de mésusage – qu'elles soient dues au manque d'expertise, à une documentation insuffisante ou à de la malveillance – représentent un vrai problème. La recherche mondiale et comparative sur le changement climatique dépend de l'accès ouvert aux données, mais la politisation de ces travaux fait hésiter les chercheurs et chercheuses de ces domaines à diffuser leurs données.

Diffuser les données

Les chercheurs et chercheuses éprouvent souvent de grandes difficultés à gérer leurs données pour leur usage présent et futur. Disposer des moyens de gérer efficacement des données est un prérequis pour les diffuser de manière à les rendre interprétables par d'autres scientifiques.

Représentations et mobilité

La difficulté à extraire les données du contexte de recherche est connue comme le problème de la « mobilité ». Rendre des données mobiles suppose de les séparer de la situation qui en a fait des données. Par conséquent, on perd une couche de signification. Consigner la provenance peut permettre de la rétablir en partie, mais le contexte précis est impossible à transférer intégralement. Les données peuvent constituer des « objets-frontières » qui délimitent les communautés. Elles prennent des sens différents à mesure qu'elles sont utilisées, décrites, documentées et interprétées d'un domaine à l'autre et d'une chercheuse à l'autre (Bishop et Star, 1996 ; Bowker, 2005 ; Kanfer *et al.*, 2000 ; Star *et al.*, 2003 ; Star et Griesemer, 1989 ; Star, 1983, 1989).

Comme expliqué au chapitre 2, les données sont des entités qui servent à mettre en évidence des phénomènes à des fins de recherche. Elles ne peuvent que rarement être séparées de leurs représentations. Que les chercheurs et chercheuses étudient un enfant dans un parc ou des expressions dans un texte chinois bouddhique, ils représentent leurs données d'une façon particulière pour prouver un phénomène donné. Une fois les représentations intégrées dans un fichier STATA, Word, R ou autre, elles perdent ce qui en a fait des preuves de phénomènes particuliers.

Les logiciels peuvent constituer une source de friction dans l'échange de données. Au CENS, par exemple, certaines équipes utilisaient R, un progiciel statistique *open source* répandu, quand d'autres préféraient Matlab ou d'autres outils. Pour échanger leurs données, ces groupes exportaient leurs données dans des tableaux Excel qui faisaient office de plus petit dénominateur commun. D'autres équipes partageaient leurs données dans des tableurs Google et des méthodes semblables. Le CENS se situe du côté de la science légère, mais Vertesi et Dourish (2011) ont observé le même problème en astronomie, où les groupes devaient réduire leurs données à leur plus petit logiciel commun pour pouvoir les échanger. De même, les données qualitatives dans des outils analytiques comme NVIVO ou Atlas.ti peuvent être exportées dans des tableurs ou des traitements de texte à condition de renoncer à leur structure. Ces réductions ont pour conséquence une perte de puissance d'analyse considérable. Les tableurs utilisés pour échanger des données sont cependant omniprésents au point d'être devenus un format courant pour le dépôt (California Digital Library, 2013).

Des normes et standards communs permettent l'utilisation de suites logicielles pour analyser, visualiser, documenter, partager et réutiliser des données. Ils aboutissent à l'interopérabilité entre matériels, logiciels, outils, protocoles et pratiques au sein de communautés et entre elles. Le revers de la médaille est que ces normes créent des silos de données et de systèmes incompatibles entre eux. Les formats représentant le plus petit dénominateur commun peuvent alors constituer le seul moyen d'échanger des données. Les silos sont un problème canonique des sciences de l'information. De nombreux mécanismes de conversion, humains ou techniques, existent, mais ils restent des traductions (Busch, 2013 ; Lampland et Star, 2009 ; Libicki, 1995 ; *Nature Staff*, 2012 ; Star, 1991).

Partager des données dans l'optique d'une réutilisation immédiate, comme lors d'échanges entre collaborateurs et collaboratrices, est plus facile que de les diffuser à l'intention d'utilisateurs et utilisatrices inconnus qui en auront besoin à une période indéterminée, potentiellement des années plus tard. Dans ce cas, la documentation et les représentations constituent la seule communication entre le créateur ou la

créatrice des données et l'utilisateur. Ce dernier peut être un autre chercheur ou bien un ordinateur. Les *data miners* (explorateurs et exploratrices de données) et les robots ont leurs propres critères de « lecture », d'indexation et d'extraction. Plus la distance entre l'utilisateur et le créateur est grande, que ce soit au point de vue temporel, disciplinaire ou linguistique, plus il est difficile d'interpréter l'argumentaire, la méthodologie, les données, les résultats et le contexte d'une étude ou d'un jeu.

Les politiques de partage des données ne s'appliquent pas de la même manière aux données numériques et aux objets matériels. Certaines politiques font en effet expressément référence aux échantillons et aux spécimens. Ceux-ci sont envoyés par la poste et constituent en ce sens une forme de partage de données entre laboratoires en sciences de la vie. On réalise parfois un partage de données en envoyant du personnel à un laboratoire, des archives ou un terrain où se trouvent des données. Des accords de transfert matériel régissent habituellement ce type de partage.

Provenance

Des questions de provenance – au sens archivistique d'« historique de conservation » et au sens informatique de « transformations par rapport à l'état d'origine » – se posent lorsqu'on interprète des données qui ont évolué à travers plusieurs mains et processus (Buneman *et al.*, 2000 ; Carata *et al.*, 2014 ; Groth *et al.*, 2012). Les données et leurs représentations tendent à correspondre à des logiciels et, souvent, à du matériel. Les réseaux de capteurs, le traitement en *pipeline* en astronomie et l'examen au microscope électronique en métallurgie sont quelques exemples de cet entrelacement inextricable des processus matériels et logiciels. Répliquer ces procédures ou en interpréter les résultats implique d'accéder au logiciel. Or, qu'ils soient *open source* ou propriétaires, les logiciels changent continuellement. Documenter les données peut nécessiter de saisir des éléments sur chaque version du logiciel et chaque configuration des plateformes où ces données ont été manipulées. L'interprétation des données de sortie peut être difficile parce que les outils d'analyse et de traitement – logiciels statistiques, mathématiques et de flux de travaux – gardent rarement une trace précise des systèmes ou des données à chaque transaction (Claerbout, 2010 ; Goble et De Roure, 2009).

Certains des logiciels nécessaires à l'interprétation des données peuvent être des outils développés sur mesure par les scientifiques pour recueillir, traiter, analyser et documenter leurs données. Les chercheurs et chercheuses écrivent du code pour analyser les phénomènes qu'ils étudient, que ce soient des modèles climatiques ou des séquences de symboles sur des papyrus. Ce code représente parfois un investissement majeur nécessitant le recours à du personnel pour la programmation, mais il consiste plus souvent en des scripts écrits dans des outils statistiques, de

visualisation ou autres. En informatique, les logiciels sont parfois « les données » qui résultent d'un projet.

Les chercheurs et chercheuses sont souvent encore plus réticents à diffuser leur code que leurs données. Les raisons en sont multiples. Pour commencer, ils s'inquiètent du caractère « brouillon » de leur code. En effet, le code utilisé en recherche est souvent mal documenté et son exécution n'est pas toujours efficiente. Conçu pour être utilisé immédiatement, il est rarement documenté pour être réutilisé, notamment dans les projets ponctuels. Un autre sujet d'inquiétude est le contrôle du programme informatique. L'investissement intellectuel du chercheur dans son code transcende le jeu de données ou la publication. Les modélisations de phénomènes ou d'autres types de code analytique représentent en effet un avantage concurrentiel. Les résultats des modèles et les paramètres sont souvent partagés sans diffuser le code qui les sous-tend.

Un aspect peu exploré de la diffusion et la réutilisation des données est l'influence des décisions de traitement préliminaires sur les interprétations d'un jeu de données. Les choix initiaux sur la manière de nettoyer et de réduire les données peuvent déterminer à quel degré la provenance des données peut être récupérée. Dans des domaines comme l'astronomie, la biologie et la métaomique, où la réutilisation est courante, les chercheurs et chercheuses nettoient et réduisent les données pour bâtir une base. Comme l'expliquent Alexander Blocker et Xiao-Li Meng (2013), les décisions de prétraitement « peuvent être périlleuses » au point de vue statistique, car la plupart sont irréversibles. Les partis pris initiaux contraignent les analyses de toutes les phases ultérieures : les réutilisations par les scientifiques en aval reposent donc sur eux. Toute inférence ne saurait être plus valide que les partis pris sur lesquels le traitement a été fondé dans les premières phases ; or, les connaître est parfois impossible. Bien des controverses scientifiques remontent au problème statistique de l'inférence sur plusieurs phases. Même des collaborateurs et collaboratrices peuvent éluder ces partis pris. La réutilisation est plus risquée encore pour les scientifiques ne faisant pas partie de la collaboration ou du domaine, puisqu'elles et ils sont moins au fait des questions à poser sur les décisions de prétraitement (Blocker et Meng, 2013 ; Meng, 2011).

Les décisions de prétraitement et les autres formes de nettoyage et de réduction des données sont prises dans l'optique de questions de recherche et de phénomènes particuliers. Lorsque ces données sont réutilisées pour des questions différentes, connaître les divers choix de prétraitement peut devenir une nécessité. Par exemple, de nouvelles théories sur le cancer suggèrent que les réponses aux traitements sont largement individuelles. Or, les cas où l'efficacité d'un médicament ou d'un autre

traitement est exceptionnelle peuvent être exclus de l'analyse des données, car les observations situées aux extrémités d'une courbe sont généralement traitées comme des aberrations statistiques. Un nouveau projet va étudier ces valeurs aberrantes en explorant des données anciennes afin de générer des hypothèses qui expliqueraient pourquoi des médicaments fonctionnent sur certaines personnes et pas sur d'autres (Kaiser, 2013). Ces décisions ont toutefois un prix. La NASA conserve ses données à plusieurs niveaux de traitement, comme expliqué au chapitre 2. Peu de scientifiques ont les moyens de tenir à jour leurs ressources en données à différents niveaux ou avec une documentation suffisante pour réanalyser ou réutiliser chaque étape du traitement.

Il vaut mieux envisager la provenance des jeux de données comme un réseau que comme une séquence linéaire de relations. En fonction du domaine, un jeu peut être lié à des objets de recherche tels que des logiciels, des instruments, des protocoles, des carnets de laboratoire, de la documentation technique et une ou plusieurs publications. Relier ces objets rend chacun d'eux plus utile et contribue à leur découvrabilité. Lier des publications aux jeux de données correspondants est particulièrement utile, parce qu'ils s'apportent mutuellement une valeur ajoutée. Les articles représentent la principale source de découverte de la plupart des jeux de données et peuvent constituer leur seule documentation publique. Créer de tels liens est un investissement dans la gestion de données, puisqu'il faut décrire chaque objet de recherche et préciser leurs relations. Des technologies comme les données liées peuvent aider à enregistrer et trouver ces relations, mais restent une solution incomplète (Bechhofer *et al.*, 2010, 2013 ; Borgman, 2007 ; Bourne, 2005 ; Parsons et Fox, 2013 ; Pepe *et al.*, 2010).

Acquérir des données à réutiliser

L'enchantement à l'idée de pouvoir réutiliser des échantillons d'ADN, des mesures climatiques, des flux de transport ou du trafic mobile et de pouvoir déterminer les récurrences d'explosion de plaques d'égout (Anderson, 2006, 2008 ; Mayer-Schonberger et Cukier, 2013) fait trop souvent oublier les risques posés par la réutilisation de données. La majeure partie du travail de recyclage de données anciennes est consacré à leur nettoyage, à leur codification et à leur vérification. Comme le remarquent Danah Boyd et Kate Crawford (2012, p. 668), « les données les plus volumineuses ne sont pas toujours les meilleures ». Augmenter la taille de la meule de foin ne rend pas l'aiguille plus facile à trouver ; les chercheurs et chercheuses se montrent donc prudents au moment de décider s'ils acquerront des données auprès d'autres scientifiques, de référentiels ou d'autres sources externes (J. Boyle, 2013 ; Eisen, 2012 ; Kwa et Rector, 2010).

Utilisations de premier plan et d'arrière-plan

Il arrive que les ressources en données ne soient pas citées dans les publications scientifiques, que des jeux de données y soient joints ou non. L'absence de renvoi ne signifie pas qu'aucune donnée n'a été utilisée dans l'étude, mais simplement qu'aucune utilisation ou réutilisation n'a été mentionnée. Les données acquises afin d'étalonner les instruments, de vérifier les mesures ou d'évaluer les conditions sur le site peuvent ne pas être jugées dignes d'une mention ni être considérées comme une « utilisation » au sens des pratiques disciplinaires. Chercher des informations dans ses fichiers personnels, des bibliothèques, des musées, des archives, des référentiels numériques, des greniers, des entrepôts ou auprès d'amis et de collègues fait pourtant partie des activités de recherches. Cette dimension de la collecte de preuves tend à ne pas être documentée explicitement dans les publications – quoique les pratiques varient en la matière. Les conclusions d'une étude peuvent dépendre de nombreuses recherches, sources, ressources, utilisations et réutilisations qui ne seront pas mentionnées, car jugées inutiles au propos.

C'est en partie pour cela que l'utilisation et la réutilisation des données sont, en pratique, difficiles à consigner. Celles et ceux qui recueillent de nouvelles données pour répondre à leurs questions de recherche peuvent aussi chercher des informations pour des raisons plus générales, « d'arrière-plan ». Par exemple, dans les études du CENS et des astronomes, ce type d'informations était tiré de référentiels de données, de relevés astronomiques, de bilans météorologiques, de documents administratifs relatifs à l'affectation des terres ou, occasionnellement, d'autres scientifiques. Bien qu'elles soient essentielles à l'interprétation des résultats de l'étude, ces sources ne sont généralement pas mentionnées ni explicitement citées dans les articles. En revanche, les nouvelles sources de données recueillies par les chercheurs et chercheuses se situaient au premier plan de leur travail et sont mentionnées (Borgman *et al.*, 2012 ; Wallis *et al.*, 2012 ; Wynholds *et al.*, 2012).

Ces éléments initiaux sur l'usage de premier plan et d'arrière-plan des données laissent à penser que la réutilisation – au sens de l'utilisation de données n'ayant pas été recueillies par les auteurs et autrices de l'étude – pourrait être plus élevée qu'il n'y paraît. Comparer ses données à des sources externes peut être considéré comme une bonne pratique et constituer une forme de savoir méthodologique implicite qui n'a pas besoin d'être mentionné dans les publications. Comme nous le verrons au chapitre 9, le taux de citations de données semble en augmentation dans certains domaines. Nous ne savons cependant pas si c'est la réutilisation des données qui augmente ou si citer ses sources de données est en train de devenir une bonne pratique répandue ; il est probablement impossible de le savoir. Les questions à poser semblent être : à quel moment quelque chose est considéré comme une donnée,

quels types de preuves méritent d'être rapportés dans les publications et quelles parties du processus de recherche peuvent être considérées comme du savoir implicite au sein d'une communauté. Au point de vue psychologique, la perception figure-fond pourrait être un facteur : certains objets avancent au premier plan en tant que données quand d'autres restent en arrière-fond en tant que contexte ou comme un simple bruit ambiant.

Interprétation et confiance

La capacité à faire confiance aux données des autres est cruciale, car les chercheurs et chercheuses jouent leur réputation sur leurs choix de données. Les publications et les réseaux personnels constituent les principaux moyens de découverte de données, notamment hors des disciplines qui ont recours à des référentiels partagés. Lorsque c'est possible, les chercheurs contactent les auteurs et autrices pour aborder les méthodes de création, de nettoyage, de traitement, d'analyse et de communication des données (Faniel et Jacobsen, 2010 ; Wallis *et al.*, 2013 ; Zimmerman, 2003, 2007). Des communautés travaillant sur des problèmes connexes peuvent verser des données dans plusieurs référentiels. Les données de l'une peuvent être cachées aux autres à cause de différences dans les métadonnées ou les mécanismes de classification ou faute de connaître les ressources qui existent hors de la communauté. Quand un chercheur trouve une publication pertinente en cherchant un élément spécifique, comme la structure d'une protéine, il peut en contacter les auteurs plutôt que de consulter des archives où l'élément recherché a pu être déposé.

Un nombre étonnamment faible de travaux, étant donné l'intérêt suscité par le sujet, ont étudié les circonstances, les motivations ou les pratiques de réutilisation de données. Faniel et Jacobsen (2010, p. 357), dans leur étude sur les spécialistes du génie parasismique, ont identifié trois questions utilisées pour évaluer le caractère réutilisable de données « 1) les données sont-elles pertinentes, 2) compréhensibles, 3) fiables ». Ces ingénieures et ingénieurs recherchent généralement des paramètres techniques auxquels ils pourront comparer leurs propres observations. Ils semblent donc chercher des « données d'arrière-plan » à des fins similaires de celles des chercheurs en astronomie et en réseaux de capteurs (Wallis *et al.*, 2012 ; Wynholds *et al.*, 2012).

Les infrastructures de la connaissance

Des investissements stratégiques dans les infrastructures de la connaissance, tels que des référentiels, des ressources humaines expertes en gestion de données, de meilleurs outils et des méthodes pour reconnaître le travail de partage, pourraient

augmenter la diffusion et la réutilisation des données. Pour être efficaces, les infrastructures de la connaissance doivent pouvoir prendre en charge les nombreuses formes de données diffusables et les multiples manières dont elles peuvent être utilisées, ainsi que trouver des moyens de transmettre les connaissances relatives à ces données aux contributeurs et contributrices et aux utilisateurs et utilisatrices. Elles doivent également permettre la conservation et l'accès aux ressources informationnelles pour une durée appropriée.

Certains des premiers rapports d'orientation sur le partage des données faisaient état du manque d'infrastructure pour soutenir la diffusion et la réutilisation. Un rapport de 1985 des National Academies of Science états-uniennes, intitulé *Sharing Research Data* (« Partager les données de la recherche »), a formulé des recommandations qui devaient modifier les pratiques des organismes de financement, des auteurs et autrices, des personnels de recherche, des rédactions de revues, des arbitres, des archives de données, des universités et des bibliothèques (Fienberg *et al.*, 1985). De même, la longue liste de principes que l'OCDE a publiée en 2007 traduit les préoccupations en matière d'infrastructure de nombreux acteurs. Bâtir des systèmes d'information – ou des infrastructures de la connaissance – qui donnent envie d'être utilisés est bien plus difficile qu'il n'y paraît (Markus et Kell, 1994). Les infrastructures ne se font pas en un jour, pas plus qu'elles ne sont statiques : elles sont des écosystèmes complexes comportant de nombreuses parties mouvantes. Elles doivent s'adapter aux évolutions des pratiques, des politiques, des technologies et des parties prenantes. Les principales composantes des infrastructures de la connaissance sont identifiables, même si toutes les relations entre elles ne le sont pas.

Référentiels, ensembles et archives

Demander aux scientifiques de diffuser leurs données pose la question de comment et où ces données seront accessibles et pour combien de temps. La réponse la plus évidente, bien qu'incomplète, est de créer davantage de référentiels où les chercheurs et chercheuses pourront déposer leurs données. Qu'on les appelle des archives, des collections, des ensembles, des systèmes de données, des banques de données, des systèmes d'information ou des référentiels, la technologie n'y constitue qu'un point de départ. Ils sont des réservoirs communs de ressources qui nécessitent des modèles de gouvernance : politiques de développement des ensembles, règles de contribution et d'accès, normes de classification et de structure de données et dispositifs de pérennisation. Les ensembles et leur gouvernance sont associés à des communautés, mais les frontières de ces dernières sont poreuses et se chevauchent.

La constitution de référentiels obéit souvent à une logique ascendante : ils trouvent leur origine dans des ensembles de recherche qui répondent aux besoins de petits laboratoires ou communautés, comme évoqué au chapitre 2. Ces types d'ensembles favorisent le partage le temps de la collaboration, mais peuvent ne pas être pérennisés. Ceux qui attirent des communautés plus vastes peuvent devenir des ensembles ressources. Enfin, ceux qui connaissent la plus grande réussite deviennent des ensembles de référence, qui bénéficient d'un financement à long terme et peuvent établir des normes dans leurs communautés (National Science Board, 2005). D'autres ensembles, qui représentent un but stratégique dans le cadre de grands projets, sont instaurés grâce à une logique descendante. C'est le cas des missions astronomiques, en particulier pour les télescopes spatiaux. L'archivage des données consomme une part relativement faible du budget des grandes missions et rend possible leur utilisation à long terme. Elles conservent leur valeur bien après la mise hors service des instruments.

Les données peuvent être représentées de multiples manières dans l'espace et le temps. Les chercheurs et chercheuses peuvent proposer leurs jeux de données à plusieurs référentiels ou à aucun. Leurs choix dépendent des exigences des organismes de financement, des revues et des contrats institutionnels, ainsi que des politiques de collecte de chaque référentiel. En effet, certains référentiels recueillent des données relevant d'un domaine de recherche particulier, d'autres se concentrent sur des types de ressources et d'autres encore sur des régions géographiques. Certains acceptent les dépôts de tout un chacun, tandis que d'autres ne prennent en charge que les données des membres de leur consortium. Les bibliothèques universitaires se sont mises à acquérir les données de projets de recherche basés dans leur institution. Cependant, la plupart d'entre elles ne peuvent accepter que de petits jeux de données qui ne seraient pas hébergés autrement ou se contentent d'administrer un service d'enregistrement qui améliore la découvrabilité.

L'engagement en matière de conservation est également très variable d'un référentiel à l'autre. Sauvegarder des données dans leur format de dépôt peut être coûteux ; mais les maintenir sur de longues périodes nécessite des investissements autrement plus importants dans la migration vers de nouveaux formats et technologies à mesure qu'ils apparaissent. Chaque référentiel spécifie les normes de qualité et de format qu'il accepte. Leur niveau d'investissement dans la vérification de la teneur et de la structure des données, dans l'augmentation des jeux par des métadonnées et de la documentation de provenance ou dans l'apport d'autres services à valeur ajoutée varie considérablement. Certains acceptent les données dans la mesure où elles répondent aux normes techniques, laissant le soin de la validation scientifique au contributeur ou à la contributrice. D'autres contrôlent la

qualité des données avant d'accepter le dépôt. Une fois les données versées, la responsabilité de leur maintenance et de leur accessibilité passe généralement des scientifiques aux référentiels. Ces derniers peuvent s'engager à conserver les données indéfiniment ou pour des périodes déterminées. Ils peuvent se soumettre à une homologation vis-à-vis de normes de qualité et de dispositifs de pérennisation, mais cette procédure est généralement facultative (Consultative Committee for Space Data Systems, 2012 ; Data Seal of Approval, 2014 ; Jantz et Giarlo, 2005 ; Digital Curation Centre, 2014).

Pour autant que les chercheurs et chercheuses soient prêts à diffuser leurs données, ils sont vraisemblablement plus enclins à les verser dans des référentiels de confiance qui conserveront leurs données et les mettront à la disposition du public indéfiniment. Les référentiels peuvent apporter une valeur ajoutée aux données grâce à des métadonnées, des informations de provenance, de la classification, des normes en matière de structure de données et de la migration. Ils peuvent également créer de la valeur en améliorant la découvrabilité des données et en élargissant leur utilisation à l'aide d'outils et de services. Il s'agit là d'investissements substantiels à réaliser pour les communautés et les universités, ainsi que pour les organismes et autorités qui les financent.

Les référentiels représentent une composante essentielle de l'infrastructure de la connaissance et un facteur primordial dans la diffusion et la réutilisation de données. Ils ne constituent toutefois pas une panacée. Les choix de conception varient selon les communautés et les circonstances. Il est plus facile d'organiser des ensembles en fonction de la structure de leur contenu qu'en fonction des questions posées par les utilisateurs et utilisatrices. Les systèmes individuels s'efforcent de garder une cohérence interne en utilisant des critères qui conviennent au contexte et aux usages perçus de la communauté. Parce que différents référentiels répondent à différents besoins, il existe peu de moyens efficaces de chercher dans plusieurs référentiels à la fois. Chaque personne et chaque robot peuvent poser une question différente sur ces données. Les requêtes évoluent au fil du temps. Les robots aussi apprennent et s'adaptent. Recycler des données pour répondre à des questions inattendues est un but plus ambitieux encore que simplement les réutiliser aux fins pour lesquelles elles ont été créées. Des mécanismes de fédération sont nécessaires et doivent, eux aussi, être régis par une gouvernance.

Certains chercheurs et chercheuses s'opposent au développement d'ensembles parce qu'ils draineraient des ressources au détriment d'autres formes de recherche, de bourses modestes et de la création d'autres formes de données. Les vastes projets à forte concentration de données peuvent être jugés contre-productifs s'ils semblent

dissuader la poursuite de nouvelles pistes de recherche axées sur la découverte. D'autres soutiennent que les données sont agrégées à une mauvaise échelle – soit trop grande, soit trop petite – pour être réutilisées (J. Boyle, 2013 ; Eisen, 2012 ; Kwa et Rector, 2010).

La granularité et la représentation sont des défis fondamentaux dans la conception de tout système d'information. L'ensemble des structures, des outils et des classifications qui accélèrent l'exploitation des données dans une communauté créeront des barrières dans une autre. Il n'y a pas de modèle unique qui conviendrait à tous, surtout pour des informations aussi malléables que les données de recherche. Lors de la conception d'infrastructures de la connaissance en général et de référentiels de données en particulier, il est essentiel de prendre en compte différents niveaux de granularité et de multiples outils d'exploitation du contenu, ainsi que de veiller à la flexibilité des moyens d'accès.

Pratiques privées

Les référentiels ne représentent pas les seuls moyens de diffuser des données pour qu'elles soient réutilisées par d'autres. Les données circulent également au sein d'informations complémentaires, en étant publiées sur un site web ou encore à la demande. Les études sur les pratiques des revues illustrent à quel point il est difficile de déterminer avec certitude la quantité de données diffusées en complément des articles (Alsheikh-Ali *et al.*, 2011). Même lorsque les revues stipulent que les données associées aux articles doivent être mises à disposition, les comités de rédaction ne tiennent que rarement des statistiques détaillées sur le respect de cette exigence. Les types, les formats et les volumes de données et de documentation varient grandement. Les auteurs et autrices peuvent fournir un identifiant de registre de données pour attester le dépôt dans un ensemble communautaire, mais parfois une déclaration comme quoi les données sont disponibles sur demande peut suffire. L'une ou l'autre de ces méthodes de diffusion peut répondre aux exigences ou aux recommandations des organismes de financement.

Comme noté plus haut, le partage privé pourrait être plus courant que l'acquisition de données à partir de référentiels (Faniel et Jacobsen, 2010 ; Pienta *et al.*, 2010 ; Wallis *et al.*, 2013). Ces formes de partage sont difficiles à repérer et à documenter, donc peu d'études et de statistiques existent. Les communications privées peuvent être très efficaces, car elles permettent aux scientifiques de discuter de la teneur, du contexte, des forces, des limites et de l'applicabilité d'un jeu de données spécifique à un phénomène.

Le partage de données entre individus, bien qu'il constitue un mécanisme courant et efficace pour la réutilisation, n'est pas applicable à grande échelle. Les chercheurs et chercheuses consacrent du temps et des efforts à rendre leurs données utiles à d'autres uniquement dans certaines circonstances. Cela peut dépendre de leur relation avec le demandeur ou la demandeuse et de la confiance qu'ils lui accordent, de ce qu'on leur demande et de l'utilisation prévue des données. Si les requêtes sont fréquentes ou pénibles, ils peuvent préférer refuser. Publier des jeux de données ou des informations à leur sujet sur des sites web est une méthode de partage qui épargne aux auteurs une partie du travail de documentation ; elle revient cependant à renoncer à maîtriser qui réutilise les données et comment.

Bien qu'imparfaites, toutes ces méthodes améliorent les possibilités de réutilisation des données et doivent être prises en charge dans les infrastructures de la connaissance.

Le partage privé et la publication sur un site web peuvent répondre aux exigences en matière de diffusion des données, mais elles ne favorisent pas particulièrement la découvrabilité, la provenance, l'utilisation et la pérennité. Le partage privé ne fonctionne que tant que les chercheurs et chercheuses disposent toujours de leurs données et sont familiers de leur contexte. De même, les jeux de données publiés sur des sites web consistent généralement en des fichiers statiques associés à des versions particulières d'un logiciel. L'accès à ces données diminue à mesure que les logiciels évoluent, que les ordinateurs sont remplacés, que les mémoires sont nettoyées et que le souvenir des détails s'estompe.

Infrastructure humaine

La capacité à divulguer, partager et réutiliser des données repose largement sur l'infrastructure humaine. Les données peuvent être diffusées sous forme de jeux non traités – des chaînes de symboles accompagnées d'une documentation minimale –, mais ne présentent alors que peu de valeur pour la réutilisation, faute d'une représentation et d'une documentation adéquate. Peu de chercheurs et chercheuses disposent du savoir-faire pour documenter leurs données en se conformant à des normes archivistiques. Des notions telles que les métadonnées et la provenance n'occupent pas une place clé dans les cours de second cycle, à part peut-être dans les métiers de l'information. Le temps nécessaire pour acquérir ces compétences est souvent vu comme du temps non consacré à la recherche. La documentation des données, quand elle est faite, est souvent déléguée à des étudiantes et étudiants de second cycle et à d'autres personnels de recherche. Ces personnes sont souvent celles qui côtoient les données de plus près : elles effectuent en effet l'essentiel de

la collecte et de l'analyse, en particulier dans les sciences exactes. Les chercheurs peuvent accepter de déléguer certaines de ces activités à des bibliothécaires et à des archivistes en fonction du temps qu'ils devront passer à travailler avec eux, du bénéfice perçu et des bénéficiaires supposés.

Les investissements visant à rendre les données diffusables et réutilisables constituent des problématiques infrastructurelles. Ils devraient être envisagés comme des questions communautaires et non comme la seule responsabilité de chercheurs et chercheuses individuels. Les institutions qui investissent dans les ressources humaines nécessaires à la gestion des données sont mieux à même de réutiliser ces dernières. Les personnes qui organisent les données et maintiennent les normes, les systèmes de classification et les outils associés sont des médiatrices entre celles qui créent les données et celles qui pourraient les réutiliser. Une grande partie de ce travail est effectué – quand il l'est – par des professionnels de l'information travaillant pour des référentiels de données et d'autres institutions (Lee *et al.*, 2006 ; Mayernik, 2015).

Ce travail est bien trop souvent invisible, comme expliqué au chapitre 3. Celles et ceux qui bénéficient du labeur des scientifiques et des professionnels de l'information, qui représentent les données pour les rendre mobiles, peuvent tout ignorer de ces investissements. L'invisibilité du travail nécessaire pour que toutes les composantes interagissent en harmonie est une caractéristique fondamentale de l'infrastructure. Elle met aussi en exergue les tensions entre parties prenantes. Celles et ceux qui investissent dans les référentiels, dans le travail interne de gestion des données et dans les autres parties de l'infrastructure n'en sont pas toujours les bénéficiaires directs. En effet, les acteurs d'autres communautés, d'autres pays et d'autres secteurs peuvent en tirer des bénéfices substantiels. C'est là que résident les défis pour la gestion de ces réservoirs communs de ressources, parmi lesquels la pérennité et le libre accès des passagers clandestins.

Problèmes insolubles

La gouvernance de réservoirs communs de ressources, tels les référentiels de données, ne représente pas une tâche impossible, mais les méthodes pour y parvenir sont loin d'être automatiques (Ostrom et Hess, 2007). Elle doit être continuellement renégociée. Certains problèmes paraissent insolubles, notamment ceux causés par les influences extérieures identifiées dans les études de cas. Ils ne peuvent être réglés, mais des compromis sont parfois possibles. Par exemple, la propriété, la licence et les droits sur les données varient selon les domaines, les pratiques, les

politiques et les juridictions. Des accords conclus en début de collaboration peuvent réduire les frictions, mais sans nécessairement les éliminer.

Certaines formes de données ne peuvent pas être partagées pour des raisons de confidentialité ou d'éthique. La majeure partie de la recherche en sciences sociales sur des sujets humains est menée dans le cadre de réglementations qui limitent la diffusion des données. Lorsque ces informations sont divulguées, elles sont anonymisées dans la mesure du possible. Si les données sont utilisées isolément, des techniques de base comme le remplacement des noms par des identifiants peuvent suffire. D'autres méthodes, comme celles employées dans les recensements, agrègent des données en des ensembles suffisamment importants pour qu'on ne puisse pas identifier les individus. Néanmoins, des personnes qui sont anonymes dans un jeu de données singulier peuvent souvent être identifiées si on en combine plusieurs. Le regroupement d'informations telles que l'âge, le sexe, le lieu de résidence, la date de sortie de l'hôpital et le nombre d'enfants rendent la réidentification étonnamment facile. Des donneurs d'ADN anonymes ont ainsi pu être réidentifiés à l'aide de bases de données généalogiques publiques (Bohannon, 2013b ; Gymrek *et al.*, 2013 ; Rodriguez *et al.*, 2013). Anonymiser les données de manière sûre tout en préservant leur intérêt pour la recherche représente un défi technique et stratégique (Ohm, 2010 ; Sweeney, 2002). Une approche consiste à ne diffuser les données de sujets humains qu'auprès de chercheurs et chercheuses qualifiés et à condition qu'aucune tentative de réidentification ne soit entreprise. Ce type de modèle suppose une supervision institutionnelle.

Les infrastructures de la connaissance par discipline

Les études de cas ont fourni une abondance de détails sur le travail scientifique des données dans de nombreux domaines de recherche. Un bref résumé des méthodes de gestion des données utilisées par chacun de ces champs, de leurs motivations et des acteurs impliqués illustrera les différences entre leurs infrastructures de la connaissance.

Dans les sciences exactes

Les sciences exactes sont loin d'être homogènes en ce qui concerne leur gestion des données, leurs pratiques de partage et leurs investissements dans les infrastructures de la connaissance.

En astronomie

En astronomie, la fourniture de données est largement descendante : des missions nationales et internationales développent et mettent en service des instruments de télescope dans l'espace et au sol. Les missions spatiales sont celles qui investissent le plus dans les ensembles de données. Les données des instruments sont transmises en continu à des stations au sol et dans des référentiels, où elles sont alors traitées dans un *pipeline*. Une fois nettoyées, étalonnées et vérifiées, elles sont diffusées à la communauté scientifique.

Chaque *data release* (divulgaration de données) de mission astronomique telle que le Sloan Digital Sky Survey, Hubble, Chandra et Planck s'accompagne généralement d'un *data paper* (publication de données) qui consigne le travail scientifique et la provenance du jeu de données. Les *data papers*, qui comportent les noms de tous les membres de l'équipe impliqués, sont proposés à des revues d'astronomie. Il s'agit des publications les plus citées dans la discipline et c'est par ce moyen que les chercheurs et chercheuses montrent qu'ils partagent leurs données.

L'astronomie dispose de l'infrastructure de la connaissance la plus complète de tous les domaines évoqués dans les études de cas. Ses réservoirs communs regroupent des télescopes, des référentiels de données, des systèmes de classification et de nombreuses autres ressources. La communauté a convenu de standards pour les structures de données, les métadonnées et les ontologies. Elle dispose de suites logicielles pour analyser les données dans ces formes normalisées. Depuis le début des années 1990, elle coordonne des efforts internationaux en matière de publication pour verser ses notices bibliographiques à l'Astrophysics Data System. Ce dernier recouvre la littérature astronomique jusqu'au ^{xix}^e siècle. Les objets célestes mentionnés dans ces publications sont catalogués dans SIMBAD, NED et d'autres systèmes de ce type. Ces derniers se recoupent étroitement, de manière que les publications pointent vers les astres et que les documents sur les objets célestes soient liés aux publications où ils sont mentionnés. Le WorldWide Telescope combine des données et de la littérature tirées de ces sources et d'autres pour créer une plateforme de visualisation de données. Spécialistes et novices peuvent contribuer à des « visites guidées » d'astres, de phénomènes et de régions célestes. Le WWT est intégré dans des outils astronomiques pour servir de plateforme scientifique et éducative (Accomazzi et Dave, 2011 ; Accomazzi, 2010 ; Eichhorn, 1994 ; Genova, 2013 ; Goodman, Fay *et al.*, 2012 ; Hanisch, 2013 ; Norris *et al.*, 2006 ; Pepe *et al.*, 2011 ; Udomprasert et Goodman, 2012 ; White *et al.*, 2009 ; WorldWide Telescope, 2012).

Malgré sa vaste infrastructure de la connaissance, la communauté astronomique n'attend pas de ses chercheurs et chercheuses qu'ils diffusent leurs propres

données, que celles-ci soient dérivées de données de référentiels ou aient été obtenues par d'autres moyens, comme des demandes de temps d'observation. Si quelques référentiels existent pour accepter des données, la plupart sont des archives de soutien aux missions. Certains chercheurs publient des données sur les sites web de leurs laboratoires, comme l'a fait l'équipe du relevé COMPLETE dans l'étude de cas. Les référentiels universitaires comme Dataverse bâtissent des capacités d'archivages spécifiques à l'astronomie (Goodman *et al.*, 2012). Un nombre relativement faible de publications en astronomie comportent des liens vers les sources de données, mais le taux de citation augmente (Pepe *et al.*, à paraître).

L'observatoire de rayons X Chandra offre un exemple de gain de valeur des données au fil du temps grâce à des investissements intensifs dans leur intendance. Le personnel de Chandra a documenté ses usages dès le lancement de la mission en 1999, élaborant des indicateurs et des moyens de valoriser les données. Le personnel chargé des archives de données fouille la littérature astronomique à la recherche de mentions de la mission, puis lit les publications avec attention pour déterminer si ses données ont été utilisées et, si oui, comment. Leurs indicateurs imposent que l'article comporte un lien explicite vers une observation et qu'une propriété quelconque en soit dérivée. Le personnel distingue les publications scientifiques sur Chandra, les publications liées à Chandra et les articles sur la conception d'un instrument. Les utilisations de données qui répondent aux critères sont ajoutées aux archives de Chandra, à l'ADS et à SIMBAD, créant des liens entre les jeux de données, les publications et les objets célestes à l'aide de métadonnées. Cette procédure ajoute de la valeur aux données par l'apport de documentation scientifique et les rend plus facilement découvrables grâce à de nouveaux points d'entrée. Le personnel de Chandra se charge de créer les liens parce que les auteurs et autrices citent rarement expressément les sources de données (Becker *et al.*, 2009 ; Blecksmith *et al.*, 2003 ; Lagerstrom *et al.*, 2012 ; Rots et Winkelman, 2013 ; White *et al.*, 2009 ; Winkelman et Rots, 2012a, 2012b ; Winkelman *et al.*, 2009).

Ces efforts investis dans l'intendance des données montrent aussi la difficulté de définir ce qu'est une « utilisation ». Comme l'ont découvert Rots, Winkelman et Becker (2012), la notion d'« observation » diffère d'un observatoire à l'autre et peut varier au sein d'une même mission astronomique. À cause de cela et d'autres ambiguïtés dans les indicateurs des référentiels, un groupe de travail international a établi une série de bonnes pratiques en matière de statistiques qui sont applicables aux archives de données astronomiques (IAU Working Group Libraries, 2013).

En science et technologie des réseaux de capteurs

Les tentatives de normalisation de la collecte de données et d'adoption de standards de métadonnées au sein du CENS ne cadraient pas avec la nature exploratoire et génératrice d'hypothèses de cette communauté de recherche. Lorsqu'une réutilisation avait lieu, elle impliquait souvent un partage privé pour s'assurer que les informations contextuelles étaient transmises avec les jeux de données. Les données les mieux à même d'être versées à des référentiels provenaient de pratiques et de protocoles normalisés, comme les données génomiques et sismiques. Certains logiciels étaient déposés dans des référentiels de code. Les équipes étaient souvent pleines de bonnes intentions en matière de diffusion, mais avaient omis de désigner un membre qui en aurait la responsabilité (Borgman *et al.*, 2012 ; Wallis et Borgman, 2011 ; Wallis, 2012).

Les chercheurs et chercheuses du CENS n'avaient pas besoin d'harmoniser des données sur le long terme, sauf au sein des équipes individuelles. La plupart des chercheurs se préoccupaient de la cohérence de leurs observations plutôt que de leur comparabilité entre sites de recherche. La comparaison à l'échelle mondiale est nécessaire pour mesurer le changement climatique et la perte de biodiversité, mais c'est là un autre type de programme de recherche. Au CENS, les réserves naturelles participantes, où l'on recueillait des observations synoptiques de la flore, de la faune et du climat, constituaient une exception (James San Jacinto Mountains Reserve, 2013 : Natural Reserve System, 2013). Elles ont servi de données d'arrière-plan pour certains types de recherche au CENS.

Il existe peu de référentiels pour prendre en charge des données exploratoires telles que celles générées par la recherche de terrain du CENS. D'ailleurs, la demande en la matière y était faible, car les scientifiques n'arrivaient pas à imaginer qui pourrait bien avoir l'usage de leurs données (Mayernik, 2011). Le CENS a conçu un petit système d'archivage de données pour encourager le partage et la réutilisation, mais les dépôts y ont été peu nombreux (Wallis *et al.*, 2010). Les communautés des sciences de l'environnement ont commencé à mettre en place des référentiels et des exigences de partage dans les derniers jours du CENS seulement (Dryad, 2013 ; Reichman *et al.*, 2011 ; Whitlock, 2011).

Une des nombreuses situations rencontrées à la fois par le CENS et par les Long Term Ecological Research Centers (LTER) a été la tentative d'adoption du standard Ecological Metadata Language (EML) afin d'améliorer le partage, la réutilisation et la comparaison des données. Millerand et Bowker (2009) ont identifié des récits concurrents sur l'acceptation de l'EML sur les sites LTER. Selon les concepteurs et conceptrices du standard, celui-ci a été adopté par l'ensemble de la communauté

des LTER et l'entreprise a donc été couronnée de succès. Les gestionnaires de l'information chargés de la mise en place de l'EML sur leurs sites de recherche ont livré une version différente : ce standard est complexe, difficile à comprendre dans sa globalité et les outils fournis étaient incompatibles avec les pratiques locales. De leur point de vue, l'EML n'a que partiellement atteint son but. L'expérience du CENS est comparable à ce dernier récit. Comme l'EML se prêtait le mieux aux activités de recherche de plusieurs équipes du CENS, il a été proposé comme standard pour un référentiel de données commun. Cependant, le manuel de deux cents pages a découragé ces petites équipes ; elles ont rejeté l'EML au motif que cette solution était trop « puissante » pour leurs besoins. La plupart ont continué d'organiser leurs données dans des dossiers hiérarchiques et à l'aide d'autres conventions internes pour nommer les éléments et les variables (Wallis *et al.*, 2010).

Les collaborations au sein du CENS se sont formées autour d'intérêts communs pour des problématiques de recherche plutôt que sur des données partagées. Les scientifiques avaient besoin de meilleurs outils pour mener leurs travaux, tandis que les spécialistes de l'informatique et de l'ingénierie avaient besoin d'applications concrètes leur permettant de concevoir et de tester des technologies. Chaque équipe serait peut-être capable de retracer la provenance de ses propres collectes de données, mais les déploiements conjoints ne peuvent être complètement reconstitués. Ce n'est pas un échec : cette issue montre plutôt les efforts considérables que les équipes ont consacré à l'apprentissage des domaines des autres pour pouvoir traiter des problématiques ensemble. Malgré leurs problèmes d'interopérabilité, elles ont tout de même pu collaborer (Edwards *et al.*, 2011 ; Mayernik *et al.*, 2011 ; Olson *et al.*, 2008 ; Wallis *et al.*, 2008 ; Wallis *et al.*, 2007).

En génomique

Bien que nous n'ayons pas développé une étude de cas complète à son sujet, la génomique offre des exemples d'infrastructures de la connaissance qui contrastent avec les autres cas en sciences exactes. Parce que les spécialistes de ce domaine ont besoin de réutiliser leurs propres données, l'offre en la matière se crée de manière ascendante à mesure que les laboratoires séquent des génomes et contribuent aux référentiels. Ces scientifiques ont pu investir des ressources massives dans des laboratoires, du matériel, des logiciels et du personnel. Elles et ils ont une motivation pour normaliser leurs procédures, ce qui produit des formes standardisées de données plus faciles à partager. Ces domaines de recherche sont extrêmement concurrentiels. Les laboratoires s'efforcent d'être les premiers à trouver un remède contre le cancer, la maladie d'Alzheimer, la grippe pandémique et d'autres affections redoutables : pour eux, des vies et des sommes importantes sont en jeu. La menace de brevetage et de contrôle qui a pesé sur le génome humain, combinée au déluge de

données génomiques, a mené à des accords historiques en matière de partage des données (Watson, 1990). Depuis, le taux de réussite des essais cliniques de médicaments a diminué et leur coût a augmenté. La pression du public pour la divulgation des résultats d'essais cliniques s'est également intensifiée. Le partage des données génomiques promet l'accélération de la découverte de médicaments, la réduction des essais redondants, un recours plus faible aux sujets de recherche et l'apport de multiples bénéfices aux malades et aux médecins.

En dépit de l'intense concurrence entre parties prenantes, la génomique développe des réservoirs communs de ressources. De nombreux laboratoires pharmaceutiques et consortiums universitaires ont commencé à partager les données de leurs essais cliniques, des séquences du génome et des structures de protéines. Cependant, certains acteurs refusent d'en faire de même et chacun peut imposer ses propres conditions au partage et à la réutilisation (Den Besten *et al.*, 2009 ; Bhattacharjee, 2012 ; Check Hayden, 2013 ; Corbyn, 2011 ; Couzin-Frankel, 2013b ; Edwards *et al.*, 2009 ; Edwards, 2008b ; Goldacre, 2012 ; Howe *et al.*, 2008 ; Nelson, 2009 ; Rabesandratana, 2013 ; Weigelt, 2009 ; Williams *et al.*, 2012).

Même en ayant divulgué ses structures de protéines, le laboratoire qui conduit les études a toujours l'avantage quand il s'agit d'utiliser ces données pour expliquer des mécanismes scientifiques. Il a en effet passé des semaines ou des mois à générer, nettoyer et réduire ses données. Seules les données finales, propres et étalonnées, sont communiquées et diffusées. Bien que des traces des protocoles et des traitements intermédiaires soient conservées, tant de données sont créées et réduites qu'il est peu probable de pouvoir répliquer exactement les mêmes étapes. Les conditions précises de chaque expérience ne peuvent être reproduites, car des différences infimes de température, d'humidité et d'autres conditions du laboratoire affectent les observations. Le laboratoire diffuse ses protocoles en même temps que ses données pour aider d'autres à les réutiliser. Cependant, on s'aperçoit que lorsque des laboratoires extérieurs essaient de répliquer ses études, de petites déviations par rapport au protocole peuvent produire des résultats très différents. Même lorsqu'on applique des normes scientifiques exigeantes, des procédés par étapes tels que la cristallisation de protéines peuvent s'avérer stochastiques.

Si la mise en commun de ces données représente une fusion d'intérêts remarquable, elle est aussi semée d'embûches, certaines prévues et d'autres non. Le Structural Genomics Consortium (2013) et Sage Bionetworks (2013) sont deux vastes collaborations à l'avant-garde du partage et de la réutilisation de données en génomique et en sciences biomédicales. Elles sont financées par différentes sources, dont des subventions et des sociétés pharmaceutiques. Pour recruter et retenir les meilleurs

chercheurs et chercheuses, ceux-ci doivent pouvoir publier leurs conclusions dans des revues de premier rang comme *Science*, *Nature* et *Cell*. C'est là que la porosité des données et des résultats apparaît. Si les données sont publiées avant la soumission d'un article les concernant, les revues considèrent parfois cela comme une « publication préalable » et peuvent refuser d'examiner le manuscrit. Pourtant, les résultats pourraient être novateurs parce qu'ils expliquent le fonctionnement biologique de la protéine et du modèle. La diffusion des données est donc un acte scientifique différent de la communication des résultats. Une société pharmaceutique peut rechercher un nouveau médicament en se fondant sur des corrélations entre l'expression du gène et la protéine en même temps qu'un laboratoire académique étudie la biologie qui explique cette corrélation.

La diffusion d'informations génétiques humaine est imprégnée de questions bio-éthiques complexes. Par exemple, les propositions de diffuser les « découvertes fortuites » révèlent d'importantes divergences d'opinions des spécialistes de la bio-éthique, des spécialistes de la recherche biomédicale, des responsables politiques, des médecins et des patientes et patients. Certains pensent que des résultats génétiques imprévus devraient être communiqués au patient, tandis que d'autres estiment que ce sont les patients ou les médecins qui doivent décider si les découvertes sans lien avec le motif de consultation devraient être intégrées au dossier (Cousin-Frankel, 2013a ; McGuire *et al.*, 2013 ; Wolf *et al.*, 2013).

Dans les sciences sociales

Comme le montrent les études de cas, la recherche en sciences sociales est elle aussi caractérisée par un ensemble composite de pratiques de partage des données. Celles-ci influent sur la capacité à diffuser, interpréter et réutiliser des données.

Dans la recherche sur Internet

En tant que parties prenantes du World Internet Project, les chercheurs et chercheuses menant l'Oxford Internet Survey ont besoin de comparer leurs données dans le temps et l'espace et sont donc motivés à maintenir une cohérence interne qui le permette. Chaque jeu de données est placé sous embargo le temps d'un cycle d'enquête pour être exploité en vue d'une publication dans des revues. Les jeux diffusés à des chercheurs qualifiés sont les observations « brutes » fournies par l'institut de sondages dans un format conçu pour STATA, un progiciel statistique courant.

En diffusant des jeux de données sur son site web avec une simple licence autorisant les usages non commerciaux, l'Oxford Internet Institute conserve une trace des demandes et des téléchargements. Il n'a cependant aucun moyen de suivre l'usage

réel. Faute de recevoir des demandes sur les caractéristiques précises des données de la part des internautes, il suppose qu'elles remplissent un rôle d'arrière-plan à des fins scientifiques ou éducatives, comme l'enseignement de techniques de sondage. Puisqu'il ne dépose pas les jeux de données dans des archives, il n'a pas besoin de leur assigner des métadonnées ou de la documentation de provenance au-delà de ce qui est nécessaire pour les besoins en interne.

Des flux Twitter et d'autres données de réseaux sociaux peuvent être obtenus auprès des sociétés qui possèdent les services concernés. Ils peuvent donner lieu à un contrat qui régit le droit des chercheurs et chercheuses à les diffuser. Si les données sont recueillies en copiant des sites web, elles sont alors plus difficiles à contrôler, même si les fournisseurs s'efforcent d'en limiter la diffusion. La possibilité de maîtriser et de diffuser des informations relatives à sa propre personne ou à d'autres relève d'une zone grise des politiques en matière d'Internet (J. E. Cohen, 2012). Les revues craignent de publier des articles fondés sur des données exclusives interdites à la diffusion. Les spécialistes de la recherche d'information s'inquiètent également de mener des recherches sur des jeux de données qui ne peuvent être diffusés à des fins de réplication.

En sociotechnique

L'équipe de recherche sociotechnique dont le travail est étudié au chapitre 6 adapte ses questionnaires, ses protocoles, ses guides de codification et ses autres méthodes de manière à ce que ses résultats soient comparables au fil du temps entre différents sites de recherche. La réutilisation des instruments, des guides de codification, des entretiens et des notes de terrain nécessite une autorisation de comités de protection des personnes. Les sujets de recherche peuvent approuver ou exclure les clauses de réutilisation des données. Dans le cas de projets collaboratifs impliquant plusieurs universités, les protocoles concernant les sujets humains doivent être approuvés par chaque établissement participant et donnent lieu à de multiples accords bilatéraux.

Les entretiens ouverts et les ethnographies peuvent être plus difficiles à anonymiser que les sondages ou les dossiers médicaux, en particulier quand les personnes sont connues au sein de leurs communautés. Les enregistrements audio des scientifiques et des technologues de l'étude de cas ne peuvent pas être anonymisés, puisque les voix y sont distinctes et reconnaissables. Les transcriptions sont à peine moins identifiables, car les détails fournis sur les centres d'intérêt, les métaphores utilisées et les formes d'expression peuvent révéler une identité. Le caviardage de ces matériaux pour les anonymiser de manière à ce qu'ils conservent suffisamment d'informations pour être réutilisables par d'autres pourrait constituer un projet de recherche à part entière. Communiquer les résultats tirés de ce type de données suppose un savant

savoir-faire : les explications sont travaillées de manière à être les plus descriptives possible en conservant les nuances d'interprétation de la communication, mais sans donner des détails qui dévoileraient l'identité des personnes impliquées. Utiliser ces données est difficile et les réutiliser plus encore.

Dans les sciences humaines

Les données des sciences humaines sont les plus indéfinissables de toutes ; pourtant, leur réutilisation augmente. Les scientifiques trouvent de nouvelles manières d'exploiter leurs données et de poser des questions inédites.

En histoire de l'art et en archéologie antiques

Le projet CLAROS est autant une infrastructure que la constitution d'un ensemble. Il réemploie des archives muséales pour qu'elles puissent être utilisées par des scientifiques, des apprenantes et apprenants et le grand public. Il leur adjoint de nouvelles fonctionnalités, comme la recherche d'images ou l'interrogation de plusieurs musées. Ces ensembles sont utiles à plusieurs types de publics. Le défi en matière de gouvernance est de répondre à tous les besoins tout en aidant suffisamment les communautés pour que celles-ci s'investissent dans la survie du projet. Ces systèmes complexes peuvent être pérennisés de plusieurs manières. Une approche consiste à en maintenir l'interactivité ; une autre à conserver une archive fermée du contenu. Les efforts investis dans l'archivage des « données internes » indépendamment du système opérationnel intégral ont constitué un facteur dans la disparition de l'Arts and Humanities Data System (Reimer *et al.*, 2007 ; Robey, 2007, 2011).

Pour la recherche sur le griffon de Pise, les spécialistes de l'histoire de l'art et de l'archéométallurgie rassemblent des données tirées d'archives et de sources secondaires du monde entier. Elles et ils examinent aussi l'objet lui-même. L'archéométallurgiste fournit des données tabulaires dans un format qui permet de comparer les métaux de cette œuvre en bronze à des objets similaires. Les publications sur le griffon de Pise comportent des notes de bas de page et des renvois à des publications et d'autres documents où des informations ont été trouvées. Des personnes ayant fourni des informations précises aux chercheurs sont parfois nommées. En revanche, on mentionne rarement les archives consultées ou les ensembles où des éléments probants ont été découverts : dans ces domaines, ces derniers peuvent être considérés comme des données d'arrière-plan. Les archives et les collections dont des ressources ont été tirées peuvent ne jamais savoir comment leurs ensembles sont employés ou comment les éléments obtenus sont réutilisés. Pour la communauté de recherche, publier les interprétations est plus important que de diffuser les données. Ce n'est que dans les

rares cas où les travaux sont contestés sur le plan de l'exactitude ou de la fraude que l'on communique des notes et d'autres traces (Wiener, 2002).

En études bouddhiques

Le spécialiste en philologie bouddhique chinoise a accumulé un vaste ensemble de ressources au cours de sa carrière, qu'il continue d'explorer. Jusqu'à récemment, il s'agissait surtout d'une collection sur papier. Maintenant qu'il est possible de saisir, diffuser et réutiliser des représentations de textes très anciens en langues non occidentales, il accumule davantage de traces numériques de ses observations. Ces possibilités étaient inimaginables il y a encore quelques années pour cette petite communauté soudée. Les motivations du philologue sont de maintenir ces ressources de façon à pouvoir les fouiller et les combiner à l'avenir. De même, son accès à la CBETA lui offre une immense mine de ressources numériques où puiser. Il rapporte les détails de ses analyses sous forme de longs tableaux dans ses publications.

La CBETA et les autres recueils entretenus par la communauté bouddhiste représentent un cas particulier de réservoirs communs de ressources. Cette religion encourage la diffusion de ses textes depuis des siècles, car cela obéit à un précepte. La pérennité de ces ressources numériques n'inquiète donc pas les savantes et savants. Le recueil, qui continue de croître au-delà des cent trente-cinq et quelques volumes d'origine, est d'une portée et d'une qualité incomparables pour les spécialistes de ce champ.

Conclusion

Les politiques des organismes de financement, des revues et des autres acteurs du travail scientifique des données se sont employées à augmenter l'offre de données réutilisables, sans guère s'intéresser aux motivations scientifiques du partage ou de la réutilisation des données ou aux investissements nécessaires dans les infrastructures de la connaissance. Les raisonnements derrière la promotion du partage reflètent les préoccupations légitimes des parties prenantes, comme la capacité à reproduire des travaux, à mettre des ressources publiques à la disposition de la population, optimiser les investissements dans la recherche et faire progresser la recherche et l'innovation. Ils conduisent souvent à l'instauration de politiques génériques qui ne reflètent pas la grande diversité du travail scientifique des données d'un domaine à l'autre et en leur sein. Les données sont à la fois des atouts et des charges. Lorsque la recherche repose sur la capacité à réutiliser ses propres données et à les mettre en commun avec d'autres scientifiques de la discipline, des motivations extrinsèques incitent au partage

et à la réutilisation. En revanche, lorsque la recherche est locale et exploratoire ou qu'elle dépend d'éléments probants accumulés sur de longues périodes, les incitations à diffuser sont moindres.

L'amalgame entre accès ouvert aux données et accès ouvert aux publications traduit une incompréhension de ce qu'est la communication savante. Les publications sont bien plus que de simples contenants pour les données : elles sont des argumentaires étayés par des preuves et les relations qui les unissent aux données sont très variées. Dans les domaines où les jeux sont distincts et où chacun correspond à une publication, la diffusion des données est envisageable. En revanche, dans les disciplines qui consistent largement en des interprétations, elle n'a pas de sens. Diffuser les données représente rarement une simple tâche annexe à la publication d'un article de revue. En effet, représenter les données d'une manière qui serve aux autres peut mobiliser des ressources substantielles, qui ne sont pas consacrées directement à la recherche.

Il faut envisager la diffusion, le partage et la réutilisation des données comme des problèmes d'infrastructures de la connaissance. Se concentrer sur l'offre de données pose la question de la demande. Or, dans la plupart des domaines, la demande de données réutilisables paraît très faible. Il est cependant très difficile de définir les utilisations et les réutilisations, car les pratiques varient considérablement. Rendre les données mobiles pour qu'elles circulent entre domaines, entre contextes et au fil du temps est extrêmement difficile. Elles ne sont pas des objets naturels faciles à transformer en marchandises et à échanger sur un marché ; ce sont des entités qui servent à mettre en évidence des phénomènes. Un même objet ou observation peut être représenté de multiples façons. Par conséquent, la même entité devient des données différentes quand elle est transférée dans un autre contexte. De légers changements dans l'interprétation, la méthode ou la pratique peuvent conduire à considérer ces entités comme des preuves légèrement différentes de phénomènes légèrement différents. Les choix effectués dans les premiers stades de la recherche, comme le prétraitement des observations avant qu'elles ne soient considérées comme des données, ont des répercussions tout au long du processus.

Les données sont souvent inséparables des logiciels, du matériel, de l'instrumentation, des protocoles et de la documentation. Pour gérer cette complexité, les scientifiques disposent de nombreux moyens. Cela implique souvent des décisions inscrites au cœur des activités de recherche et qui font partie d'un savoir implicite difficile à consigner ou à transmettre à autrui. Parfois, on part du principe que certains détails sont connus de la communauté et peuvent donc être omis de la documentation et des publications. Les jeux de données peuvent être

extraits de leur contexte technologique et réduits au plus petit outil commun pour être interprétés. Rendre les données mobiles suppose de prendre en compte ces subtilités et la perte de sens induite par la traduction.

Les infrastructures de la connaissance peuvent faciliter l'échange d'informations entre communautés et en leur sein. Pour les chercheurs et chercheuses, gérer ses propres données représente souvent la plus grande difficulté. Savoir les exploiter est pourtant un prérequis pour les partager. Au sein des infrastructures de la connaissance, les progrès en matière de méthodes, d'outils et de ressources humaines pour la gestion des données devraient permettre de tirer le meilleur parti de leur diffusion. Des moyens communs de représenter des objets de recherche, comme des méta-données et des mécanismes de classification, constitueraient une autre fonctionnalité utile pour de nombreuses communautés. Lorsque les communautés parviennent à se mettre d'accord sur des représentations et d'autres standards d'échange d'informations, elles disposent d'une base pour bâtir des ensembles partagés. Ces derniers ont plus de chances d'émerger dans des communautés qui ont besoin de combiner des ressources informationnelles. Ces ensembles peuvent répondre aux besoins de communautés petites ou grandes, pour de longues ou courtes périodes.

Néanmoins, les limites des communautés de recherche sont poreuses et en mutation perpétuelle. Une chercheuse peut être membre de plusieurs communautés et changer d'appartenance en fonction des priorités et des fronts de recherche. Les normes, les pratiques et les représentations qui facilitent le partage des données au sein d'une communauté peuvent aussi créer des barrières au partage entre disciplines. Quand ces barrières sont surmontées, c'est souvent par la communication privée entre chercheurs. Les discussions personnelles et le partage de connaissances au sein de groupes représentent des composantes essentielles de l'infrastructure de la connaissance. Cependant, ils ne sont pas généralisables à grande échelle. Chaque scientifique ne peut partager ses données en tête à tête qu'avec un nombre limité de personnes. De plus, le partage privé n'augmente pas au fil du temps : les données et leur savoir s'effritent rapidement si on n'en prend pas soin.

Il est crucial de reconnaître le niveau d'investissement à réaliser dans les infrastructures de la connaissance si l'on veut que le partage, la diffusion et la réutilisation des données de la recherche fonctionnent. La valeur d'un jeu de données ou d'un ensemble peut n'apparaître que bien après sa création. Pour s'en rendre compte, il faut investir massivement dans des ressources pécuniaires, techniques et humaines. Les infrastructures de la connaissance ont besoin de l'implication d'acteurs multiples et concurrents, souvent de part et d'autre des frontières nationales. Les coûts et les bénéfices des investissements qui y sont effectués sont inégalement répartis. Elles

ont toutes besoin d'être régies par une gouvernance, d'être maintenues le temps nécessaire et de gérer les passagers clandestins. Déclarer que les scientifiques doivent diffuser leurs données au nom du bien commun est, au mieux, simpliste : c'est une responsabilité communautaire. Il faut élargir le débat sur ce que le partage, la diffusion et la réutilisation des données sont censés accomplir. Si deux chercheurs veulent en apprendre plus sur leur travail scientifique des données respectif, alors le partage privé est le moyen le plus rapide. Si les ressources en données doivent être gérées pour profiter aux générations futures, des investissements majeurs dans la technologie, les ensembles, la main-d'œuvre et la gouvernance sont nécessaires. Les buts sont sans doute un mélange de ceux-là et de beaucoup d'autres. Les parties prenantes et les intérêts catégoriels abondent, mais il en va de même des formidables possibilités d'optimisation de l'usage des ressources existantes.

9. Reconnaissance, attribution et découvrabilité des données

Introduction

« Si les publications sont les étoiles et les planètes de l'univers scientifique, les données en sont la "matière noire" : influente, mais largement ignorée dans notre cartographie. »

— Groupe de travail sur les normes et pratiques en matière de citation de données du CODATA-ICSTI¹, « Out of Cite, Out of Mind »

Les données resteront la « matière noire » de la communication savante si elles ne sont pas décrites, conservées et rendues découvrables. La citation de données figure parmi les priorités des bibliothèques, des maisons d'édition, des sociétés savantes, des référentiels et des organismes de financement. Mentionner les créateurs et créatrices des données est devenu une urgence pour les organisations autour de la recherche d'information, comme CODATA, la Research Data Alliance, le Board on Research Data and Information (États-Unis), le Joint Information Systems Committee (Royaume-Uni) et DataCite. Des groupes de travail internationaux se sont formés. Les parties prenantes se sont rassemblées dans des colloques et des groupes d'étude dédiés à cette question. Des manifestes ont été publiés et des normes sont en cours d'élaboration (Altman et King, 2007 ; Crosas *et al.*, 2013 ; Institute for Quantitative Social Sciences, 2011 ; Research Data Alliance, 2013).

En surface, la citation de données s'apparente à un problème technique simple : adapter les mécanismes existants pour les références bibliographiques. Or, ce sont précisément les efforts pour y parvenir qui ont exhumé de vieux débats sur pour qui, pour quoi, comment, pourquoi et quand il fallait reconnaître une quelconque forme de contribution scientifique. Les pratiques savantes de gestion et de reconnaissance du mérite, d'attribution de la responsabilité et de découverte des publications se sont accumulées graduellement au cours des siècles. Bien qu'imparfaite, l'infrastructure de la connaissance qui en a résulté est suffisamment robuste pour remplir ces fonctions pour les publications anciennes comme nouvelles ; ainsi, les travaux de Galilée peuvent être trouvés et cités dans les réseaux numériques décentralisés d'aujourd'hui. À cette infrastructure s'adjoignent des indicateurs pour mesurer la productivité des chercheurs et chercheuses, l'influence des revues, des maisons

1. CODATA-ICSTI Task Group on Data Citation Standards and Practices.

d'édition et des pays, ainsi que le flux des idées circulant entre disciplines et au fil du temps (Borgman et Furner, 2002 ; Borgman, 1990 ; Cronin, 1984, 2005 ; Kurtz et Bollen, 2010).

Créditer les créateurs et créatrices de données est un défi bien plus complexe qu'il n'y paraît, malgré les efforts qui y sont consacrés. Les mécanismes techniques de citation ne sont que des caractéristiques superficielles des infrastructures de la connaissance où ils sont intégrés. Des conventions sociales sous-tendent cette pratique, qu'il s'agisse de renvoyer à des publications, à des données, à des documents, à des pages web, à des personnes, à des endroits ou à des institutions. Pour chaque publication, les auteurs et autrices sélectionnent des objets dignes d'être cités. Ce choix est basé sur des pratiques encore mal comprises : qui choisit de citer quoi, comment, quand et pourquoi ? Les méthodes de citation varient grandement d'un champ à l'autre, comme le montrent les guides de styles disparates des publications en sciences exactes, sciences sociales, sciences humaines et droit. Les citations renvoient à des publications de nombreux styles, en mentionnant ou non la liste complète des auteurs, le titre de l'article, les numéros de page ou les identifiants chiffrés. Elles peuvent être explicites ou voilées et d'importantes sources de preuves peuvent n'être jamais mentionnées. Les références bibliographiques partent largement du principe que les objets sont des unités fixes, stables et complètes. Or, aucune de ces caractéristiques ne peut être présumée des données.

Les mécanismes de reconnaissance des contributions, d'attribution et de découverte des objets de recherche sont indissociables de la communication savante ; pourtant, les études sur les pratiques sociales de citation font cruellement défaut. Les méthodes de citation s'apprennent par imitation sans être enseignées. Les pratiques émergent et évoluent au sein des communautés de manière indépendante, voire isolée ou opposée à celles des communautés voisines. L'attribution du crédit auctorial varie également grandement d'un domaine à l'autre, ce qui crée des conflits dans les collaborations et de la confusion de part et d'autre des frontières disciplinaires. Déterminer qui devrait se voir attribuer le mérite d'un produit ou d'un processus dépend en partie desquels sont jugés dignes de reconnaissance.

Des choix pragmatiques, effectués par des praticiennes et des praticiens peu au fait des grands principes de reconnaissance scientifique, d'attribution, de découverte, d'identité, de pérennité et de contrôle bibliographique, peuvent semer le chaos dans les infrastructures de la connaissance. Les producteurs de bases de données sont connus pour réorganiser l'ordre des auteurs et autrices dans les articles afin d'améliorer l'efficacité de leurs algorithmes de tri. On sait que les maisons d'édition modifient des Digital Object Identifiers (DOI) pour des raisons marketing. Des

rédacteurs et rédactrices omettent les initiales du second prénom des auteurs cités plutôt que de vérifier les noms dans les sources d'origines. Bref, il n'y a pas de limites aux possibilités d'altération de l'intégrité des citations de publications et de données. De petites décisions peuvent avoir de lourdes conséquences lorsqu'elles se propagent d'un domaine à l'autre et au fil du temps.

Le choix des normes, des politiques et de mises en œuvre en matière de reconnaissance des créateurs et créatrices des données porte de lourds enjeux. La réussite des méthodes de citation de données dépend de leur adoption par la communauté, c'est-à-dire par les chercheurs et chercheuses qui écrivent les publications et y mobilisent les données comme preuves. Elle dépend aussi des investissements réalisés dans les infrastructures de la connaissance pour rendre la citation de données réalisable et avantageuse. Décrire et organiser les données de façon à les rendre découvrables pourrait nécessiter des moyens humains considérables. Le présent chapitre étudie l'intégration des pratiques de citations dans la théorie et la pratique de la recherche et propose une vision plus vaste de l'attribution des données dans le cadre des infrastructures de la connaissance.

Principes et problèmes

Parmi les problématiques de la citation de données, on trouve : la répartition de la reconnaissance des contributions de multiples parties à des données spécifiques ; le cadre juridique pour la licence, la propriété et le contrôle ; la granularité des objets à citer ; le traçage de la provenance sur de longues périodes ; la maintenance de l'intégrité et de la vérifiabilité des données ; l'intégration à des mécanismes de contrôle bibliographique existants ; l'intégration à des normes et technologies de réseaux numériques anciens ; la découvrabilité par des êtres humains et des machines ; la gestion et l'intendance des données ; la facilitation du partage et de la réutilisation ; l'identité des individus et des organisations associées aux données ; l'intérêt des mécanismes de citations pour des usages accessoires, comme l'évaluation et les politiques et enfin, la prise en compte des pratiques disparates des différentes disciplines et parties prenantes. C'est là une liste bien longue de questions pour quelque chose d'apparence aussi simple que la « citation de données » (Borgman, 2012b).

Un point de départ fécond consiste à évaluer le système de communication savante où la citation bibliographique s'insère, ce qui permet de démêler quelque peu l'écheveau théorique qui tараude la problématique de la citation de données. Ces éléments peuvent être appliqués à la reconnaissance des contributions, à l'attribution et à la découverte de données. À ce carrefour, nous pourrions explorer de nombreuses

questions, dont nous allons choisir les plus saillantes : comment citer, pourquoi citer, comment reconnaître les contributions, comment attribuer la responsabilité, comment identifier les personnes et les objets, comment mettre en œuvre la citation de données, le rôle des citations dans la reconnaissance et l'attribution, les différences de principe entre référence bibliographique et citation de données et la disparité des préoccupations des parties prenantes en matière de communication savante.

Les parties prenantes préoccupées par les politiques et les infrastructures des sciences ont formulé des séries d'exigences en matière de citation de données. L'ensemble de principes le plus complet publié à ce jour a été promulgué par un groupe de travail international (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013).

1. Principe de statut : les citations de données doivent se voir accorder la même importance que la citation d'autres objets dans les documents de recherche.
2. Principe d'attribution : les citations doivent faciliter l'attribution du mérite et de la responsabilité juridique à toutes les parties à l'origine de ces données.
3. Principe de pérennité : les citations doivent être aussi durables que les objets cités.
4. Principe d'accessibilité : les citations doivent faciliter l'accès aux données elles-mêmes, ainsi qu'aux métadonnées et à la documentation associées afin que les êtres humains comme les machines puissent faire un usage éclairé des données référencées.
5. Principe de découvrabilité : les citations doivent permettre la découverte des données et de leur documentation.
6. Principe de provenance : les citations doivent faciliter l'établissement de la provenance des données.
7. Principe de granularité : les citations doivent permettre les descriptions les plus fines nécessaires à l'identification des données.
8. Principe de vérifiabilité : les citations doivent comporter suffisamment d'informations pour identifier les données sans équivoque.
9. Principe des standards de métadonnées : les citations doivent employer des standards de métadonnées largement acceptés.
10. Principe de flexibilité : les méthodes de citations doivent être suffisamment souples pour prendre en compte la variété des pratiques d'une communauté à l'autre, mais sans différer au point de compromettre l'interopérabilité des données entre communautés.

Dès la parution du rapport du CODATA-ICSTI, d'autres groupes ont entrepris de discuter cette série de principes et de les affiner. À l'heure où nous écrivons, les

représentantes et représentants de bibliothèques, de maisons d'édition, d'organismes de politique scientifique, de référentiels de données et d'autres secteurs sont parvenus à un consensus sur un ensemble semblable, mais plus concis, de huit principes : importance, reconnaissance et attribution, preuve, identification unique, accessibilité, pérennité, versionnage et granularité, et enfin interopérabilité et flexibilité (Data-citation Synthesis Group, 2014). Des groupes de mise en œuvre sont aussi en train de se former.

Parvenir à un consensus sur ces principes a nécessité plusieurs années de discussion. Ils représentent des règles de fonctionnement fondées sur le besoin de mécanismes viables pour la reconnaissance et la découverte, mécanismes qui pourraient s'incarner dans les nombreux acteurs responsables de la dimension institutionnelle de la communication savante. À l'inverse, la présente réflexion part du comportement des scientifiques d'un point de vue théorique et empirique. Les chercheurs et chercheuses poursuivent des myriades de pistes de recherche par une multitude de méthodes, s'appuyant sur des preuves nouvelles comme anciennes. Ces sources de preuves peuvent être statiques ou dynamiques, claires ou contestées, simples ou complexes et rares ou abondantes. Les publications fondées sur ces sources peuvent porter le nom d'un seul auteur ou autrice ou de milliers. Les critères de paternité varient grandement d'un champ à l'autre, tout comme les critères qui déterminent si des actes ou des objets sont dignes d'être cités. Les cadres techniques des citations bibliographiques sont, au mieux, fragiles, après des siècles d'évolution pour prendre à charge la diversité des pratiques. Ce sont notamment les cadres techniques qui s'adaptent aux pratiques et non l'inverse. Instaurer un cadre technique avant de demander aux scientifiques de l'adopter est une démarche risquée. Il est plus prometteur de partir des pratiques de reconnaissance, d'attribution et de découverte des objets de recherche.

Théorie et pratiques

Le rôle des citations dans la recherche a éveillé la curiosité des sociologues à partir du milieu du ^{xx}^e siècle, donnant lieu à une littérature foisonnante sur la communication savante et la bibliométrie, y compris plusieurs revues et séminaires dédiés. La bibliométrie, c'est-à-dire l'étude des relations au sein de la littérature publiée, est toutefois bien plus ancienne. Certains la font remonter aux talmudistes du Moyen Âge (Paisley, 1990) ; d'autres, sur la base d'analyses textuelles jugées analogues aux pratiques contemporaines, estiment qu'elle existait déjà plusieurs siècles avant notre ère. En ce qui concerne l'usage moderne des citations pour la découverte d'informations, on en situe généralement l'origine aux indices de citation Shepard en droit.

Apparu à la fin du ^{xix}^e siècle, le système Shepard utilisait des pastilles et, plus tard, un index imprimé pour relier des affaires à des décisions ultérieures ayant conduit à les confirmer, les casser ou les corroborer. À la fin du ^{xx}^e siècle, ces liens ont été automatisés. Des affaires peuvent désormais être « shépardisées » (*shepardized*) dans LexisNexis pour déterminer leur statut juridique actuel.

En 1955, Eugene Garfield a conçu le Science Citation Index (SCI) en inversant les listes de références d'articles de revue pour permettre la recherche en fonction des citations reçues (Garfield, 1955). Au milieu des années 1960, le SCI, généré automatiquement, était publié au format papier ; dès le début des années 1970, il était devenu une base de données en ligne. Le Social Sciences Citation Index et l'Arts and Humanities Citation Index ont suivi. L'établissement de liens entre citations et le suivi de renvois d'un objet à l'autre font partie des techniques utilisées par les moteurs de recherche d'aujourd'hui.

Style et substance : comment citer

Les mécanismes de citation sont tellement ancrés dans les pratiques des auteurs, des autrices et du lectorat qu'ils sont employés sans qu'on se préoccupe de leurs principes et présupposés sous-jacents. La langue commune a tendance à confondre des concepts distincts. Par exemple, on fait une référence (*to make a reference*), alors qu'on reçoit une citation (*to receive or to accrue citations*). Le document référençant, ou citant, contrôle la forme de la citation reçue. L'auteur citeur peut décrire le document cité de façon complète et précise en se servant du style de référence le plus couramment utilisé dans les domaines des auteurs cités. À l'inverse, il peut aussi mal orthographier leurs noms, les omettre ou en changer l'ordre, introduire des erreurs dans le titre, la date, le volume, le numéro, les numéros de page ou d'autres éléments et employer un style bibliographique différent. Les auteurs qui suivent habituellement les consignes stylistiques de l'American Psychological Association (APA), par exemple, pourront voir leurs publications décrites différemment lorsqu'elles sont citées dans des revues juridiques, de sciences exactes ou de sciences humaines. Une fois créées, les erreurs et les variantes tendent à proliférer. Dès lors que la publication – ou le jeu de données – est lâchée dans la nature, les auteurs n'ont guère de contrôle sur la façon dont elle est citée, utilisée ou interprétée, comme nous l'avons vu au chapitre 8.

Lorsqu'un auteur ou une autrice cite une autre publication dans la liste des références, les notes de bas de page ou la bibliographie, une relation se crée entre le document citant et le document cité. Dans un monde purement papier, la relation du document citant au document cité est unidirectionnelle. Il a fallu attendre que le système Shepard

et le Science Citation Index viennent inverser les listes de références pour que les citations soient traitées comme des relations bidirectionnelles. Dans un monde purement numérique, ces relations peuvent devenir des liens automatiques. Leur efficacité dépend de l'exactitude de la référence et de la capacité à identifier de manière unique les objets citants et cités. Elle peut aussi dépendre de la participation des médias de publication à l'infrastructure technique qui permet l'établissement de liens. Ces infrastructures sont, à leur tour, fondées sur des décisions de génie logiciel concernant les structures de données symétriques, qui ont émergé au terme de longs débats dans les sciences numériques. Aucune de ces relations n'est facile à établir et les mécanismes qui permettent le lien sont invisibles aux utilisateurs et utilisatrices. L'auteur et le lectorat découvrent des liens du citant au cité – ou du cité au citant – qui fonctionnent en toute fluidité. Ils sont aussi confrontés à des liens brisés, mal orientés, inexistants ou qui les amènent à une page d'autorisation ou d'accès payant.

Comprendre pourquoi certains liens fonctionnent et d'autres pas suppose une connaissance poussée du système de la communication savante et des technologies qui le rendent possible. Pour la grande majorité des utilisateurs et utilisatrices, le système est opaque. L'invisibilité de l'infrastructure la rend facile à utiliser, mais camoufle son intrication. Il faut, pour appliquer aux données les principes et les mécanismes de l'établissement de liens entre publications, dévoiler un peu de cette complexité.

Sont aussi invisibles les accords sur les métadonnées et les styles de présentation qui sous-tendent le processus de citation. Presque tous les styles bibliographiques s'accordent sur des éléments basiques : auteur ou autrice, titre de la publication et date de parution. Au-delà, les métadonnées ont tendance à varier en fonction du genre du document : volume, numéro et numéros de pages pour les articles de revues, maison et lieu d'édition pour les livres. Ces métadonnées sont cohérentes avec les systèmes de contrôle des maisons d'édition et avec les systèmes de catalogage et d'indexation mis en œuvre par les bibliothèques, lesquels représentent deux acteurs clés dans le système de la communication savante. Les métadonnées de localisation et d'identification, comme les URN (Uniform Resource Names) et les DOI (Digital Object Identifiers), sont venues plus tard, de même que les métadonnées spécifiques à des domaines de recherche, à des systèmes et à des mécanismes de classification.

Le choix des métadonnées est distinct de celui du style où elles sont présentées dans la référence. Les guides de rédaction courants tels que ceux de l'American Psychological Association, de la Modern Language Association, du droit (*Bluebook*) et du Council of Science Editors, varient à plusieurs égards : ordre des métadonnées, utilisation des noms complets ou des initiales des auteurs et autrices, abréviation

des titres de revue et inclusion d'autres métadonnées. Certains styles demandent une numérotation dans le texte et les listes de références ; d'autres utilisent des renvois insérés dans le corps du texte et suivent l'ordre alphabétique. Les outils de gestion bibliographique comme Zotero, EndNote et Mendeley enregistrent des métadonnées descriptives et permettent l'ajout de tags et de notes. Ils peuvent présenter les métadonnées sous forme de références bibliographiques dans des milliers de styles. Les technologies pour lier documents citants et cités, telles que CrossRef, sont indépendantes du style de citation (Council of Science Editors and the Style Manual Committee, 2006 ; CrossRef, 2009, 2014 ; EndNote, 2013 ; Mendeley, 2013 ; American Psychological Association, 2009 ; Harvard Law Review Association, 2005 ; Zotero, 2013 ; Modern Language Association of America, 2009).

Ce ne sont là que quelques exemples de l'invisibilité de l'infrastructure qui sous-tend la citation bibliographique. Celle-ci évolue depuis des siècles et ses racines n'apparaissent qu'aux spécialistes. Ce système étant suffisamment robuste pour la plupart de ses usages, la fragilité de ses fondations n'est pas flagrante. Il soutient adéquate-ment la découverte d'informations, les fonctions basiques de mention des sources et l'attribution de la paternité. Cependant, ses fondations se fissurent lorsque les références sont employées à des fins annexes, telles que le décompte des citations pour évaluer l'influence d'auteurs et d'autrices, de revues, de maisons d'édition et de pays, ou la cartographie des flux de connaissances à travers le temps, l'espace géographique et les frontières disciplinaires, ou encore la citation de données.

Théories du comportement citateur : quoi citer, quand et pourquoi

Les questions de qui choisit quoi citer, quand et pour quelles raisons représentent les domaines les plus problématiques et les moins explorés du processus de citation. Elles constituent aussi un terrain prometteur pour bâtir une théorie de la communication savante. La recherche sur les utilisations et les réutilisations des données devrait façonner un système robuste pour la citation.

Sens des liens

Dès le départ, traiter les liens entre publications comme une base pour la découverte, la reconnaissance, l'attribution et les mesures d'évaluation s'est avéré problématique. Cette conception part du principe que la relation a du sens et que ce sens peut être comptabilisé de manière objective. La critique de l'usage des citations comme des indicateurs quantitatifs pour cartographier la recherche, formulée par David Edge (1979), trouve encore un large écho aujourd'hui. Son analyse nuancée d'historien des sciences révèle que les citations ne constituent pas une mesure objective de

l'influence d'un chercheur, d'une chercheuse, d'un écrit ou d'un projet. On ne peut comprendre le choix des documents référencés dans une publication quelconque qu'en examinant de près le processus de recherche – et encore. Même les récits *a posteriori* sont suspects, car les chercheurs peuvent justifier leurs choix après coup.

L'article d'Edge a constitué un cri de ralliement pour les spécialistes de la bibliométrie et beaucoup lui ont répondu (MacRoberts et MacRoberts, 1989, 2010 ; McCain, 2012 ; White, 1990). Howard White (1990) admettait que les conflits quant à la validité des indices de citations ne se résoudraient pas de sitôt sur le plan empirique. Il en a caractérisé les perspectives de manière succincte (p. 91) : « D'un côté, nous avons des personnes qui ne veulent penser que de manière biographique, en termes d'intérêts singuliers et de particularités individuelles. De l'autre, des personnes prêtes à faire confiance à des données extrêmement agrégées, existant à un haut degré d'abstraction, et à y chercher des schémas ».

Plus loin, White décrit ces distinctions comme des vues de la réalité respectivement « au sol » et « aériennes ». Il juge ces perspectives incommensurables. Dans la vision agrégée, on voit des schémas qui ne peuvent être perçus depuis le sol, mais le risque est de ne pas pouvoir les interpréter sans bien comprendre ce qui se trouve réellement en bas. Télescopes et microscopes sont tous deux nécessaires pour voir les données, mais ils offrent des points de vue contrastés.

Un modèle « aérien » valable exige une meilleure compréhension des pratiques locales. Cependant, malgré des décennies de recherche, on sait peu de choses sur ce qui pousse les auteurs et autrices à choisir ce qu'ils citent dans chaque communication, article, livre ou autre document. Si les citations peuvent être comptées et cartographiées comme autant d'indicateurs objectifs de relations, cela suppose que les auteurs choisissent leurs références parmi toutes les sources possibles de documentation, sélectionnant toutes celles, et seulement celles, qui sont nécessaires au propos. L'ensemble des références listées dans chaque article devrait donc être optimal, c'est-à-dire représenter une trace nécessaire et suffisante de toutes les sources pertinentes pour l'article. Un autre présupposé est que les références constituent des descriptions exactes et complètes des objets cités. Or, en pratique, aucun de ces présupposés n'est vrai ou généralisable.

Les vues aérienne et depuis le sol des citations sont sans commune mesure non seulement en raison de différences théoriques, mais aussi du fait de leurs méthodologies. Agrégées, les citations sont utilisées pour cartographier les flux d'informations entre champs, communautés et pays. Elles servent aussi à estimer l'influence des revues, des universités et d'autres grandes organisations. Là où un problème surgit,

c'est lorsque ces statistiques agrégées sont employées pour tirer des conclusions sur les individus au sein de ce groupe, ce qui constitue ce qu'on appelle une « erreur écologique » (Babbie, 2013 ; Robinson, 1950). Les citations ou tout autre indicateur sont distribués de manière inégale au sein de n'importe quel groupe. Les communautés influentes ne sont pas composées d'auteurs et d'autrices d'influence égale à l'échelle individuelle. Un petit nombre d'articles très cités peuvent fausser le décompte d'un groupe. De bonnes revues publient de mauvais articles et inversement. Utiliser les statistiques agrégées d'un groupe comme variable proxy du comportement ou de l'impact des individus qui le composent n'est pas correct au point de vue statistique.

Sélectionner des références

Les pratiques sociales en matière de sélection de références sont loin de l'idéal d'objectivité décrit plus haut. Les références sont utilisées pour étayer l'argumentaire d'une publication. Bien qu'il arrive que les auteurs et autrices citent des publications qui contredisent leurs résultats, dans l'ensemble ils tendent à faire davantage référence aux éléments qui corroborent leur propos qu'aux résultats divergents ou non concluants. Les auteurs lisent beaucoup de choses qu'ils ne citent pas ; ils citent aussi parfois des choses qu'ils n'ont pas lues, que ce soit des classiques d'une discipline, des éléments référencés dans d'autres articles ou des travaux de directeurs ou directrices de départements ou d'autres personnes en position d'autorité.

Les éléments jugés dignes d'une citation par les auteurs et autrices varient selon les publications, les médias, les lectorats et bien d'autres facteurs. Les auteurs lisent le plus attentivement lorsqu'ils rédigent des mémoires, des états de l'art et des ouvrages. Autrement, ils peuvent choisir ce qu'ils connaissent le mieux ou ce qui est à portée de main sur leur bureau ou leur ordinateur, plutôt que de faire une vaste revue de la littérature. Ils peuvent surreprésenter la revue à laquelle ils proposent l'article dans leurs références, afin de positionner leur argumentaire pour cette communauté et ce comité de lecture. Le phénomène inverse est la citation forcée, lorsque d'autres auteurs exigent une reconnaissance explicite sous forme de référence bibliographique (Cronin, 2005). Le choix des références peut ne refléter que les lectures du coauteur qui a effectué la revue de la littérature au lieu de celles de l'ensemble des auteurs. La longueur de la liste peut être contrainte par un nombre limite de pages ou de références accordé par la revue.

Les pratiques de citation sont aussi une affaire de personnalité. Certains auteurs et autrices appliquent le rasoir d'Occam et ne sélectionnent que les références nécessaires et suffisantes pour l'argumentaire de leur communication. D'autres sont expansifs et parsèment leur article de renvois pour piquer l'intérêt du lectorat et le

pousser à approfondir les sujets. D'autres encore utilisent les références pour se défendre de possibles accusations de plagiat ou de fraude.

On donne des références pour de nombreuses raisons, positives comme négatives ; cependant, chacune compte pour une unité lors d'une évaluation ou de la cartographie d'un réseau bibliométrique. Qu'elles consignent des faits, étayent des arguments, en réfutent, apportent un contexte méthodologique, identifient ou valident des travaux antérieurs des auteurs que l'article vient prolonger, rendent hommage à un modèle ou identifient des travaux connexes dans la revue visée, toutes les références sont traitées comme des actes équivalents.

Théorie et modélisation du comportement de citation

Voyant les références réduites à des décomptes simplistes, les scientifiques ont appelé de leurs vœux une théorie englobante du comportement de citation (Cronin, 1981, 1984 ; Zhang *et al.*, 2013). Faute de théories générales, beaucoup se sont efforcés de catégoriser les raisons pour lesquelles on fait une référence. Ben-Ami Lipetz (1965) a le premier élaboré une classification des citations. Son but était de réduire le « bruit » lors de la recherche d'informations à l'aide de citations, puisque certaines sont plus pertinentes que d'autres vis-à-vis du contenu de l'article. Lipetz a donc proposé une typologie de 29 « indicateurs relationnels » pour décrire les rapports entre publications citantes et citées. Ces indicateurs étaient regroupés en quatre catégories : « contribution scientifique originale ou intention de l'article citant », autre type de contribution, « identité relationnelle entre les articles » et « caractère de la contribution scientifique du papier cité dans le papier citant ». Les données sont mentionnées dans deux catégories : « transformation des données » et « données cumulées »². La distinction des citations de données parmi les autres types de références a donc constitué une préoccupation dès les débuts de la recherche d'information scientifique.

Les nombreuses tentatives de catégorisation des citations diffèrent en fonction de l'objectif de l'étude – certaines cherchent à développer des théories de la communication savante, d'autres à améliorer les systèmes de recherche – et de la littérature sur laquelle elles se fondent. Par exemple, les pratiques de citation dans les sciences humaines sont très différentes de celles des sciences exactes. Pour l'heure, les catégorisations du comportement de citation sont si diverses du point de vue théorique, méthodologique, des questions de recherche et de la littérature qu'aucune théorie

2. NdT : Traductions tirées de Évelyne Broudoux, « Autorité scientifique et épistémique à l'épreuve de la mesure des citations », *Études de communication* [En ligne], 48 | 2017, mis en ligne le 1^{er} juin 2017, consulté le 9 octobre 2020. URL : <http://journals.openedition.org/edc/6841> ; DOI : <https://doi.org/10.4000/edc.6841>

générale sur comment, quand et pourquoi les auteurs et autrices citent n'a pu être édiflée. Les nombreuses typologies ne peuvent être fondées en une seule du fait de leurs différences de périmètre, de présupposés et d'objectifs. Une théorie générale du comportement de citation bibliographique a peu de chance de s'affirmer vu le peu que l'on connaît sur les variations des pratiques.

Les auteurs et autrices, les objets, les données et les relations qui les unissent peuvent être modélisés formellement. SCoRO, la Scholarly Contributions and Roles Ontology (Ontologie des contributions et rôles scientifique) [Shotton, 2013], par exemple, est fondée sur les normes du web sémantique et décrit des classes, des propriétés d'objets, des propriétés de données, des individus nommés et d'autres caractéristiques des contributions scientifiques. Cette ontologie comporte plus d'une centaine de catégories d'individus nommés, qui établissent des distinctions fines entre fournisseur d'accès, agent, analyste de données, chercheuse, réviseur du manuscrit, ayant droit, scientifique, collecteur de fonds, etc. SCoRO ne trouve pas son origine dans des modèles théoriques de la communication savante : ses racines sont techniques, le projet visant à fournir une liste exhaustive de catégories et de relations pouvant être utilisées dans la publication sémantique. Des classifications similaires, mais plus simples, sont également développées par des auteurs et des maisons d'édition (Harvard University and Wellcome Trust, 2012).

La catégorisation varie en fonction de qui assigne : l'auteur, une étudiante, un indexeur, une éditrice, un lecteur qui annote le document ou une chercheuse étudiant les pratiques de citations. Les indexeurs et indexeuses – qu'ils soient des humains ou automatiques – peuvent catégoriser des citations en fonction de leur signification apparente, mais ne peuvent saisir l'intention des auteurs et autrices. Ces derniers citent ce qu'ils considèrent comme pertinent pour leur article, mais éprouvent souvent des difficultés à affiner l'objectif de chaque référence.

Les auteurs et autrices ne peuvent sélectionner des citations potentielles que dans l'univers des ressources à leur disposition ; or, ces univers sont très variables. Certains auteurs ont accès aux plus grandes bibliothèques du monde, alors que d'autres ne disposent que de ressources informationnelles minimales et d'un accès restreint aux publications récentes. La croissance de l'accès ouvert aux publications engendre une plus grande équité dans le réservoir des ressources accessibles aux scientifiques et à la population étudiante et modifiera sans doute les schémas de citation. Le nombre de références au sein des articles continue de grandir. On ne peut que rarement présumer que la liste des références d'une publication quelconque représente l'ensemble optimal de sources nécessaires et suffisantes pour son contenu.

Citer les données

Une chose que l'on sait sur la citation de données est qu'elle est rare. Les études montrent que seul un faible pourcentage d'articles comportent des données dans les bibliographies ou les notes de bas de page, même si leur nombre augmente depuis quelques années. Là encore, il est difficile de comparer les résultats parce que les citations peuvent être présentées sous forme de renvois bibliographiques, d'URN, de mentions dans le texte ou par d'autres moyens. Il est donc compliqué de comparer les chiffres. Dans les articles, les données sont souvent rapportées ou incluses sous forme de tableaux et de figures. Certains domaines publient des « *data papers* » (publications de données) et des « *instrument papers* » (publications d'instruments) pour reconnaître des contributions spécifiques. Ces publications, très citées, servent de proxy pour la citation des données proprement dites. À l'inverse, les auteurs et autrices ont aussi recours à des données sans les citer, comme les exemples de premier plan et d'arrière-plan évoqués au chapitre 8.

Déterminer ce qui constitue une citation ou une utilisation de données est complexe et dépend du contexte. Comme nous l'avons vu dans l'exemple de l'observatoire de rayons X Chandra au chapitre 8, les archivistes de données astronomiques ne sont pas d'accord sur la définition d'« observation », qui représente l'unité fondamentale pour la classification. La communauté a fini par parvenir à un consensus international sur les bonnes pratiques de statistiques d'utilisation des données. Leur mise en œuvre repose sur le travail des professionnelles et professionnels de l'information, qui se chargent de créer des liens entre publications et jeux de données, puisque les auteurs et autrices citent rarement de manière explicite les données qu'ils utilisent.

Clair ou contesté : qui crédite-t-on ? À qui attribue-t-on ?

La seule métadonnée sur laquelle tous les styles de citation sont d'accord est le *créateur* ou la partie à l'origine de l'objet cité. Dans la plupart des cas, le créateur est un auteur ou une autrice, ou plusieurs. Dans d'autres, il peut s'agir d'un organisme tel qu'un comité, une collaboration de recherche ou une agence, par exemple la National Science Foundation (NSF). Dans d'autres cas encore, les parties à l'origine des données peuvent être des contributeurs et contributrices, des relecteurs et relectrices, des traducteurs et traductrices, des interprètes, des artistes, des conservateurs et conservatrices, des illustrateurs et illustratrices, etc. Les mentions de responsabilité sont devenues si complexes que d'aucuns ont suggéré de les traiter comme des génériques de films, avec de longues listes d'individus et d'organisations recensés par catégories.

Nommer l'auteur cité

L'étape la plus simple du processus de citation est la création de la liste de références à inclure dans la publication. Chaque référence crédite les créateurs et créatrices en les nommant. Pour le citeur ou la citeuse, les noms de ceux qui ont créé le document cité peuvent sembler clairs : ils correspondent à la mention de responsabilité telle qu'elle figure sur le document. Dans les articles de revue d'aujourd'hui, les mentions d'auteur sont explicites et faciles à restituer dans les citations. Dans d'autres cas, le citeur doit décider au mieux qui doit être reconnu pour sa contribution. Le fameux rapport de la NSF sur la cyberinfrastructure est cité sous les noms de « rapport Atkins », « Blue Ribbon Panel Report on Cyberinfrastructure », qui est une forme abrégée de son nom complet, *Revolutionizing Science and Engineering through Cyberinfrastructure : Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure*, ou par les noms des auteurs du panel (Atkins *et al.*, 2003).

Pour maintenir des points d'entrée cohérents, les catalographes se réfèrent à des règles qui s'étendent sur des centaines de pages. En revanche, les citeurs et citeuses ont tendance à prendre des décisions *ad hoc* en matière de reconnaissance, bien que des consignes existent dans la plupart des manuels de publication. Lorsque les références sont inexactes, ces variantes prolifèrent dans les moteurs de recherche et les bases de données bibliographiques. En dépit de la cohérence des mentions d'auteur dans chaque publication, on trouve des renvois aux publications de Christine L. Borgman avec ou sans l'initiale et parfois avec le patronyme germanique « Borgmann ». Les auteurs et autrices aux noms courants peuvent être confondus quand les initiales sont omises : Clifford A. Lynch et Cecilia P. Lynch deviennent ainsi « Lynch, C. ». Les variantes des noms, des dates, des numéros de pages, des DOI, des URN et d'autres détails des publications mettent à mal la découverte d'information. L'exactitude des statistiques de citations s'en ressent.

Pour compliquer encore davantage les choses, la même personne peut publier ou être connue sous différents noms, comme nous l'évoquerons un peu plus loin à propos des questions d'identité. Les créateurs et créatrices de documents et de jeux de données fournissent souvent une forme de citation à privilégier, ce qui contribue à la cohérence sans toutefois la garantir. La reconnaissance du créateur du contenu de pages web et de littérature grise est moins uniforme encore. Les moteurs de recherche peuvent résoudre certains types d'ambiguïté dans les références, mais certaines formes peuvent avoir si peu en commun qu'il est difficile de voir qu'elles pointent vers le même objet.

L'honneur ultime est peut-être de voir ses idées mentionnées, mais sans être pour autant nommé dans une référence bibliographique. Une fois que des idées sont

communément acceptées, des citations peuvent être oblitérées par incorporation (McCain, 2012), ce qui brouille la piste intellectuelle. La notion de « diffusion des innovations », par exemple, est souvent évoquée sans référence à Everett M. Rogers, qui a forgé l'expérience (Rogers, 1962). Cette oblitération est fréquente dans la recherche et pas toujours intentionnelle. Ce qui est de notoriété publique dans un champ peut constituer une découverte dans un autre. Les scientifiques, comme les étudiants et étudiantes, peuvent ignorer les origines d'une idée. Dans certains domaines des sciences humaines, notamment, les auteurs et autrices font des références voilées aux idées d'autrui, car ils partent du principe que le lectorat participe à une conversation déjà entamée au sein de la communauté. Retrouver la source d'une idée peut tenir aussi bien de l'enquête historique et de la criminalistique que de la bibliométrie.

Négocier la reconnaissance auctoriale

Déterminer quelles parties ont le droit d'être nommées comme autrices d'une publication ou d'un jeu de données est un processus autrement plus complexe. La paternité et les autres formes de responsabilité sont des conventions sociales, qui varient selon les personnes, les laboratoires, les communautés, les médias et les époques. Jusqu'au milieu des années 1950, la plupart des publications n'étaient signées que d'une personne. À mesure que le nombre d'auteurs et d'autrices par article augmentait, la responsabilité d'une publication s'est diluée (Wuchty *et al.*, 2007). À la fin des années 1990, les articles comptaient souvent plusieurs auteurs, voire des centaines ; la proportion des publications d'un seul auteur continue de décroître (Davenport et Cronin, 2001 ; King, 2013).

Les collaborateurs et collaboratrices négocient qui est nommé en tant qu'auteur dans chaque article et dans quel ordre. Avec l'augmentation de la valeur du crédit auctorial et des citations, ces négociations se sont faites plus litigieuses. Rédiger le propos, recueillir des données, compiler la bibliographie, analyser les données ou fabriquer des instruments peut donner ou non le droit d'être nommé comme auteur. On confère parfois une paternité de courtoisie aux directeurs et directrices de départements ou de laboratoires qui ont recueilli les fonds, indépendamment de leur implication dans cette publication en particulier. La paternité peut être partagée entre différents articles : certains membres d'une collaboration seront nommés dans les *data papers*, d'autres dans les *instruments papers* et d'autres enfin dans les articles scientifiques, comme c'est le cas en astronomie. En recherche par sondages, celles et ceux qui ont rédigé les questionnaires peuvent ne pas être nommés comme auteurs dans chaque article ; pourtant, ce sont eux qu'il sera important de contacter en cas de réutilisation des données.

Le statut de premier auteur peut tourner selon un ordre préétabli ou être assigné à celle ou celui qui « en a besoin » pour un entretien d'évaluation, une recherche de poste ou un exercice national d'évaluation. Dans certaines disciplines, la position de premier auteur est la plus importante ; dans d'autres, c'est la dernière place qui est la plus prestigieuse. Les noms peuvent être classés par ordre alphabétique, parfois en deux listes : la première pour les étudiantes et étudiants et la seconde pour les professeures et professeurs. Les personnes indiquées comme auteurs ou autrices chargés de la correspondance peuvent être les plus influentes, quel que soit l'ordre d'apparition des noms.

Les publications comprennent souvent des sections de remerciements qui identifient des contributions autres que la paternité. Ces remerciements sont rarement pris en compte dans les évaluations bibliométriques. Les quelques travaux concernant leur rôle dans la communication savante confirment que des remerciements spécifiques peuvent représenter une importante documentation des relations entre personnes et idées (Cronin et Franks, 2006 ; Cronin, 1995). Ainsi, les sources de financement sont souvent nommées, mais pas sous forme de citations, pas plus qu'elles ne sont reconnues comme autrices. Une ontologie des sources de financement, contenant une taxonomie de plus de quatre mille noms normalisés de financeurs, est en cours de développement pour que les auteurs, les autrices et les maisons d'édition puissent taguer leurs sources dans les publications (CrossRef, 2013). En standardisant la forme de référence, les organismes de financement et les maisons d'édition espèrent améliorer le traçage des publications qui résultent d'un financement spécifique.

Lorsque les auteurs et autrices sont peu nombreux, la paternité de chaque article se négocie en interne entre les collaborateurs et collaboratrices. En revanche, dans les domaines où chaque publication est signée par un grand nombre d'auteurs, comme en médecine, en biologie et en physique, les maisons d'édition édictent des règles qui précisent ce qui donne droit à la qualité d'auteur (Committee on Publication Ethics, 2013 ; International Committee of Medical Journal Editors, 2013)³. On peut demander à tous les auteurs d'approuver le manuscrit final avant publication. Certaines revues demandent des déclarations ou mentions comportant différentes catégories qui identifient les contributions de chaque auteur à la recherche ou à la rédaction. Ces contributions peuvent comprendre la collecte de données, leur analyse, la rédaction et ainsi de suite, mais les catégories sont moins précises que la centaine de rôles de SCoRO (Shotton, 2013).

Dans des domaines comme la physique des particules, la paternité peut être collective. Par exemple, le premier article du CERN sur le boson de Higgs mentionne

3. Pour une version française des recommandations de l'ICMJE, voir Maisonneuve, 2019.

comme auteur « The Atlas Collaboration », suivi d'une liste de 2 392 noms (Aad *et al.*, 2012). Les critères de paternité sont spécifiés explicitement par la collaboration du CERN. Ceux-ci sont fixés pour une période prédéfinie afin que ceux et celles qui ont contribué au projet à ses débuts voient leur travail reconnu. Par conséquent, des personnes décédées peuvent être autrices (Mele, 2013). L'article d'Aad (*et al.*, 2012) « est dédié à la mémoire de nos collègues de l'ATLAS qui n'auront pu voir pleinement les effets et l'importance de leurs contributions à l'expérience ». Cependant, se montrer explicite n'empêche pas la controverse. Peter Higgs, qui a donné son nom au boson, n'est qu'un des quelques théoriciens qui ont suggéré son existence (Cho, 2012). Le prix Nobel de physique a été attribué à Higgs et à François Englert et non à la collaboration Atlas, comme certains l'espéraient.

Le nombre d'auteurs et d'autrices par article a crû plus lentement dans les sciences humaines, où la signature unique reste la norme dans de nombreux domaines. Les articles à plusieurs mains peuvent être un assemblage de sections signées par un seul auteur. Par exemple, en histoire de l'art et archéologie antiques, que nous avons décrites au chapitre 7, les articles à plusieurs mains sont organisés en unités distinctes, chacune signée d'un ou deux auteurs (Contadini *et al.*, 2002 ; Faoláin et Northover, 1998 ; Kurtz *et al.*, 2009).

Responsabilité

Les discussions d'aujourd'hui sur la qualité d'auteur raniment de très vieux débats sur la responsabilité dans l'invention d'idées ou de la création de documents. La notion de responsabilité individuelle et collective a varié au cours des siècles selon les cultures et les contextes (Eisenstein, 1979 ; Fitzpatrick, 2011). Autrefois, les écrits religieux, les œuvres d'art et les autres artefacts du patrimoine culturel ne portaient ni signature ni date. Depuis, la recherche s'efforce d'identifier leurs origines et leur provenance, qui peuvent rester contestées des siècles durant. Les scientifiques cherchent à reconstituer des textes qui ont été copiés, fusionnés, divisés, glossés, annotés, corrigés et traduits sur de longues périodes. Les textes étaient aussi transmis oralement, changeant légèrement au gré des répétitions et des rédactions, comme dans le cas des études bouddhiques au chapitre 7.

Les personnes qui consignent des idées contribuent à l'histoire documentée qui traverse le temps. Elles ne comptent pas seulement des savantes et savants, mais aussi des moines, des scribes, des boutiquières et boutiquiers et des employés et employés administratifs, qui créent des traces. Du travail que nous considérerions aujourd'hui comme du plagiat peut, dans d'autres contextes, être perçu comme de la recherche à part entière. Les frontières entre la rédaction, la correction et la copie s'estompent au fil du temps. De nombreux rôles peuvent s'avérer importants lorsqu'on retrace la

responsabilité de la création de documents. Un *tradent* (« passeur »), par exemple, est un producteur de texte qui transmet des vérités spirituelles, parfois de manière anonyme (Mayer, 2010). Les savants dont le nom a été oblitéré par l'incorporation de leurs idées dans le canon moderne (McCain, 2012) sont membres du plus vaste groupe des contributeurs et contributrices inconnus du savoir. C'est pour toutes ces raisons, et d'autres encore que certains préfèrent redéfinir la qualité d'auteur ou paternité (*authorship*) comme qualité de contributeur (*contributorship*) (Harvard University et Wellcome Trust, 2012).

Reconnaissance pour la création de données

La capacité à trouver et réutiliser des données s'améliore lorsque les parties responsables de leur création peuvent être identifiées. Cependant, la notion de responsabilité de la création des données n'est pas mieux comprise que celle de paternité dans le canon bouddhique. Les compilations de données, par exemple les cartes de navigation, les tables de logarithmes et les registres de recensement, restent utiles malgré l'anonymat des responsables de leur création. Les rôles des auteurs et autrices, relecteurs et relectrices, compileurs et compilatrices, contributeurs et contributrices, collectionneurs et collectionneuses sont difficilement séparables. La reconnaissance et l'attribution sont, dans certaines circonstances, distinctes, au sens où les parties utilisatrices des données peuvent avoir la responsabilité légale d'attribuer la source par une citation particulière (Pearson, 2012). La distinction entre création d'information et compilation de faits a des implications juridiques en matière de droit d'auteur.

La paternité au sens des publications ne s'applique pas bien aux données, pour les raisons présentées aux chapitres 3 et 8. Les publications sont des argumentaires présentés par des auteurs et autrices, alors que les données sont des preuves qui servent à étayer l'argumentaire. Ces données peuvent provenir de nombreuses sources et beaucoup de personnes, d'instruments et de procédures peuvent les avoir affectées en route. Une publication constitue un objet singulier, compréhensible par lui-même pour le lectorat visé. Les données ne peuvent être comprises isolément : elles dérivent leur sens d'un contexte et d'objets connexes, comme des protocoles, des logiciels, une instrumentation, une méthodologie et les publications qui les décrivent. Il est souvent impossible de nommer précisément l'ensemble des personnes responsables de la création d'un jeu de données, encore moins de les mentionner dans une liste d'auteurs et autrices. Lorsque les jeux de données sont vastes, comme un relevé astronomique, on rédige des *data papers* pour les expliquer afin que d'autres puissent s'en servir. Les citations de ces données pointent vers les *data papers* et non directement vers les jeux de données.

Quand nous avons interrogé pour la première fois les chercheurs et chercheuses du CENS sur la paternité de leurs données, il est apparu clairement qu'ils ne se retrouvaient pas dans cette terminologie (Wallis *et al.*, 2008). Les données sont associées à des publications, mais il ne s'agit pas d'une correspondance biunivoque. Un seul jeu de données peut donner lieu à de multiples articles et un même article peut s'appuyer sur plusieurs jeux de données. Une étude plus approfondie a révélé que la principale raison pour laquelle les données n'étaient pas versées à des référentiels était l'absence d'accord sur la responsabilité au sein de l'équipe. Il n'était pas clair qui, de la chercheuse ou du chercheur principal, de l'étudiante ou étudiant chargé de l'analyse des données ou d'un autre membre de l'équipe, devait assumer la responsabilité de la diffusion ou de la publication des données (Wallis *et al.*, 2013). Les étudiants et chercheurs postdoctoraux qui recueillaient et analysaient les données connaissaient le mieux leurs caractéristiques et leur provenance. Les chercheurs principaux sont légalement responsables du projet ; ce sont eux qui sont les auteurs chargés de la correspondance sur la plupart des articles. L'auteur chargé de la correspondance peut être la personne ayant l'adresse la plus stable et pas nécessairement la personne connaissant le mieux les détails de la conduite de la recherche.

La thèse de Jillian Wallis (2012) représente à ce jour le travail le plus complet sur les questions de paternité et de responsabilité de la gestion des données, bien qu'elle ne concerne qu'un seul centre de recherche, à savoir le CENS. Wallis a étudié comment les chercheurs et chercheuses percevaient leur responsabilité par rapport aux données, comment les tâches de gestion des données étaient réparties dans l'équipe, de quelles tâches les individus étaient tenus pour responsables et à quel niveau. Bien que les détails des tâches de gestion varient selon les sujets de recherche et les équipes, elle a observé différents schémas de responsabilité au sein des six équipes étudiées. La responsabilité changeait parfois d'épaules au cours d'un projet, lorsque les données étaient transmises d'une personne à une autre. La paternité des articles et la responsabilité de la gestion de données étaient entremêlées à bien des égards. Dans tous les cas, la « responsabilité des données » était une notion vague qui nécessitait souvent de longues discussions pour l'expliquer (Wallis, 2012, p. 174).

Attribuer les données facilite la réutilisation dans la mesure où les individus qui en sont responsables peuvent être contactés. Lorsque les données sont découvertes par le truchement des publications qui les décrivent, les auteurs et autrices sont le premier point de contact. Le mérite des données est donc associé au mérite des publications. Parce que les scientifiques tirent des avantages des références à leurs publications, la plupart préfèrent qu'on cite ces dernières plutôt que leurs jeux de données.

Lorsque les jeux de données sont cités indépendamment de publications, des questions de traçabilité se posent. Le problème de la provenance persiste à mesure que les jeux de données sont fusionnés et explorés. Le problème de l'inférence sur plusieurs phases, évoqué au chapitre 8, se pose lorsque des parties en bout de chaîne ont besoin de comprendre les premières phases de traitement des données. Les registres de provenance doivent accompagner les jeux de données sur plusieurs générations, ce qui suppose un investissement considérable dans la conservation. La provenance est aussi une affaire de reconnaissance des contributions. Bien que certains chercheurs et chercheuses placent leurs données dans le domaine public en abandonnant leurs droits, la plupart des créateurs et créatrices souhaitent être crédités lors des usages ultérieurs. Les registres de provenance peuvent comporter des contrats, comme des licences qui précisent ce que l'on peut faire avec des données, à qui il faut les attribuer et comment (Ball, 2012 ; Guibault, 2013).

Nom ou numéro : questions d'identités

Les références pointent vers des objets spécifiques, qu'il s'agisse de publications, de jeux de données, de personnes, d'endroits, de pages web, de documents, de logiciels, de processus informatiques ou d'autres entités. Dans l'idéal, ils sont identifiés de manière unique afin d'établir une relation précise entre l'entité citante et l'entité citée. Les identifiants et les fiches de métadonnées, qui constituent les formes usuelles de représentation des objets de recherche, doivent permettre aux gens et aux machines de repérer, découvrir et trouver des objets citants et cités. Par ailleurs, l'identification doit perdurer aussi longtemps que l'objet pour que la citation reste exacte et découvrable. En pratique, ni l'identité ni la pérennité ne sont absolues. Les gens changent de noms, les documents changent de version, les objets numériques changent d'emplacement lorsqu'ils sont transférés sur un autre ordinateur et changent de formes lorsqu'ils sont migrés sur de nouvelles générations de logiciels et cessent d'être identiques à l'octet près. Comme nous l'avons évoqué au chapitre 3, le document de référence est bien moins stable dans un environnement numérique que sur papier, qu'il soit une publication ou des données. Pour reprendre le point de vue de Herbert Van de Sompel (2013), il peut être nécessaire de bâtir des infrastructures non pas pour une seule « version de référence », mais pour plusieurs.

Identifier des personnes et organisations

L'identité et la pérennité sont des problèmes épineux dont les racines conceptuelles sont profondes. Les individus endossent de nombreux rôles et identités, malgré les efforts de Facebook pour les fondre en une : auteur, correcteur, étudiant, enseignant, employeur, employé, parent, enfant, ami, collègue, supérieur hiérarchique, citoyen,

frère, conducteur, membre et ainsi de suite. Les personnes peuvent porter sur elles de nombreuses formes d'identification, chacune arborant un identifiant unique au sein d'un espace de nom : permis de conduire, passeport, badge d'entreprise, carte électorale, cartes d'assurance, cartes de crédit, cartes de paiement, cartes grand voyageur, cartes de fidélité et cartes de club de sport. Chaque tentative de création d'une identification universelle se heurte aux mêmes problèmes : comment identifier une personne dans un but donné, qui est concerné et dans quelles circonstances l'identification est utilisable. Un permis de conduire est exigé pour louer une voiture, un passeport est nécessaire pour traverser une frontière internationale. Ces formes d'identification ne sont pas interchangeables à ces fins, mais toutes seront acceptées au moment de vérifier un billet d'avion pour un vol intérieur.

Établir une forme cohérente et pérenne d'un nom de personne est un chemin semé d'embûches. La notion même de nom de famille est très contestée et l'ordre des noms dépend largement du contexte. Faire cohabiter des variantes contemporaines de noms est difficile ; faire de même avec des variantes historiques exige des connaissances considérables. Les noms légaux apparaissent dans les jeux de caractères de leur langue d'origine. Quand ils sont translittérés en anglais, les idéogrammes chinois et les diacritiques hongrois disparaissent. En Asie et en Europe centrale, le nom de famille précède généralement le prénom. Ainsi, Berend Ivan en Hongrie et Ivan Berend aux États-Unis sont une seule et même personne. Les étudiantes et étudiants asiatiques en échange en Occident adoptent fréquemment des prénoms occidentaux, ainsi Ding Jian peut-il devenir James Ding ; les Occidentales et Occidentaux peuvent choisir un nom asiatique quand ils travaillent sur place. Dans les traditions latines et hispaniques, les conjoints et les enfants prennent des noms de famille composés afin de rendre hommage à leurs parents et à leur partenaire. Dans certaines régions, des préfixes et suffixes apparaissent à l'occasion d'un mariage ou même d'un diplôme. Les époux prennent parfois le titre honorifique de leur conjoint. Les individus s'identifient eux-mêmes par des noms appropriés à chaque occasion : signature d'une œuvre, document légal ou activité sociale. Ce ne sont là que quelques exemples de la variation des noms et de leurs usages au fil du temps, des langues, des régions et des contextes. Tout système d'information présumant que chaque personne dispose d'un nom unique et pérenne est voué à l'échec. Le défi consiste à regrouper de multiples versions de noms de manière suffisamment fiable pour que les systèmes puissent remplir leurs fonctions (Borgman et Siegfried, 1992).

Tout espace de nom où des personnes sont identifiées dispose de règles d'admissibilité, d'identification et d'application. Les noms d'auteurs et d'autrices ne font pas exception. Les règles de catalogage des bibliothèques comprennent des critères pour identifier les auteurs, les éditeurs et éditrices, les illustrateurs et illustratrices

et les autres contributeurs et contributrices. Ces règles instaurent une cohérence au sein des catalogues et entre les indexeurs. La cohérence interne peut cependant se faire au prix du conflit entre systèmes. Comme les règles de catalogage découlent de règles sociales, elles reflètent des cultures nationales et régionales et varient selon les pays et les continents. Depuis que l'automatisation des bibliothèques s'est accélérée dans les années 1960, les règles internationales se sont harmonisées, mais présentent toujours des variations. Les noms d'auteurs et d'autrices sont normalisés dans chaque pays par les bibliothèques nationales, mais les auteurs de livres sont mieux traités que ceux d'articles de revue.

Les fichiers de nommage des bibliothèques instaurent une forme privilégiée ou de référence parmi les variations des noms et des renvois. Par exemple, les auteurs et autrices qui publient sous pseudonyme peuvent être intégrés sous leur nom légal avec des références sous pseudonyme ou vice versa, selon le nom sous lequel ils sont le mieux connus. Ainsi, Samuel Langhorne Clemens publiait sous le nom de Mark Twain. Les entrées de catalogues Clemens renvoient vers Twain, du moins aux États-Unis. J. K. Rowling, autrice de la série « Harry Potter », a plus tard écrit sous un nom de plume, qu'elle a tenté de garder secret. Les registres sur les livres de la série sont saisis sous l'entrée J. K. Rowling avec des renvois depuis les variantes, comme Joanne Kathleen Rowling. Les catalographes détermineront un jour s'il faut ajouter des renvois depuis ses pseudonymes secrets.

Identité et découverte

Les systèmes d'information peuvent être divisés en deux grandes catégories : ceux qui organisent l'information au moment de son intégration et ceux qui l'organisent au moment de la recherche. Les systèmes de catalogage des bibliothèques appartiennent à la première catégorie : on investit dans l'infrastructure pour établir, coordonner et maintenir des formes d'entrée cohérentes. Ces lourds investissements sont rentabilisés à très long terme par la découvrabilité et la gestion de l'information. Les moteurs de recherche appartiennent à la deuxième catégorie : ils s'efforcent de désambiguïser et de mettre en correspondance les formes variantes au moment de la recherche. Cette dernière approche pose problème à grande échelle, puisque les formes que la machine ne parvient pas à mettre en correspondance sont renvoyées à l'individu qui effectue la recherche. Les longues listes de documents comportent des doublons en raison de la variation des noms d'auteurs et d'autrices, des titres d'articles, des dates et des autres descripteurs. Ces listes s'allongent avec la prolifération des revues scientifiques et l'augmentation du nombre d'auteurs par article. Les noms courants, tels que Smith, Jones, Garcia, Chen, Lee ou Nguyen sont difficiles à désambiguïser. Une autre difficulté est la croissance de la recherche automatique, où il n'y a pas d'être humain dans la boucle pour désambiguïser des noms semblables grâce à d'autres indices.

À mesure que le problème d'échelle de la désambiguïsation des noms s'accélère, on recherche des solutions techniques et politiques. Les systèmes et services qui créent des identifiants uniques pour les auteurs et autrices pourraient organiser les ressources au moment de leur intégration, soit en créant une trace, soit en la convertissant. Le VIAF, l'ORCID et l'ISNI sont des initiatives plus ou moins coordonnées pour normaliser les formes des noms. Le VIAF, c'est-à-dire le fichier d'autorité international virtuel ou Virtual International Authority File (2013), est un projet mené par les bibliothèques nationales et hébergées par l'OCLC (Online Computer Library Center). L'ORCID, pour Open Researcher and Contributor ID (Identifiant ouvert pour chercheur et contributeur), est conduit par le secteur de l'édition (Haak *et al.*, 2012 ; Open Researcher and Contributor ID, 2011). Le Code international normalisé des noms ou ISNI, pour International Standard Name Identifier, est hébergé par l'OCLC, mais vise à l'identification bien au-delà de l'auteur et du contributeur. Cette norme ISO est également employée pour les artistes, les interprètes et les autres formes d'ayant droit (International Standard Name Identifier International Agency, 2013).

Le VIAF est une initiative institutionnelle qui permet aux bibliothèques et à d'autres organisations d'employer des formes d'entrées préétablies dans leurs systèmes. Les auteurs et autrices, qu'ils soient vivants ou morts, ne participent pas directement au VIAF et la plupart ignorent son existence. L'ORCID, en revanche, dépend largement de la participation des auteurs et des institutions identifiés. Les individus sont encouragés à s'enregistrer pour obtenir un identifiant et ensuite s'attribuer leurs publications, créant ainsi une bibliographie en ligne de leur travail. Les maisons d'édition participantes mettent en œuvre l'ORCID en demandant aux auteurs d'inclure leur numéro ORCID au moment de soumettre leur article. Les universités et les autres organisations sont incitées à revendiquer les publications de leurs auteurs, créant des bibliographies et proposant d'autres services, comme des bases de données d'expertise professorale. L'ORCID se préoccupe essentiellement des auteurs contemporains, tandis que l'ISNI se consacre surtout à établir des identifiants pour les personnes et les traces qui figurent déjà dans des bases de données.

Dans la mesure où le VIAF, l'ORCID, l'ISNI et d'autres services sont adoptés et mis en place, l'infrastructure de gestion de l'information associée aux noms de personnes s'en trouvera consolidée. Leur réussite dépendra de leur capacité à gérer des problèmes épineux en matière d'identification des noms, de confiance, de coopération et de flexibilité. Aucune entité unique ne saurait établir la confiance à elle seule. De nombreux auteurs et autrices sont soupçonneux des initiatives des maisons d'édition ou d'autres autorités centrales. Certains d'entre eux consacrent d'importants efforts au maintien de sites personnels et à la gestion de leur présence en ligne. D'autres ne le souhaitent pas ou n'en

sont pas capables. Beaucoup préféreraient que les bibliothécaires se chargent d'entretenir leur présence bibliographique au sein de leur institution.

D'autres questions, plus larges, concernent les personnes qui ont l'autorité d'établir des identifiants, qui les entretiendront et qui les corrigeront. Certains de ces défis vont au cœur de la communication savante : qui a le droit de revendiquer une publication ? Des personnes non listées en tant qu'autrices peuvent-elles s'attribuer une publication ? Les universités peuvent-elles conférer des publications à des membres, passés ou présents, du corps enseignant, du personnel universitaire ou de la population étudiante ? Qui peut revendiquer les publications de personnes décédées ? Comment résoudre les conflits ? Les individus peuvent-ils maintenir des identités multiples ? Une scientifique peut-elle tenir ses publications de recherche à part de ses écrits de fiction ? L'adoption dépend aussi de qui met en place et entretient le système d'identification. Elle peut réussir dans la mesure où les maisons d'édition, les universités, les bibliothèques, les archives de données et les autres parties prenantes des systèmes opérationnels investissent les ressources humaines et techniques nécessaires. Mais dans la mesure où l'adoption dépend de l'investissement des chercheurs et chercheuses dans la prise en charge leur identité, ce n'est pas sûr. La situation la plus comparable serait l'adoption des référentiels institutionnels, où le taux de contribution des auteurs est faible. Leur réussite a largement reposé sur l'investissement des bibliothèques dans l'acquisition, le catalogage et le dépôt de publications au nom des auteurs affiliés.

Identifier des objets

Il n'est pas plus facile d'attribuer un identifiant unique à des objets de recherche qu'à des personnes ou à des organisations. La découverte d'informations dépend de l'identification d'éléments de manière unique et de leur association avec des éléments connexes. Une recherche sur le mot « Hamlet » qui remonte des centaines de traces est rarement utile, surtout lorsque les résultats comprennent aussi bien des pièces de Shakespeare que des hameaux (*hamlet* en anglais). Les auteurs et autrices doivent choisir quelle version d'une œuvre ils citent et dans quelle traduction, comme pour les références à Borges et Galilée que nous avons faites dans les chapitres précédents. Chaque partie prenante et chaque espace de nom ont leurs propres méthodes pour gérer les relations et les renvois. Par exemple, les principes de catalogage des bibliothèques sont fondés sur une hiérarchie des œuvres, des expressions, des manifestations et des éléments qui peuvent être intégrés dans les systèmes de recherche (Mimno *et al.*, 2005).

Les livres semblent être les objets les plus stables à identifier. Cependant, ils existent non seulement en de nombreux exemplaires, mais aussi en différents formats – relié

ou de poche –, en de multiples éditions numériques et sous différentes traductions : chacune de ces versions dispose d'un numéro unique dans l'espace de noms du Numéro international normalisé du livre ou ISBN (pour International Standard Book Numbers) [International Standard Book Number Agency, 2013]. Les variantes, par exemple un film, une pièce, une édition jeunesse ou une réédition par une autre maison, exigent de nouveaux ISBN. Les bibliothèques cataloguent chaque ouvrage avec suffisamment de métadonnées pour le distinguer des œuvres proches. Les bibliothèques de prêt affinent encore cette distinction en attribuant un numéro local unique à chaque exemplaire pour que ceux-ci puissent être empruntés par les usagers et usagers. Ces derniers disposent par ailleurs de numéros uniques de cartes de bibliothèque, qui sont propres à l'institution.

De même, les revues se voient attribuer un numéro international normalisé des publications en série ou International Standard Serial Number (ISSN) qui les identifie de manière unique (International Standard Serial Number International Centre, 2013). Les revues changent parfois de nom, ce qui aboutit à un nouvel ISSN. Par exemple, la revue *American Documentation* est devenue le *Journal of the American Society for Information Science (JASIS)*, puis le *Journal of the American Society for Information Science and Technology (JASIST)*, et enfin le *Journal of the Association for Information Science and Technology (JASIST)*. Malgré la continuité de la société savante et du comité de rédaction, la revue dispose de quatre ISSN. Les indicateurs de citations reviennent à chacun des noms et numéros séparément.

Les ISBN et ISSN, conçus à l'origine pour les publications papier, ne sont pas suffisamment précis pour gérer des articles de revue individuels. Les Digital Object Identifiers ont été créés par le secteur de l'édition à la fin des années 1990 pour assigner un identifiant unique et pérenne à des publications (Paskin, 1997, 1999). Ils ont été largement adoptés pour les articles de revue et attribués rétrospectivement à des documents plus anciens. Un DOI est une sous-catégorie de Handle, un système d'identifiants uniques et pérennes pour des ressources en ligne (Corporation for National Research Initiatives, 2013). À mesure que l'usage des DOI s'est démocratisé, il a perdu en cohérence, les identifiants servant parfois à faire référence à un article, parfois à citer des tableaux ou des figures individuels au sein d'une publication et, parfois à renvoyer à des données. Les formes alternatives des publications, comme les prépublications dans des référentiels, peuvent recevoir un DOI indépendant de celui de l'article publié. Les DOI sont aussi adoptés dans d'autres secteurs, comme celui du cinéma, ce qui mène à une application moins cohérente. Les mérites des DOI, URL, URN et des autres systèmes d'identification des objets numériques sont chaudement débattus (Altman et King, 2007 ; Van de Sompel et Lagoze, 2009 ; Van de Sompel *et al.*, 2012).

Derrière les débats sur le choix des identifiants d'objets se cache l'épineux problème de la granularité. Quelle unité de la publication faut-il citer ? La réponse se fait moins claire maintenant que les articles apparaissent en de multiples versions et que des sous-sections, comme des tableaux ou des figures, en sont citées individuellement (Cronin, 1994 ; Nielsen, 2011). La citation bibliographique répond à certains de ces problèmes en référant des publications complètes ou en effectuant un « lien profond » vers des numéros de page grâce aux notes de bas de page. Les références à des publications complètes, comme le préconise le guide de style que nous employons ici (*Chicago Manual of Style*, 2010), créent une liste unique d'ouvrages cités à la fin de la publication. L'usage d'expressions comme *ibid.* ou *op. cit.*, courantes dans les styles de citation du droit et des sciences humaines, conduit à décrire le même objet différemment dans la première référence, les références suivantes sur la même page et les références ultérieures dans le reste du document citant. Dans ce cas, une bibliographie listant tous les éléments en note peut être fournie à la fin de la publication, ou non. Lorsque les notes de bas de page se réfèrent seulement à des portions d'un document par des numéros de page, il peut être impossible d'identifier l'objet complet cité. Après avoir longtemps été un identifiant stable des publications papier, les numéros de page sont, dans les objets numériques, souvent vides de sens. En effet, la longueur de la page peut dépendre de la taille et de la forme de l'écran où elle est affichée ; si numéros de page il y a, ils sont assignés par l'appareil de visionnage.

Les annexes jointes aux articles de revue constituent un autre domaine où les unités de données posent question. De nombreuses revues, notamment dans les sciences exactes, demandent des informations complémentaires nécessaires à l'interprétation, à la vérification ou à la reproduction de la recherche. Ces annexes, qui peuvent comprendre des jeux de données, sont le plus souvent disponibles exclusivement en ligne par un lien depuis l'article. La prolifération des compléments a suscité des inquiétudes quant à la notion d'un rapport de recherche se suffisant à lui-même (Maunsell, 2010). Pour compliquer encore les choses, les moteurs de recherche indexent rarement ces ressources, ce qui diminue leur découvrabilité. Les bonnes pratiques en matière d'informations complémentaires promulguées par les organismes de normalisation internationaux distinguent entre le contenu intégral, le contenu annexe et le contenu connexe (National Information Standards Organization, 2013).

Chaque référentiel dispose de ses propres règles sur le périmètre d'un jeu de données ou d'une autre unité de dépôt (Gutmann *et al.*, 2009). Citer des données hébergées dans un référentiel est le cas le plus simple au point de vue de la granularité et le premier évoqué par DataCite. DataCite est une organisation internationale à but non lucratif qui cherche à faciliter la découverte, l'utilisation et la réutilisation des données. Ses partenaires comptent des bibliothèques nationales, des bibliothèques de

recherche, des sociétés savantes, des associations professionnelles, des organismes de normalisation et la DOI Foundation (Brase *et al.*, 2014 ; DataCite, 2013).

Les données sont particulièrement difficiles à identifier parce qu'elles peuvent consister en de nombreux types d'objets et en de nombreuses versions : échantillons physiques, traces numériques, jeux de données à divers niveaux de traitement, carnets de laboratoires, guides de codification, notes de terrains, documents archivistiques, photographies, annotations, etc. Le problème d'unité se démultiplie lorsque ces objets et d'autres éléments numériques tels que des conférences, des *slides*, des tableaux, des figures, des vidéos, des tweets et des billets de blogs reçoivent des identifiants uniques.

Les rapports entre ces nombreux objets sont rarement hiérarchiques. Ils constituent plutôt un réseau essentiel à la compréhension de la provenance d'un jeu de données. Les modèles formels de ces relations, comme l'Object Reuse and Exchange (ORE), peuvent aider à lier et découvrir les jeux, mais réclament un travail important (Pepe *et al.*, 2010). Les objets tirés de la recherche au sein du CENS pouvant être représentés dans ORE sont présentés à la figure 9.1.

Les données existent en de nombreuses unités et à de nombreux endroits et peuvent être citées de nombreuses façons à de nombreuses fins. Le principe de granularité évoqué plus haut encourage les auteurs et autrices à citer « les descriptions les plus fines nécessaires à l'identification des données » (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). Citer des unités de petite taille, tels des tableaux, des cellules de tableaux, des figures et sous-parties de figures peut faciliter la traçabilité, surtout si elles sont susceptibles d'être replacées dans le contexte d'unités plus grandes. En revanche, citer des données transmises en continu (*streaming*), où les jeux ne constituent qu'un cliché d'un instant T, est un tout autre défi. La capacité à identifier des données de manière unique et pérenne tout en facilitant la cohabitation d'éléments connexes est un problème classique du catalogage des bibliothèques, des pratiques archivistiques, de la recherche d'information et de la citation de données (Agosti et Ferro, 2007 ; Renear *et al.*, 2010 ; Svenonius, 2000).

Théorie et technologie : les citations comme actions

Désormais, les méthodes de citation s'incarnent dans les technologies permettant de créer, découvrir, chercher, fouiller, compter et cartographier des citations. Un simple clic sur la barre d'un explorateur peut créer une notice bibliographique dotée de métadonnées complètes pour présenter des références dans les styles de différentes revues. À l'heure où nous écrivons, Zotero intègre 6 789 styles

bibliographiques (Zotero, 2014). Les renvois d'un article à l'autre sont des liens cliquables. Les citations sont des termes de recherche pour trouver des articles. Les décomptes de citations d'auteurs et d'autrices produisent l'indice h, l'indice g et d'autres indicateurs de l'influence scientifique. Les décomptes de citations de revues deviennent les facteurs d'impact des revues (JIF, pour Journal Impact Factor), qui servent à classer les médias de publication et se matérialisent dans des listes de périodiques où publier pour obtenir un poste ou une promotion. Les cartes de citation servent à modéliser le flux des idées et l'influence des universités et des pays.

Figure 9.1. Exemple de cycle de vie scientifique du Center for Embedded Networked Sensing

Crédit : Pepe et al., 2010.

Certaines de ces technologies s'appuient sur les théories de la communication savante. D'autres sont des solutions techniques pour la gestion d'objets numériques, élaborées sans se référer aux diverses origines ou à la longue histoire du contrôle bibliographique. Dans un cas comme dans l'autre, la conséquence possible est que le code du logiciel détermine ce qu'on peut citer ou non, comment les citations sont

réalisées et ce qu'on peut en faire. Lawrence Lessig a expliqué comment le code peut verrouiller des pratiques et barrer la route à des influences essentielles, telles que les normes sociales, les marchés et le droit (Lessig, 1999, 2001, 2004). Les choix préliminaires sont importants : or, nous sommes aux débuts de l'élaboration des pratiques de reconnaissance, d'attribution et de découverte des données. Les choix faits dans les premiers jours des technologies, comme les claviers des machines à écrire, ont une influence à long terme que leurs inventeurs et inventrices n'auraient pu imaginer (David, 1985 ; Mullaney, 2012).

Dans un monde papier, les citations sont des liens stables entre des objets fixes. Dans un monde numérique, elles lient des objets mouvants. Ni l'objet citant ni l'objet cité ne peuvent être fixés indéfiniment en une forme ou en un emplacement. Pour utiliser les citations aux fins de la reconnaissance, de l'attribution et de la découverte, il faut imposer des notions de fixité à l'infrastructure. Des identifiants uniques et pérennes, par exemple, sont essentiels pour maintenir la provenance. Si les objets changent de version, de nouveaux identifiants et liens sont nécessaires. En retour, il faut prévoir des règles de versionnage pour établir quel degré de changement constitue une nouvelle version. En génie logiciel, le contrôle de version est codifié. Les domaines de recherche reposant sur les mégadonnées, comme l'astronomie, codifient souvent les versions sous forme de *data releases*. Cependant, dans la plupart des contextes, le contrôle de version est une affaire de pratiques locales.

Risques et récompenses : les citations comme monnaie d'échange

L'un des principaux arguments en faveur de la citation de données est que créditer les chercheurs et chercheuses les incitera à partager leurs données. Cette hypothèse, bien que répétée à l'envi, reste à prouver. Les citations de données peuvent bien être appréciées, en particulier quand les jeux sont largement utilisés. Cependant, le crédit qu'offrent les publications est tellement plus précieux que certains chercheurs découragent la citation de données. Ils préfèrent que leurs articles soient cités en substitut de leurs données. Les intérêts des scientifiques dans la citation de données semblent varier selon le but de celle-ci. Par exemple, l'équipe chargée de suivre l'utilisation des données de Chandra a découvert que les chercheurs étaient prêts à apporter leur aide à l'établissement de liens entre données et publications lorsque ces liens apportaient une valeur ajoutée scientifique aux documents. Ils étaient bien moins enclins à consacrer du temps à la citation et aux liens quand il s'agissait de rendre compte de leur travail (Winkelman et Rots, 2012a ; Winkelman *et al.*, 2009).

Plus les indicateurs de publication et de citation sont utilisés dans le recrutement, la promotion et l'évaluation, plus ils sont scrutés. Tout indicateur peut être détourné, en particulier des mesures uniques comme le décompte des citations. Des auteurs et autrices peuvent se citer eux-mêmes, leurs collègues, leurs étudiantes et étudiants et leurs mentors et réduire le nombre de références à leurs concurrentes et concurrents. La production académique peut être « saucissonnée » en petites unités pour augmenter le nombre de publications et de citations. La paternité honoraire et d'autres méthodes d'augmentation des taux de citation peuvent être difficiles à repérer ; c'est l'une des raisons qui ont poussé la déontologie en matière de publication à se codifier (Committee on Publication Ethics, 2013). La citation de données peut être détournée de la même façon, en particulier du fait du problème de la granularité. Pourquoi citer un jeu de données quand on peut citer individuellement cent ou cent mille objets ?

Les faiblesses des indicateurs de citation sont bien connues, car ils sont étudiés depuis l'apparition des facteurs d'impact et autres indices. Ils impliquent trop souvent un raisonnement fallacieux, commettant l'erreur écologique d'appliquer des caractéristiques du groupe aux individus qui s'y trouvent. Les citations de revues ne se répartissent pas de manière homogène dans les articles qui y figurent : en général, ce sont quelques articles très cités qui en sont responsables. Lorsque les articles étaient unis dans des numéros de revue, la corrélation entre citation d'un article et citation de la revue était plus forte. Elle s'est affaiblie depuis que les articles peuvent être cherchés indépendamment de la revue (Lozano *et al.*, 2012). Le JIF ou facteur d'impact de revue, tel qu'il est calculé par Thomson Scientific (anciennement l'Institute for Scientific Information, aujourd'hui Thomson Reuters), fait partie des indicateurs les moins prédictifs de l'influence scientifique (Bollen *et al.*, 2009). Cependant, il reste l'un des plus utilisés pour évaluer les revues et les chercheurs et chercheuses, en dépit même des objections de comités de rédaction de revues très citées (Alberts, 2013 ; PLoS Medicine Editors, 2006). Créé au sein des sciences exactes, le JIF prend en compte les deux dernières années. Or, le délai de citation dans les sciences humaines et sociales est souvent bien plus long, ce qui rend le JIF d'autant moins valide dans ces disciplines (Borgman, 2007).

Les problèmes posés par les indicateurs de citation bibliographique ont mené aux *webmetrics*, *webometrics* ou indicateurs du web, qui appliquent les méthodes bibliométriques aux documents et aux liens sur Internet (Ingwersen, 1998 ; Thelwall *et al.*, 2005). Des chercheurs et chercheuses bien au fait de problèmes de fiabilité et de validité de la bibliométrie ont élaboré des modèles d'influence plus large que ceux pouvant être calculés à partir des bases de données des maisons d'édition. D'autres ont cherché à intégrer le décompte de la communication savante

informelle dans l'évaluation des scientifiques. L'*Altmetrics Manifesto* propose des indicateurs alternatifs (*altmetrics*) de l'influence et de la productivité scientifique (Priem *et al.*, 2010). Ils comprennent les téléchargements, les mentions dans des blogs, les annotations et tags et les apparitions sur les médias sociaux tels que Twitter et Reddit. Un jeune secteur économique s'est développé autour de ces indicateurs alternatifs et propose des décomptes aux maisons d'édition et aux autres médias, qui les affichent avec les articles. Auteurs, autrices, lecteurs et lectrices peuvent désormais voir combien de fois un article a été vu, cité, mentionné ou partagé et suivre ces liens (Chamberlain, 2013 ; Fenner, 2013 ; Thelwall *et al.*, 2013 ; Yan et Gerstein, 2011).

Ces unités distinctes de communication savante sont utiles pour découvrir des objets connexes, mais leur validité en tant qu'indicateurs alternatifs de la productivité scientifique est discutable. *Stricto sensu*, un tweet annonçant la parution d'un article de revue constitue une citation de cet article. Malgré ses défauts, la citation bibliographique est fondée sur la pratique scientifique historique de mention des sources de preuve et d'influence. On sait bien moins de choses sur le sens des mentions dans la communication informelle ou dans la citation de données. Modeler les pratiques d'évaluation et de reconnaissance autour des données sur celles du contrôle bibliographique revient à reprendre ces hypothèses inépuisables.

La valeur des décomptes de citations, en particulier ceux émanant de revues indexées par l'Engineering Index de Thomson Reuters et Elsevier ont subi une inflation telle que la paternité s'achète et se vend pour de larges sommes. Une enquête menée par le magazine *Science* a révélé un « marché noir académique florissant » en Chine, où les noms auctoriaux sont remplacés dans les articles quelques jours à peine avant leur parution. Dans d'autres cas, les articles s'autoplagient par des traductions du chinois à l'anglais pour être repropasés à des revues anglophones. Auteurs et autrices, revues, comités de rédaction, agentes et agents et autres acteurs sont impliqués dans diverses manigances, où les tarifs peuvent atteindre le salaire annuel d'une professeure (Hvistendahl, 2013). L'enquête de *Science* s'est concentrée sur la Chine, où les chercheurs et chercheuses peuvent être largement récompensés s'ils publient dans ces revues et où les publications dans le Science Citation Index ont été multipliées par six depuis 2000. On ignore l'étendue de la fraude ailleurs, mais maisons d'édition comme scientifiques reconnaissent l'existence de possibilités pour détourner les décomptes de citations.

Les chercheurs et chercheuses se sont sentis menacés par l'omniprésence de la citation et des autres indicateurs dans l'évaluation de la recherche, au point de signer la Déclaration de San Francisco, ou DORA (Declaration on Research Assessment).

Lancée par l'American Society for Cell Biology, cette initiative a depuis été soutenue par de nombreux scientifiques, revues et associations professionnelles. On a pu lire des éditoriaux au sujet de la DORA dans un vaste éventail de revues scientifiques et de titres de presse. La DORA offre des lignes directrices aux multiples parties prenantes de la communication savante et appelle à des méthodes plus nuancées et reposant sur des bases plus larges (Declaration on Research Assessment, 2013). Des projets comme ACUMEN (Academic Careers Understood through Measurement and Norms) promeuvent des démarches d'évaluation de la productivité et de l'influence scientifique plus holistiques, prenant notamment en compte le rôle des données (Research Acumen, 2013).

Trop de ces indicateurs se contentent de compter ce qui est facile à compter. Il est surprenant de voir combien de scientifiques, d'organismes de financement, de responsables politiques, de maisons d'édition, de bibliothèques et d'autres institutions impliquées prennent les décomptes de citations, les *altmetrics* et d'autres indicateurs pour argent comptant plutôt que de les soumettre aux exigences de la preuve scientifique.

Conclusion

La citation de données apporte une solution à un problème mal défini. Lui appliquer le modèle de la citation bibliographique au motif que publications et données méritent un statut équivalent est une erreur. Le véritable problème est de rendre les données découvrables. Les publications sont et resteront les étoiles et les planètes de l'univers scientifique. Les méthodes d'attribution ont seulement besoin d'éclairer suffisamment les données pour qu'elles cessent d'être de la matière noire. Les nœuds d'un réseau n'ont pas besoin d'avoir la même valeur. Des amas d'étoiles coexistent avec des régions intergalactiques désertes. Notre besoin essentiel est celui de véhicules qui puissent suivre le lien entre objets de recherche liés et ainsi leur permettre d'être découverts, explorés et combinés.

Les archives de données, les maisons d'édition et les bibliothèques sont des acteurs clés de la découverte de données parce que leurs services facilitent la gestion et la réutilisation. Une infrastructure de citation de données robuste résultera d'investissements massifs en ressources, dont des professionnelles et professionnels de l'information pour représenter les données d'une manière qui les rend citables et découvrables. Peu d'auteurs et d'autrices scientifiques font montre d'une diligence et d'une précision de puriste dans leurs références bibliographiques. Encore moins sont susceptibles de devenir des expertes et experts de la citation de données.

La citation bibliographique est vue comme la référence absolue, à laquelle la citation de données devrait aspirer. En réalité, elle est une infrastructure de la connaissance fragile qui remplit à peine les fonctions pour lesquelles elle est prévue. L'infrastructure a évolué et s'est adaptée à la transformation des pratiques et des technologies au fil des générations de savantes et savants. Elle fonctionne mieux sur le plan de la découvrabilité que de la reconnaissance et de l'attribution ou de la cartographie de la circulation des idées. Elle devient plus fragile à chaque nouvelle fonction qui lui est imposée. Les indicateurs de citation n'ont, quand ils sont appliqués à la productivité scientifique, jamais été soumis à des normes rigoureuses d'inférence statistique, de fiabilité ou de validité. Il est facile de les détourner ou de les falsifier. Pourtant, ces indicateurs sont toujours ancrés dans le système de récompense scientifique. La citation de données est un moyen de reconnaître le mérite d'une personne qui a sélectionné les données, les a recueillies, compilées, nettoyées, traitées, analysées, gérées, interprétées, explorées, combinées, mises sous licence, équipées, extraites, visualisées, présentées ou qui a joué avec. Elle n'est pas une fin en soi. Le fond du problème est de comprendre les nombreux rôles associés aux données et de parvenir à un consensus au sein des communautés sur ceux qui méritent d'être crédités et du meilleur moyen de le faire. La reconnaissance améliore alors la découvrabilité et la réutilisation. Une infrastructure de la connaissance solide qui intègre la reconnaissance des contributions aux données doit prendre en compte les galaxies d'acteurs variés et en concurrence, sans oublier les systèmes de motivation et de récompense de celles et ceux qui recueillent, créent, analysent, interprètent et présentent des éléments probants fondés sur les données, c'est-à-dire les scientifiques.

10. Que garder et pourquoi ?

Introduction

Les concepts de *big data*, de *little data* et parfois de *zéro data* restent mal compris. Les efforts entrepris pour améliorer la gestion des données, leur partage, la reconnaissance et l'attribution sont pleins de bonnes intentions, mais les parties prenantes ne sont d'accord ni sur les points de départ, ni sur les buts poursuivis, ni sur le chemin menant des uns aux autres. Faute de consensus sur les sortes d'entités qui constituent des données, il est encore difficile d'établir des politiques de partage, de diffusion, de dépôt, de reconnaissance des contributions, d'attribution, de citation et d'accès qui prennent en compte la diversité du travail scientifique des données dans l'ensemble des domaines. La pérennisation de l'accès aux données est une entreprise difficile et onéreuse, dont les coûts et bénéfices sont inégalement répartis entre les parties prenantes. Que garder et pourquoi : ces questions sont inséparables du qui, du comment, du pourquoi, du pour qui et du pour combien de temps. L'individu, qu'il soit scientifique, étudiant, bibliothécaire, archiviste, chargé de recherche ou éditrice aura, au mieux, une vision de fourmi de ce casse-tête élephanesque.

Certaines données méritent sans aucun doute d'être conservées indéfiniment et leur valeur est manifeste au moment de leur création. D'autres peuvent valoir la peine d'être gardées au cas où elles présenteraient un intérêt plus tard, soit seules, soit dans le cadre d'agrégats. Beaucoup ont une valeur éphémère, qui peut apparaître ou non au départ. Il est difficile de distinguer entre ces différents cas et surtout de le faire suffisamment tôt pour que les données soient enregistrées et sauvegardées avant de disparaître. Il est plus difficile encore de déterminer ce que sont ces données, puisqu'elles n'ont pas d'essence propre. Leur utilisation future peut dépendre de leur représentation, des phénomènes considérés et de l'évolution de ces représentations et phénomènes. Trouver de nouveaux usages à ces données peut maximiser leur valeur, mais ces investissements sont les plus spéculatifs de tous. Les chercheurs et chercheuses ont peu de raisons de conserver des données au cas où quelqu'un en voudrait pour un motif quelconque, sous une certaine forme, à un certain moment.

Bâtir des ensembles de données constitue la manière la plus évidente de mettre celles-ci à la disposition de futurs utilisateurs et utilisatrices. Les bibliothèques, les archives, les musées et les référentiels de données appliquent des principes professionnels pour sélectionner et estimer les objets à intégrer dans leurs collections. Les fonds de ces institutions de mémoire sont caractéristiques de la longue traîne : 20 %

des ressources reçoivent 80 % de l'usage. La politique de sélection s'appuie sur le suivi des utilisations, mais ces 20 % utiles changent continuellement. Certains objets connaissent une popularité immédiate qui décline progressivement. D'autres ont un usage limité au départ, mais suscitent une vague d'intérêt plus tard. D'autres encore semblent utilisés de manière aléatoire. Enfin, quelques objets, vierges de tout usage, attendent encore d'être découverts. Même les institutions les plus prestigieuses ne parviennent pas toujours à entrevoir la valeur future. La bibliothèque de Bodley (université d'Oxford) est par exemple célèbre pour avoir vendu son exemplaire du *Premier folio* de Shakespeare, apparemment parce qu'un bibliothécaire a considéré que le troisième folio, publié en 1664, pouvait le remplacer. Au début du ^{xx}^e siècle, elle a de nouveau acquis son *Premier folio* du ^{xvii}^e siècle pour une grosse somme. À l'occasion du 449^e anniversaire de Shakespeare, elle a publié un fac-similé numérique du précieux objet (University of Oxford, 2013a).

Pérenniser l'accès aux données de la recherche représente un problème d'infrastructure de la connaissance qui concerne tous les acteurs de la communication savante. Une question connexe, la préservation numérique, présente une infrastructure et des caractéristiques d'échelle similaires. Un panel international chargé d'effectuer des analyses économiques de la préservation numérique a identifié quatre difficultés structurelles : « 1) des horizons à long terme, 2) des parties prenantes dispersées, 3) des motivations faibles ou mal coordonnées et 4) un manque de clarté quant aux rôles et responsabilités des parties prenantes ». Le panel a émis trois recommandations : « formuler une description des bénéfices convaincante », « créer des incitations claires à préserver au nom de l'intérêt public » et « définir les rôles et les responsabilités des parties prenantes pour permettre une allocation constante et efficace des ressources à la préservation tout au long du cycle de vie numérique » (Berman *et al.*, 2010, p. 1-2).

Tout cela représente un défi de taille quand il s'agit de préservation numérique ou des données de la recherche. La sauvegarde des représentations numériques des données n'est qu'un des aspects de la pérennisation de leur valeur. Qu'elles soient numériques ou non, les données ne peuvent pas être considérées isolément. Elles sont indissociables des méthodes, des théories, des instruments, des logiciels et du contexte de recherche. Pérenniser l'accès aux données de la recherche suppose une conservation des objets individuels et des relations qui les unissent. Les moyens pour y parvenir varient selon les domaines, comme l'ont montré les études de cas. Si l'on reformule le problème des *big data*, *little data* et *zéro data* en termes de données ouvertes, le défi consiste à les rendre découvrables, utilisables, intelligibles et interprétables et de maintenir ces conditions pour une durée raisonnable. Une première étape consiste à examiner la distribution de ces problèmes parmi les scientifiques

et les communautés de recherche. Une deuxième est de considérer les types de données qui, quel que soit le domaine, méritent ce degré d'investissement, et les personnes qui en bénéficieraient. Une troisième étape consiste à déterminer qui serait prêt à effectuer ces investissements et où ceux-ci auraient les plus grands bienfaits. La description des bénéfices de l'accès aux données dépend d'un ensemble de facteurs étroitement interconnectés qui varie selon les parties prenantes. Nous allons jauger ces défis dans les termes des provocations présentées au chapitre 1.

Retour aux provocations

Ce qui mérite d'être gardé dépend de là où l'on se tient. Dans un petit nombre de champs, de grands projets de recherche développent des ensembles de données dans le cadre de leurs missions. Les observations qui ne sont pas saisies sont considérées comme des données perdues. Certains champs disposent de référentiels de données dans lesquels les scientifiques peuvent verser leurs données, mais beaucoup n'en ont pas. Dans la plupart des cas, ce sont les chercheurs, les chercheuses et les laboratoires individuels qui décident quoi garder, comment et pour combien de temps. Tandis que certains scientifiques conservent chaque note, document, ouvrage, article, objet numérique et échantillon physique recueillis au cours de leur carrière, beaucoup aimeraient disposer de meilleurs moyens pour gérer les données qu'ils jugent dignes d'être gardées. Cependant, l'attitude envers la diffusion de données tend à basculer à l'approche de la retraite. Pour assurer leur postérité intellectuelle, des scientifiques offrent parfois des collections longtemps conservées à des archives ou des référentiels pour une intendance durable.

De nombreux acteurs tireraient un bénéfice de la conservation et de la mise à disposition opportune d'un plus grand nombre de données. Aucune des parties prenantes du travail scientifique des données – scientifiques, population étudiante, universités, bibliothèques, archives, musées, organismes de financement, maisons d'édition, entreprises, contribuables, responsables politiques, patientes et patients, grand public, etc. – n'a un grand poids à elle seule. Ces défis sont collectifs et doivent être relevés comme des problématiques des infrastructures de la connaissance. Plus les acteurs sont nombreux autour de la table, plus les discussions ont des chances d'être sérieuses.

Droits, responsabilités, rôles et risques

La responsabilité des données de la recherche est diffuse puisque de nombreuses parties les manipulent de leur origine à leur interprétation, comme nous l'avons présenté dans la première provocation :

La reproductibilité, le partage et la réutilisation des données sont des questions débattues depuis des décennies, voire des siècles. Nous intéresser à qui possède les données de la recherche, les contrôle, y a accès et les pérennise permettra de déterminer comment exploiter leur valeur et qui pourra le faire.

Celles et ceux qui mènent les recherches ne détiennent pas nécessairement les droits de toutes les données qu'ils utilisent. Les droits légaux sur les données peuvent être associés à des instruments – comme en astronomie –, à des sociétés privées – comme dans l'utilisation des médias sociaux – ou avec des compilateurs, compilatrices, éditeurs et éditrices – comme dans la version numérique du canon chinois bouddhique. Au sein d'un laboratoire, la responsabilité de la collecte et de la gestion initiale peut incomber aux étudiantes et étudiants de deuxième cycle, mais la maintenance est laissée à la chercheuse ou au chercheur principal. Délibérément ou par défaut, l'entretien des données peut retomber sur les personnes qui ont l'adresse la plus stable plutôt que sur celles qui connaissent le mieux leurs origines. L'établissement de la propriété légale est un problème épineux en raison de la pléthore de pratiques, de parties et de juridictions impliquées. Que la propriété soit claire ou non, la responsabilité de la collecte, de l'analyse et de la gestion peut incomber à de multiples parties. L'expression juridique « la possession vaut titre » s'applique bien aux données de la recherche : ceux qui détiennent quelque chose, que ce soit un terrain ou des données, sont réputés propriétaires jusqu'à preuve du contraire. Or, les propriétaires peuvent faire ce qu'ils souhaitent de leur bien dans les limites de l'éthique et de la loi.

Stocker les données coûte cher, que ce soit sous forme physique ou numérique. Déjà en 2007, on créait l'information numérique plus rapidement qu'on ne pouvait produire d'appareils de stockage (Berman, 2008). Les chercheurs et chercheuses doivent se montrer sélectifs dans ce qu'ils conservent, comme toute personne qui crée ou utilise des ressources informationnelles. Les scientifiques sont généralement contraints par l'espace physique dans leurs bureaux, leurs laboratoires et leurs foyers. Les universités et les consortiums de recherche fournissent des serveurs partagés pour le stockage numérique, mais le coût peut en être facturé aux projets de recherche ou aux départements. Le stockage sur serveur distant (dans le *cloud*) est un modèle alternatif, mais encore insuffisamment fiable pour les données de recherche (Kolowich,

2014). Bien que les tarifs du *cloud* aient chuté, des projections indiquent que le coût du stockage à long terme de données stagne ou augmente (Rosenthal, 2010, 2014). La pérennisation de l'accès aux données de la recherche suppose que les informations nécessaires à leur interprétation – protocoles, guides de codification, logiciels, spécimens, métadonnées, standards, etc. – soient également maintenues. Ces ressources sont souvent plus volumineuses que les données elles-mêmes.

Quand il est question de rangement, les chercheurs et chercheuses ne sont vraisemblablement pas plus consciencieux que le reste de la population. Des visites impromptues dans les bureaux professoraux révèlent que le rangement par empilement est plus courant qu'une organisation méticuleuse. Même les membres de l'élite scientifique, comme les évaluateurs et évaluatrices de la revue *Science*, stockent la plus grande part de leurs données en local ; seuls 7,6 % des 1 700 personnes interrogées en archivent la majorité dans des référentiels communautaires (*Science Staff*, 2011). Peu d'institutions imposent des dispositions en matière de gestion de données, ce qui aboutit à des démarches *ad hoc* qui varient selon les laboratoires, voire selon les individus. Les scientifiques déposent leurs données lorsque c'est obligatoire, comme dans le cas des séquences de génomes, des documents d'essais cliniques et des enregistrements sismiques. Elles et ils garderont les données qu'ils pensent réutiliser un jour, mais savoir s'ils les conserveront suffisamment bien pour les réutiliser est une autre question. Lorsque les étudiantes, les étudiants et le personnel quittent un projet, leur expertise peut être irrémédiablement perdue. À mesure que les logiciels sont mis à jour et les ordinateurs remplacés, les fichiers de données peuvent cesser d'être interprétables, surtout si les données ne sont pas migrées vers les technologies de nouvelle génération. Les octets se corrompent, les liens se brisent, les réfrigérateurs à spécimens se vident et les documents se perdent dans les déménagements et les mises à niveau informatiques.

Il faut envisager la gestion des données comme une affaire institutionnelle plutôt que comme la seule responsabilité du chercheur ou de la chercheuse. Le niveau d'assistance fourni varie largement selon les domaines et les institutions. Dans certains champs, les chercheurs ont d'emblée accès à des suites logicielles, de normes techniques et de référentiels où acquérir et déposer des données, alors que d'autres communautés ne disposent de rien de tout cela. Les universités et les autres institutions de recherche sont parfois bien dotées en professionnelles et professionnels de l'information pour aider à la gestion des données, à l'assignation de métadonnées, à la migration vers de nouvelles plateformes, au dépôt, à la découverte et aux types de coordination, mais elles restent une exception. La véritable question est la distribution inégale des coûts et des bénéfices. Dans de nombreux secteurs, l'intendance des données est considérée comme une mission sans financement avec peu

de retombées directes pour l'institution. On ignore qui paiera le prix du stockage et de l'intendance des biens numériques. Pour les personnels de recherche, les bibliothèques, les départements et les chercheurs des universités, les données peuvent représenter une charge autant qu'un atout.

Partage des données

Le partage d'informations entre êtres humains ou entre machines est une activité complexe, comme formulé dans la deuxième provocation :

Transmettre des connaissances au fil du temps et dans différents contextes est difficile. Certaines formes et représentations de données sont aisément partagées d'une discipline à une autre, d'un contexte à un autre et d'une époque à une autre, mais beaucoup d'autres ne peuvent pas l'être. Il est nécessaire de comprendre quelles fonctionnalités sont importantes ou non afin d'inspirer les pratiques et politiques scientifiques et guider les investissements dans les infrastructures de la connaissance.

Les connaissances ont plus de chance d'être communiquées intactes lorsque les individus interagissent directement et de façon synchrone. Ils peuvent ainsi se poser mutuellement des questions et mettre leurs intentions au clair. Les tâches peuvent être expliquées, les pratiques montrées, les compétences identifiées et le savoir tacite rendu plus explicite. Plus l'on s'éloigne de l'interaction en face à face, plus il y a de médiation. Les métadonnées et les autres formes de documentation deviennent alors nécessaires pour découvrir, interpréter et utiliser les données. Cette documentation prend tout son sens au sein d'une communauté de recherche, où elle améliorera les échanges internes, mais peut aussi créer des barrières entre communautés. La capacité à partager la recherche dépend donc largement de qui gère les données pour quelles communautés et à quelles fins. Pour revenir à la dimension économique du chapitre 4 et aux exemples des études de cas, certains types de données de la recherche peuvent aisément être traités comme des réservoirs communs de ressources, alors que d'autres restent des biens privés. Certaines données sont diffusées en tant que biens publics et d'autres seront vendues comme biens de club. La différence réside dans l'« emballage » ou le traitement des données et non dans les caractéristiques propres de celles-ci. On peut obtenir la même série d'observations dans chacune de ces conditions économiques, quoique représentée différemment.

Un exemple de jeu de données disponibles de plusieurs manières en astronomie illustre les difficultés à évaluer les conditions de partage. Les observations peuvent être classées dans trois catégories générales selon la façon dont les chercheurs

ou chercheuses les ont acquises : ensembles de données, données recueillies à la source par les chercheurs et données dérivées. La NASA est la plus importante investisseuse dans les réservoirs communs de ressources de l'astronomie, en partenariat avec de nombreux autres organismes états-uniens et internationaux. Créer et gérer des ensembles de données fait partie de ses missions. Ces référentiels emploient des scientifiques, des spécialistes des données et des ingénieurs et ingénieures logiciels. Des investissements sont réalisés dans la migration, les outils d'exploitation et le personnel pour aider les usagères et usagers à découvrir, acquérir, utiliser et interpréter des données. Les astronomes font un grand usage de ces référentiels et les explorent des années après l'enregistrement des observations.

Les données recueillies à la source, la deuxième catégorie, sont des observations effectuées directement par les chercheurs et chercheuses en astronomie. Comme dans le cas du relevé COMPLETE au chapitre 5, certains astronomes rédigent des demandes de temps d'observation pour utiliser eux-mêmes les télescopes. Les observations ainsi obtenues peuvent être traitées, ou non, dans le *pipeline* associé à l'instrument. D'autres astronomes fabriquent leurs propres instruments pour effectuer des observations et conçoivent leurs propres *pipelines*. Ces données, une fois nettoyées et étalonnées, peuvent être versées à des référentiels ou à des ensembles institutionnels. D'autres sont diffusées directement sur des sites web de projet ou à la demande. Cependant, la vaste majorité des données recueillies à la source, même en astronomie, semblent rester sous le contrôle du chercheur.

Les données dérivées, qui représentent la troisième catégorie, sont celles obtenues d'archives ou combinées à partir de plusieurs sources. Lorsque les scientifiques utilisent des données tirées d'ensembles organisés comme l'observatoire de rayons X Chandra ou le Sloan Digital Sky Survey, elles et ils les transforment pour répondre à leurs propres questions de recherche. Souvent, ils les compareront aux données venant d'autres ensembles ou à des données recueillies à la source. Le relevé COMPLETE, par exemple, fusionne des observations tirées de différents référentiels avec des données recueillies à la source. Les données dérivées peuvent être transmises à d'autres référentiels, être diffusées directement ou, plus fréquemment, rester sous le contrôle des chercheurs et chercheuses.

Ce modèle tripartite peut s'appliquer à d'autres champs qui entretiennent des réservoirs communs de ressources. Le niveau d'investissement dans la conservation des données et la centralisation des ressources varie selon les référentiels. Certains de ces ensembles communs émanent d'une source unique. Ainsi, le Sloan Digital Sky Survey est constitué d'observations issues d'un seul télescope qui ont été recueillies de manière synoptique pendant de longues années. La CBETA a démarré

en numérisant un seul corpus du canon bouddhique chinois. D'autres ensembles acceptent des données de sources disparates en se reposant sur l'assurance qualité des créateurs et créatrices. Certains référentiels disposent de suffisamment de personnel pour vérifier les jeux de données, ce qui leur permet de n'accepter que ceux qui se conforment à leurs normes de qualité. Le personnel peut apporter une valeur ajoutée grâce à des métadonnées, de la documentation supplémentaire, la migration vers de nouveaux formats et une assistance technique. Bien qu'il y ait parfois des doublons entre ensembles, il y a surtout beaucoup de lacunes. Les scientifiques peuvent ne pas connaître les ensembles des domaines voisins et la découverte reste un problème.

Pour dériver des données de référentiels ou d'autres ressources, les scientifiques s'appuient sur les informations de provenance disponibles. La suite des analyses, des interprétations, de la gestion et de la documentation de provenance se trouve entre leurs mains. Les données dérivées ne semblent pas avoir plus de chances de trouver une demeure pérenne que les données recueillies à la source. Peu d'archives, qu'elles abritent des données ou autre, sollicitent le dépôt de produits dérivés de leurs ressources, en particulier s'ils ont été mêlés à des sources différentes. En effet, les archivistes montrent souvent une réticence à prendre en charge des objets dont elles ou ils ne peuvent vérifier la provenance ou les opérations qu'ils ont subies. Quelques ensembles de données, comme le Sloan Digital Sky Survey, acceptent les données dérivées après examen.

On notera que le partage des données est encore moins systématique dans les domaines où peu de réservoirs communs de ressources existent. Dans des champs comme la science et la technologie des réseaux de capteurs, les partenaires n'avaient d'autres choix que de maintenir leurs propres ressources en données. Celles-ci étaient recueillies à la source lors de déploiements avec des questions de recherche, des technologies et des protocoles variables. Les équipes mettaient en commun leurs ressources lorsque le besoin s'en faisait sentir, mais celles-ci pouvaient se révéler malgré tout incompatibles selon les tâches et les projets. Dans la recherche sur les médias sociaux, les scientifiques pouvaient dériver leurs données d'un flux commun, mais les transformer ensuite pour traiter leurs propres questions de recherche. Que les données de laboratoires différents soient comparables ou non, les contrats ayant permis leur obtention pouvaient prohiber le partage. L'exigence de plans de gestion des données de la part des organismes de financement peut conduire les scientifiques à prendre conscience de la valeur potentielle de leurs données. Cependant, à moins que des moyens efficaces de partage, comme des référentiels, des outils et du personnel, ne viennent derrière, ces plans pourraient échouer à promouvoir le partage et la réutilisation des données.

Publications et données

Les diverses analogies entre données et publications ont tendance à embrouiller les discussions en matière de politiques et de pratiques plutôt qu'à les éclairer, comme dit dans la troisième provocation :

En dépit de la prolifération des formes et des genres, les fonctions des publications scientifiques restent stables. Les données servent des objectifs différents des articles, des ouvrages et des communications. Traiter les données comme des publications risque de renforcer le poids des intérêts catégoriels au détriment de l'expérimentation de nouveaux modèles de communication savante. Il convient d'envisager les fonctions des données dans la recherche du point de vue de différentes parties prenantes.

Comme nous l'avons vu tout au long du présent ouvrage, les scientifiques rédigent des publications pour avancer des arguments ; les données sont les preuves qui étayent ces arguments. Reconnaissance, attribution et paternité vont de pair avec les publications, mais ne s'appliquent pas bien ni facilement aux données. Certaines publications sont riches en données, avec un propos minimaliste en guise de fil rouge. D'autres sont d'abord un argumentaire où les données ne sont mentionnées qu'en passant, voire pas du tout. Cependant, quelles que soient les proportions respectives des éléments probants et des arguments, les publications sont bien plus que des lots de données. Extraire ces dernières des premières pour en faire des marchandises indépendantes revient à leur ôter une bonne partie de leur signification. Les publications sont conçues comme des unités indépendantes, interprétables par un lectorat familier du domaine. Elles peuvent être un moyen de découvrir les données et inversement : les liens entre publications et données peuvent donc leur permettre de s'enrichir mutuellement. Cependant, à trop se focaliser sur ces relations, on risque de pétrifier une relation binaire entre les données et les publications. Conserver les données aux fins de la reproductibilité pourrait forcer les jeux de données à adopter une taille proportionnée aux publications, au lieu de les représenter d'une manière qui permette une exploitation plus générale.

L'accès ouvert aux publications scientifiques est viable pour deux raisons spécifiques à la recherche, comme l'explique Peter Suber (2012a, chapitre 3). Les chercheurs et chercheuses possèdent les droits sur leurs publications, au moins au départ, et sont motivés pour diffuser celles-ci le plus largement possible, car ils écrivent pour accroître leur influence plutôt que leurs revenus. Il n'en va pas de même pour les données, ce qui a des répercussions sur bien d'autres aspects de la gestion et des politiques en la matière. Les scientifiques accueillent à bras ouverts

les méthodes qui améliorent la découvrabilité de leurs publications, surtout si cela a pour effet d'augmenter les références à celles-ci.

Les modèles économiques de l'édition scientifique se sont métamorphosés au cours des dernières décennies. La plupart des étapes du processus de publication, de la proposition de l'article à la diffusion et à l'accès, se font maintenant en numérique. Des identifiants uniques et pérennes, comme les DOI, sont attribués au moment de la parution, ainsi que de manière rétroactive à des ressources plus anciennes.

Les référentiels assignent des DOI, des Handles et d'autres identifiants à des jeux de données. Les auteurs et autrices reçoivent, quant à eux, des identifiants uniques et pérennes comme ORCID (Open Researcher and Contributor ID, identifiant ouvert pour auteur et contributeur). Ces identifiants et d'autres peuvent être utilisés pour relier des objets numériques à des technologies telles que Crossref, Object Reuse and Exchange ou ResourceSync, comme nous l'avons vu au chapitre 9. Ce sont ces mêmes progrès technologiques qui ont contribué au mouvement pour l'accès ouvert et fait tomber les barrières pour créer une maison d'édition. Des maisons grandes et petites offrent de nouveaux services d'exploration de données, exploitant l'intégration des objets numériques pour fournir des rapports personnalisés sur les performances des revues, des universités, des départements universitaires, des référentiels, des organismes de financement et des chercheurs et chercheuses.

Les bibliothèques et les archives sont expertes dans la constitution et la gestion de collections. Les bibliothèques se consacrent surtout aux ressources publiées et les archives à des objets uniques. Les bibliothèques de recherche recueillent leur fonds dans l'ensemble des domaines de la connaissance et l'approfondissent en fonction du programme de leurs universités. Suivant les règles d'accréditation, les programmes doctoraux bénéficient de collections plus complètes que les programmes de premier cycle. Les fonds archivistiques sont généralement plus modestes et plus focalisés sur un domaine ou un type de ressources. Des collections riches dans un domaine particulier attirent scientifiques et population étudiante dans une université : elles constituent donc des atouts essentiels pour toute institution. Des chercheurs et chercheuses de nombreuses disciplines des sciences exactes et des technologies ne souhaitent utiliser que les services numériques de leurs bibliothèques universitaires. À l'inverse, dans de nombreux champs des sciences humaines, les chercheurs dépendent largement des ressources papier et archivistiques. Ils attendent de leur bibliothèque qu'elle pérennise des collections historiques pointues et qu'elle fournisse des espaces physiques pour les consulter. Chercheurs comme étudiants ont besoin à la fois de ressources physiques et numériques, de bâtiments pour les utiliser et de personnel pour les assister dans

la découverte et l'interprétation. Tous les utilisateurs et utilisatrices ont besoin de l'expertise des professionnelles et professionnels de l'information pour sélectionner, recueillir, organiser et mettre à disposition des ressources informationnelles, que leur travail leur soit visible ou non.

L'un des défis institutionnels à relever est le compromis entre la préservation et l'accès. Les formes les plus efficaces de sauvegarde, comme le stockage de copies papier sous une montagne, sont difficiles d'accès. À l'inverse, les formes d'accès les plus commodés, comme les images de pages en basse résolution, n'offrent pas une préservation correcte. Les institutions font souvent les deux, ce qui suppose un système double pour certains types de biens. La distinction est plus marquée encore pour les données de la recherche. Les fichiers numériques peuvent être préservés dans des archives fermées que l'on peut restaurer en cas de perte, mais ne permettent pas ou mal l'accès. Fournir un accès à ces fichiers d'une manière qui pérennise leur valeur pour la recherche requiert des systèmes interactifs en ligne, un dispositif technique adapté au domaine, de la puissance de calcul et un savoir-faire du personnel pour aider les usagers et usagères à exploiter ces ressources. Ces activités exigent une expertise dans le domaine et un investissement continu dans la conservation.

Même en mettant de côté les problèmes plus vastes de la préservation numérique, qui sont nombreux, la plupart des universités sont mieux placées pour gérer des archives fermées ou de petites collections de ressources spécialisées que pour assumer la responsabilité de la pérennisation de l'accès à de vastes collections disciplinaires. Celle-ci nécessite des investissements infrastructurels au niveau des communautés de recherche, des consortiums universitaires ou des États. Pour la plupart des domaines, l'agrégation de ressources en données est le moyen d'exploitation le plus viable.

D'autres distinctions entre publications et données sont pertinentes pour déterminer ce qu'il faut garder et pourquoi. L'une d'elles est que les publications peuvent exister en de nombreux exemplaires et dans de nombreux ensembles, mais qu'elles n'ont besoin d'être cataloguées qu'une seule fois. Les bibliothèques ont commencé à se répartir la charge du catalogage au début du ^{xx}e siècle avant de s'appuyer sur ces partenariats pour élaborer des services numériques partagés (Borgman, 2000). Les données de la recherche se rapprochent davantage des ressources archivistiques, car chaque jeu est unique et requiert ses propres métadonnées et registres de provenance. Un travail supplémentaire est nécessaire pour décrire des éléments uniques ou les fusionner dans des structures communes. Néanmoins, dans tous les cas, les collections gagnent en valeur à mesure qu'elles grandissent. Les bibliothèques

universitaires signent des accords consortiaux sur ce que chacune devra collecter, ce qui encourage la concentration des ressources et offre un accès aux membres de la communauté. Il est possible de faire de même avec les ensembles de données.

Les bibliothèques et les archives disposent de politiques d'acquisition complémentaires. Dans le jargon professionnel, les bibliothèques sélectionnent et les archives estiment. Toutes deux acquièrent des ressources qu'elles comptent garder indéfiniment. Les bibliothèques effectuent davantage de désherbage et de retraits des rayons à mesure que l'information se périme ou que de nouvelles éditions apparaissent. Décider quand se défaire d'un objet peut être un choix plus difficile que celui de l'acquérir. Pour ce qui est des ressources publiées, les bibliothèques passent des accords de dernier exemplaire (*last copy agreement*), où une institution s'engage à assurer la conservation d'un ouvrage de manière que les autres puissent éliminer leurs exemplaires. Cependant, ces accords ne sont pas applicables à des objets uniques tels que les données. Certaines peuvent cesser d'être utiles en quelques mois, d'autres en plusieurs décennies. La planification de la rétention est particulièrement problématique en matière de données. Pour fixer la date d'expiration, les scientifiques s'en remettent souvent aux bibliothécaires et vice versa, ce qui aboutit à une impasse.

On aura noté les différences d'investissement dans les publications et les données dans les études de cas. Tous les domaines de recherche cherchent à pérenniser l'accès aux publications. En revanche, la pérennisation des données est pour le moins inégale. En astronomie – qui dispose de l'infrastructure de la connaissance la plus complète parmi les études de cas –, les publications, la classification des objets et les données sont gérées séparément. La littérature disciplinaire est conservée dans l'Astrophysics Data System (ADS), tandis que le CDS et NED cataloguent des objets célestes mentionnés dans les publications. Ces trois institutions travaillent en étroite collaboration et créent des liens entre objets et publications. Cependant, les observations des missions spatiales sont mieux conservées que celles des missions au sol. Ce sont les professionnelles et professionnels de l'information de l'ADS, du CDS, de NED et des référentiels de données astronomiques qui établissent le plus de liens entre objets célestes, publications et données. Les bibliothèques astronomiques fournissent également des liens, qui ajoutent de la valeur à leurs propres fonds. Ce sont les auteurs et autrices de publications astronomiques qui lient le moins, puisqu'ils citent d'autres publications, mais rarement des données (Accomazzi et Dave, 2011 ; Pepe *et al.*, à paraître). En astronomie comme ailleurs, la capacité des infrastructures de la connaissance à lier solidement publications et données pour les rendre découvrables repose sur l'investissement dans des professionnels de l'information, qui effectuent le travail nécessaire.

Accès aux données

Fournir un accès aux données est plus difficile encore qu'offrir un accès ouvert aux publications, comme formulé dans la quatrième provocation :

Les travaux de recherche se diffusent plus largement grâce à des mouvements tels que l'édition en accès ouvert, l'ouverture des données et le logiciel libre. Les finalités différentes des données et des publications dans la recherche influent sur les incitations à diffuser, ainsi que sur les moyens et les pratiques de diffusion. L'ouverture de l'accès aux données a des répercussions encore mal comprises sur les personnels de recherche, les bibliothèques, les universités, les organismes de financement, les maisons d'édition et les autres acteurs.

La capacité à découvrir des données et à y accéder est plus importante dans les champs qui ont investi dans des réservoirs communs de ressources. Les domaines où les données sont d'emblée mises en commun ou agrégées sont ceux où existent des incitations à constituer ces ressources. L'astronomie, la biologie, la biomédecine et la métaomique pour les sciences exactes, la recherche par sondage dans les sciences sociales et les corpus textuels dans les sciences humaines en sont des exemples manifestes. Dans tous ces domaines, on peut comparer et combiner les données. Les chercheurs et chercheuses sont prêts à déposer les leurs en échange d'un accès au réservoir commun. Ces ressources n'en doivent pas moins être régies par une gouvernance. La pérennisation et les passagers clandestins représentent des défis permanents. Démarrer des ensembles de recherche est difficile ; mobiliser les institutions pour qu'elles en fassent des ensembles ressources ou de référence l'est plus encore. Tous ne survivent pas. Même les référentiels les plus solides prévoient des plans de relève qui précisent le traitement final de leurs données en cas d'arrêt des financements. Transformer les passagers clandestins en membres actifs nécessite une gestion ingénieuse.

La plupart des données de la recherche prennent la poussière sous forme de vieux fichiers avant d'être détruites ou abandonnées. Les raisons en sont nombreuses, comme nous l'avons illustré tout au long du présent ouvrage. Les chercheurs et chercheuses, si on les laisse à eux-mêmes, documenteront généralement leurs données juste assez pour répondre aux besoins de leurs projets de recherche immédiats ou prévus. Pour améliorer l'accès, davantage de données doivent être conservées sous une forme réutilisable, ce qui nécessite de changer les motivations extrinsèques. Pour la plupart des scientifiques, le problème fondamental est de mieux gérer leurs propres données. Elles et ils ont besoin d'outils, de services et d'assistance pour les

archiver de manière à pouvoir les réutiliser eux-mêmes, ce qui augmente les chances que leurs données puissent être utiles à d'autres ultérieurement.

Les régimes d'autorisation sont importants. Les scientifiques ont besoin de solutions de gestion qui leur permettent de garder la maîtrise de leurs ressources. De nombreux types de contrôles peuvent s'avérer nécessaires selon les travaux et les données, y compris l'observance de périodes d'embargo, la mise sous licence, des accords de collaboration et des réglementations sur les sujets humains. Si les chercheurs et chercheuses peuvent archiver leurs données dans des systèmes fiables et conformes, qui leur soient accessibles quand ils en ont besoin, ils peuvent accepter de les exposer publiquement plus tard. Une fois stockées dans un référentiel, les données peuvent être mises à la disposition de collaborateurs et collaboratrices ou du public en changeant les paramètres de permission. Des systèmes comme Dataverse et SciDrive fonctionnent sur le principe que l'archivage des données est une étape préliminaire à l'ouverture de l'accès (Crosas, 2011 ; Drago *et al.*, 2012 ; Goodman *et al.*, 2014 ; SciDrive, 2014). La documentation est parfois loin d'être optimale, mais ces démarches permettent de conserver des données qui seraient sinon perdues et augmentent leurs chances de devenir découvrables. Lorsque les données sont plus faciles à garder, la motivation à les rendre plus réutilisables peut augmenter. De même, des investissements dans la conservation peuvent améliorer la découvrabilité. En effet, les données stockées de manière fiable sont plus faciles à citer. Les publications pourraient rester le principal moyen de découverte des données, d'une part parce qu'elles les décrivent de manière plus complète et d'autre part parce que les scientifiques préfèrent la citation de publications à la citation de jeux de données.

Les données constituent à la fois une charge et un atout. Les garder coûte cher, mais les risques de mésusage, d'interprétation erronée et de responsabilité légale n'incitent pas à les diffuser. Dans une affaire récente, une demande officielle de divulgation a provoqué des mois de négociation, de longues consultations juridiques et plus de publicité que ce que les parties impliquées auraient souhaité. Des chercheurs et chercheuses en génie parasismique d'une grande université ont étudié des bâtiments en béton à Los Angeles qui risquaient de s'effondrer. Les publications subventionnées par la National Science Foundation expliquaient ce qu'ils avaient découvert, mais ne comprenaient pas de liste des bâtiments. Les autorités municipales de Los Angeles en ont demandé une afin d'évaluer la sécurité de ces constructions. Les chercheurs et la direction de l'université ont d'abord refusé au motif que les propriétaires des bâtiments concernés pourraient leur intenter un procès. L'université opérait en effet un *distinguo* entre les données recueillies à des fins scientifiques et l'évaluation sismique de chaque bâtiment. Finalement, les parties ont convenu de divulguer une liste expurgée et d'adopter des formulations juridiques sur les risques sismiques (Lin *et*

al., 2014a, 2014b ; Smith *et al.*, 2013 ; Xia *et al.*, 2013). Les inquiétudes des parties prenantes, quoique divergentes, étaient toutes légitimes.

L'utilisation de données de génie parasismique dans les politiques publiques n'est qu'un exemple d'usage imprévu des publications et des données. Une solution plus générale au problème de la réutilisation est d'autoriser l'analyse computationnelle, ou *mining* (fouille, exploration), de larges corpus de publications, de données et d'autres objets numériques. Plutôt que d'anticiper les questions risquant d'être posées, on pourrait laisser les futurs fouilleurs et fouilleuses écrire leurs propres algorithmes, éventuellement à l'aide d'API (Application Programming Interfaces). Cette approche a été proposée sous diverses formes (Bibliographic Services Task Force, 2005 ; Bourne, 2005 ; Bourne *et al.*, 2011 ; Shotton *et al.*, 2009). Sa faiblesse est la perte de contexte, la perte de provenance et la difficulté à maintenir les relations entre objets nécessaires à l'interprétation des résultats. L'ouverture des données au sens où elles seraient librement réutilisables (Open Data Commons, 2013) est une condition nécessaire à la recherche, mais non suffisante. Des données vraiment ouvertes au sens de la flexibilité, de la transparence, de la conformité au droit, de la protection de la propriété intellectuelle, de la responsabilité formelle, du professionnalisme, de l'interopérabilité, de la qualité, de la sécurité, de l'efficacité, de la responsabilité de rendre compte et de la pérennité, mettent la barre bien plus haut (Organisation for Economic Co-operation and Development, 2007).

Acteurs et talents

Les scientifiques s'attirent davantage de crédit pour la collecte ou la création de nouvelles données que pour l'exploitation de données existantes. Si le système de récompense se met, même lentement, à accorder davantage de valeur à la réutilisation, de nouvelles compétences et infrastructures seront nécessaires, comme présenté dans la cinquième provocation :

Les infrastructures de la connaissance évoluent afin de prendre en compte l'ouverture des données, leur usage intensif dans la recherche, les nouvelles technologies, les médias sociaux et les transformations des pratiques et des politiques. Certaines parties prenantes y gagnent tandis que d'autres y perdent. Coûts, bénéfices, risques et responsabilités sont redistribués. De nouvelles formes d'expertise sont nécessaires, mais leur application varie selon les contextes et les domaines de recherche.

Les réponses des parties prenantes s'opposent sur des questions fondamentales concernant les données à garder et pourquoi : Pourquoi conserver un jeu de données ? Quels sont les critères pour décider quelles données garder ? Qui décide

quelles données valent la peine d'être conservées ? Pour qui faut-il les garder ? Quelles fins et quels usages peuvent être prévus ? Sous quelles formes faut-il les conserver et avec quelles informations auxiliaires ? Combien de temps ? Qui les gardera ? Qui investira dans la pérennisation à court et long terme des données jugées intéressantes ? Qui y aura accès ? Quelles politiques d'accès, d'utilisation et de réutilisation faut-il prévoir ? Quels outils, technologies, installations et ressources humaines sont nécessaires pour maintenir l'utilité d'une ressource en données ?

Lorsque les données sont recueillies à la source ou dérivées de ressources disponibles, comme nous l'avons vu un peu plus haut, la responsabilité de leur gestion, de leur conservation et de leur diffusion incombe généralement au chercheur ou à la chercheuse. La gestion de données nécessite de combiner une expertise dans le domaine de recherche et un savoir-faire dans l'organisation et la conservation de l'information. Une expertise technique considérable peut aussi être requise selon les particularités des données, leur contexte et leurs usages. Être spécialiste d'un domaine de recherche ne fait pas nécessairement de quelqu'un une experte ou un expert de la gestion de données. Il est rare que ces compétences soient enseignées dans les programmes disciplinaires de second cycle.

On ignore jusqu'à quel point les chercheurs et chercheuses sont prêts à porter le fardeau de la gestion de données. Beaucoup, si ce n'est la majorité, voient le temps et les ressources consacrés à gérer leurs données comme autant d'énergie perdue pour leurs travaux de recherche. Ils peuvent préférer déléguer ces tâches aux personnels des bibliothèques et des archives, bien qu'il faille du temps pour développer ce genre de partenariats. Les bibliothèques peinent déjà à assurer leurs missions et ne considèrent pas toutes que la gestion des données en fait partie. Quant aux maisons d'édition, elles se chargeraient plus volontiers de l'indexation et de l'établissement de liens vers les référentiels que de la conservation des données.

Les programmes académiques en science des données traitent de l'analyse, la gestion, l'organisation, la conservation et l'accès de différentes façons. Une étude des National Academies états-uniennes sur la main-d'œuvre en préservation numérique montre combien il est difficile de définir l'éventail des compétences nécessaires (Hedstrom *et al.*, 2014). L'une des principales gageures en matière de personnel est que l'essentiel des activités de gestion de données représente un travail invisible. Les professionnelles et professionnels de l'information, les ingénieures et ingénieurs logiciels, les programmeurs et programmeuses scientifiques, les concepteurs et conceptrices d'instruments et autres expertes et experts techniques soutiennent les fondements de la recherche. Leurs compétences sont souvent dévalorisées et les débouchés sont peu clairs. Il est difficile

de recruter des talents avec des contrats « subventionnés », sans sécurité de l'emploi ni perspectives de promotion. Les investissements dans l'infrastructure humaine seront cruciaux pour réussir à conserver et exploiter les données de la recherche. Si elles veulent pérenniser leurs infrastructures de la connaissance, les communautés de recherche doivent embaucher ces professionnels et leur proposer des carrières.

La combinaison de savoir-faire nécessaires en gestion, conservation et préservation numérique des données variera selon les parties prenantes. Les universités, les bibliothèques et les laboratoires ont besoin de ces types d'expertise dans l'ensemble de leurs organisations. Les référentiels de données, qui sont eux-mêmes soutenus par des organismes de financement, des communautés de recherche ou d'autres institutions, ont besoin de personnel doté de compétences professionnelles dans la discipline, dans la gestion de données et dans la technologie. De nouveaux acteurs, publics comme privés, arrivent sur scène. L'Union européenne investit dans OpenAIRE (Open Access Infrastructure for Research in Europe), une infrastructure qui prend en charge publications, données et autres contenus dans ses référentiels. Zenodo, qui est établi au CERN, est une composante d'OpenAIRE qui accepte les objets de recherche absents des autres référentiels. L'Australie a inscrit la gestion des données dans son code déontologique pour la recherche et a commencé dans ce but une infrastructure nationale qui comprend des référentiels, du personnel et différents partenariats avec des institutions (Australian National Data Service, 2014 ; National Health and Medical Research Council, 2007 ; Open Access Infrastructure for Research in Europe, 2014 ; Schirrwagen *et al.*, 2013 ; ZENODO, 2013).

Les référentiels institutionnels des universités et les référentiels en accès ouvert tels qu'arXiv et SSRN continuent de se concentrer sur les prépublications, les rééditions et la littérature grise. Certains acceptent les jeux de données en complément de documents textuels. D'autres, comme Dataverse, sont expressément conçus pour eux (ArXiv.org, 2013 ; Crosas, 2011 ; G. King, 2013 ; Social Science Research Network, 2014). Des sociétés commerciales comme SlideShare et FigShare prennent en charge un vaste éventail d'objets de recherche, leur attribuent des DOI et les rendent plus découvrables. Thomson Reuters a lancé le Data Citation Index pour consigner les jeux de données se trouvant dans des référentiels, mais n'héberge pas de données. Ce ne sont là que quelques-uns des nombreux acteurs qui proposent des services à valeur ajoutée pour les objets de recherche. Certains ont des objectifs à court terme, tandis que d'autres s'efforcent de pérenniser l'accès des ressources.

Passé, présent et avenir des infrastructures de la connaissance

Bâtir des infrastructures de la connaissance s'apparente au problème de l'œuf et de la poule, comme présenté dans la sixième et dernière provocation :

Les infrastructures de la connaissance se développent et s'adaptent au fil des générations de savantes et savants. Elles ont besoin d'une vision à long terme pour leur conception et leurs politiques ; or le financement de la recherche se fait sur des cycles courts. Des investissements substantiels dans l'infrastructure sont nécessaires afin d'acquérir, de pérenniser et d'exploiter les données de la recherche aujourd'hui et demain. Ces investissements seront controversés, car les choix faits aujourd'hui décideront de quelles données et autres ressources informationnelles nous disposerons demain et bien après.

Les tensions sont nombreuses, en particulier dans « le long maintenant de l'infrastructure technologique » (Ribes et Finholt, 2009). Les individus, les projets et les organisations fonctionnent sur des échelles de temps différentes et la contradiction dans les buts peut ne pas apparaître immédiatement. Le financement de la recherche fonctionne généralement sur des cycles de cinq ans ou moins. De nombreuses subventions ne sont attribuées que pour un ou deux ans. Même les référentiels de données sont financés par des subventions de recherche qui doivent être régulièrement renouvelées. Partir du principe que « si nous le mettons en place, les usagers et usagères viendront » est une stratégie risquée. La subvention peut être épuisée à peine le prototype terminé. Lorsque le financement cesse, les collaborations peuvent se dissoudre sans heurts, en emportant leurs ressources et leur expertise pour les projets suivants. Cependant, il arrive aussi que les spécialistes soient laissés pour compte, que les données se retrouvent à l'abandon, que les technologies soient dispersées et que des éléments critiques de l'infrastructure ne soient plus supervisés. On ignore encore trop de choses sur le début, le milieu et la fin des infrastructures de la connaissance (Cummings *et al.*, 2008 ; Duderstadt *et al.*, 2002 ; Edwards *et al.*, 2007, 2011, 2013 ; Lee *et al.*, 2006 ; Olson *et al.*, 2008 ; Ribes et Jackson, 2013).

Financer une infrastructure de recherche n'est pas la même chose que de financer des travaux de recherche. Il peut en effet être difficile de persuader les gouvernements et les organismes de financement d'investir dans les technologies, les personnes et les services, comme l'ont appris à leurs dépens le programme Cyberinfrastructure aux États-Unis, le programme eScience au Royaume-Uni, le National Data Service en Australie et d'autres projets ailleurs. Bien que les détails soient très variables, beaucoup de scientifiques approuvent ces investissements, car elles et ils se rendent compte

que les technologies partagées et les réservoirs communs de ressources excèdent les capacités des universités et des chercheurs et chercheuses. D'autres ne les soutiennent pas et combattent des investissements qui leur apparaissent comme des solutions descendantes et centralisées. Ces initiatives s'inscrivent sur le long terme, mais doivent faire avec des cycles de financements courts et le renouvellement politique. Elles sont composées de nombreuses composantes mouvantes, qui évoluent chacune à leur propre rythme : les technologies progressent vite, alors que les universités et les maisons d'édition évoluent bien plus lentement. Les *start-ups* peuvent être plus réactives, mais n'investissent pas nécessairement pour le long terme.

La capacité à conserver et exploiter les données dépend aussi d'investissements dans la gestion de données. Les chercheurs et chercheuses semblent plus motivés à prendre soin de leurs données lorsque ces investissements apportent une valeur ajoutée à la recherche que lorsque le but est de vérifier leur travail. Les scientifiques et les autres acteurs prêts à investir dans la gestion peuvent décider conjointement quelles données valent la peine d'être gardées. S'il semble que leur valeur chute rapidement, il y a moins de raisons de les conserver. Maintenir les données anciennes, le matériel et les logiciels est difficile et coûteux. Certaines données ne peuvent être récupérées qu'en effectuant plusieurs couches d'émulation ; or ce genre d'opération se justifie difficilement (Brooks, 1975 ; Jackson et Buyuktur, 2014 ; Lee *et al.*, 2006 ; Mayernik, 2015 ; Segal, 2005 ; Winkelman et Rots, 2012a).

Les données qui ne présentent une utilité qu'à court terme peuvent être déposées ou publiées sur des sites web afin de les diffuser et de les rendre accessibles. En revanche, les objets de recherche qui possèdent une valeur à long terme devraient être confiés à des institutions stables, comme des bibliothèques, des archives, des musées ou des référentiels dotés de fonds suffisants. Certains chercheurs et chercheuses s'adaptent rapidement aux nouvelles technologies. D'autres se montrent prudents et n'adoptent que celles dont le support semble sûr. De même, certains entreprendront leurs données dans des logiciels propriétaires et d'autres uniquement dans des formats ouverts. Les silos sont partout et l'interopérabilité reste un but lointain.

Conclusion

Il n'existe pas de réponse toute faite à la question « quelles données garder ? », parce qu'il n'y a pas non plus de réponse toute faite à la question « qu'est-ce qu'une donnée ? ». Malgré leurs désaccords, la majorité des scientifiques aimeraient disposer de meilleurs moyens de gérer ce qu'elles et ils considèrent comme leurs données.

Une amélioration de la gestion mènerait probablement à des données plus pérennes et donc à de meilleurs moyens de découvrir et partager les données. Cela suppose des investissements considérables, qui ne peuvent reposer sur les seules épaules des chercheurs et chercheuses. L'amélioration de l'accès aux données requiert que les communautés de recherche, les organismes de financement, les universités, les maisons d'édition et les autres parties prenantes investissent dans les infrastructures de la connaissance. La technologie, les politiques et les pratiques s'entremêlent sur de nombreux points. Unifier les nombreuses composantes mouvantes des infrastructures de la connaissance nécessite d'investir dans les personnes qui les font tenir ensemble par leur travail invisible.

Aujourd'hui, les scientifiques peuvent recueillir les données, les découvrir, les chercher, les analyser et les diffuser à des échelles jusque-là impossibles. Certaines données valent la peine d'être gardées pour toujours ; d'autres n'ont qu'une valeur éphémère. Pour certaines, il est plus facile de les recréer au besoin que de les conserver. Au cours de l'histoire humaine, tout garder n'a jamais été une possibilité. On ne peut prévoir parfaitement les futurs usages d'une information. La capacité à identifier de nouveaux éléments probants au milieu d'informations anciennes représente l'essence même de nombreuses formes de recherche. La pérennisation des données est un objectif autrement plus ambitieux que leur simple stockage et sauvegarde. Le défi consiste à rendre les données découvrables, utilisables, évaluables, intelligibles et interprétables, le tout pour de longues périodes. Les parties prenantes peuvent ne pas s'accorder sur les types de données au sein d'un domaine qui méritent ce niveau d'investissement, ou sur les publics qui leur trouveraient un intérêt. Le plus difficile est de déterminer qui serait prêt à effectuer ces investissements dans l'intérêt d'autres parties. Faire l'article de l'accès aux données revient à faire l'article des infrastructures de la connaissance. Une vision à très long terme est nécessaire, mais y parvenir est difficile en raison du réseau complexe des parties prenantes dans l'ensemble des domaines de recherche, des communautés et des pays. Pour reprendre l'hypothèse du présent ouvrage, la valeur des données réside dans leur usage. À moins que les parties prenantes ne parviennent à s'accorder sur ce qu'il faut garder et pourquoi et à investir le travail invisible nécessaire à la pérennité des infrastructures de la connaissance, les *big data* et les *little data* deviendront bien vite *zéro data*.

Bibliographie

- Aad, G., T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. A. Abdelalim, O. Abidinov *et al.* 2012. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC". *Physics Letters [Part B]*, 716 (1), p. 1-29. <https://www.sciencedirect.com/science/article/pii/S037026931200857X?via=ihub>.
- Abbate, Janet. 1999. *Inventing the Internet*. Cambridge, MA : MIT Press.
- Accomazzi, Alberto. 2010. "Astronomy 3.0 Style". *Astronomical Society of the Pacific Conference Series*, 433, p. 273-281.
- Accomazzi, Alberto et Rahul Dave. 2011. "Semantic Interlinking of Resources in the Virtual Observatory Era". *Astronomical Society of the Pacific Conference Series*, 442, p. 415-424. <https://arxiv.org/pdf/1103.5958.pdf>.
- Acropolis Museum. 2013. "The Frieze". <http://www.theacropolismuseum.gr/en/content/frieze-0>.
- Agosti, Maristella et Nicola Ferro. 2007. "A Formal Model of Annotations of Digital Content". *ACM Transactions on Information Systems*, 26 (1). <https://dl.acm.org/doi/10.1145/1292591.1292594>.
- Agre, Philip E. 1994. "From High Tech to Human Tech: Empowerment, Measurement, and Social Studies of Computing". *Computer Supported Cooperative Work*, 3 (2), p. 167-195. <https://link.springer.com/article/10.1007/BF00773446>.
- Ahn, Christopher P., Rachael Alexandroff, Carlos Allende Prieto, Scott F. Anderson, Timothy Anderton, Brett H. Andrews, Éric Aubourg *et al.* 2012. "The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey". *Astrophysical Journal*, 203, p. 21. <https://iopscience.iop.org/article/10.1088/0067-0049/203/2/21>.
- Akyildiz, I. F., W. Su, Y. Sankarasubramaniam et E. Cayirci. 2002. "Wireless Sensor Networks: A Survey". *Computer Networks*, 38 (4), p. 393-422. <https://www.sciencedirect.com/science/article/abs/pii/S1389128601003024>.
- Alabaster, Jay. 2013. "Library of Congress Saves 500 Million Tweets Per Day in Archives". *Computerworld*. http://www.computerworld.com/s/article/9235421/Library_of_Congress_saves_500_million_tweets_per_day_in_archives.
- Alarcón, Enrique. 2000. "Corpus Thomisticum". <http://www.corpusthomisticum.org>.
- Alberts, Bruce. 2012. "The End of 'Small Science'?". *Science*, 337 (6102), p. 1583. <https://science.sciencemag.org/content/337/6102/1583.summary>.
- Alberts, Bruce. 2013. "Impact Factor Distortions". *Science*, 340 (6134), p. 787. <https://science.sciencemag.org/content/340/6134>.
- Allen, Erin. 2013. "Update on the Twitter Archive at the Library of Congress". <https://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/>.

- Allon, Mark. 2007. "Recent Discoveries of Buddhist Manuscripts from Afghanistan and Pakistan: The Heritage of the Greeks in the North-west", in *Memory as History: The Legacy of Alexander in Asia*, Himanshu Prabha Ray et Daniel T. Potts (éd.). New Dehli : Aryan Books International.
- Allon, Mark. 2009. "Recent Discoveries of Buddhist Manuscripts from Afghanistan and Pakistan and Their Significance", in *Art, Architecture and Religion along the Silk Roads*, Ken Parry (éd.), p. 133-178. Turnhout : Brepols.
- Alsheikh-Ali, Alawi A., Waqas Qureshi, Mouaz H. Al-Mallah et John P. A. Ioannidis. 2011. "Public Availability of Published Research Data in High-Impact Journals". *PLoS ONE*, 6 (9). <https://doi.org/10.1371/journal.pone.0024357>.
- Altman, Micah et Gary King. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data". *D-Lib Magazine*, 13 (3/4). <http://www.dlib.org/dlib/march07/altman/03altman.html>.
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers et David J. Lipman, 1990. "Basic Local Alignment Search Tool". *Journal of Molecular Biology*, 215 (3), p. 403-410. <https://www.sciencedirect.com/science/article/pii/S0022283605803602>.
- American Psychological Association. 2009. *Publication Manual of the American Psychological Association*. 6^e éd. Washington, DC : APA.
- Ancient Lives. 2013. <http://ancientlives.org>.
- Anderson, Chris. 2004. "The Long Tail". *Wired*. <https://www.wired.com/2004/10/tail/>.
- Anderson, Chris. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More*. New York : Hyperion.
- Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired*. <https://www.wired.com/2008/06/pb-theory/>.
- Anderson, David P., Jeff Cobb, Eric Korpela, Matt Lebofsky et Dan Werthimer. 2002. "SETI@home: An Experiment in Public-Resource Computing". *Communications of the ACM*, 45 (11), p. 56-61. <https://cacm.acm.org/magazines/2002/11/6966-setihome/fulltext>.
- Anderson, Robert J. 1994. "Representations and Requirements: The Value of Ethnography in System Design". *Human-Computer Interaction*, 9 (3), p. 151-182. <https://www.ics.uci.edu/~corps/phaseii/Anderson-RepresentationRequirements-HCI.pdf>.
- Archaeology Data Service. 2013. <http://archaeologydataservice.ac.uk>.
- Arms, William Y. 2002. "What Are the Alternatives to Peer Review? Quality Control in Scholarly Publishing on the Web". *Journal of Electronic Publishing*, 8 (1). <https://quod.lib.umich.edu/j/jep/3336451.0008.103?view=text;rgn=main>.
- Aronova, Elena, Karen S. Baker et Naomi Oreskes. 2010. "Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long-Term Ecological Research (LTER) Network, 1957–Present". *Historical Studies in the Natural Sciences*, 40, p. 183-224. <https://online.ucpress.edu/hsns/article-abstract/40/2/183/105628/Big-Science-and-Big-Data-in-Biology-From-the?redirectedFrom=fulltext>.
- Arts & Humanities Research Council. 2012. "Technical Plan". <http://www.ahrc.ac.uk/Funding-Opportunities/Research-funding/RFG/Application-guidance/Pages/Technical-Plan.aspx>.
- Arts and Humanities Data Service. 2008. <http://www.ahds.ac.uk>.

- ArXiv.org. 2013. "ArXiv.org e-Print Archive". <http://arxiv.org>.
- Arzberger, Peter, Peter Schroeder, Anne Beaulieu, Geoffrey C. Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul F. Uhler et Paul Wouters. 2004. "An International Framework to Promote Access to Data". *Science*, 303, p. 1777-1778.
- Aspesi, Claudio et Amy Brand. 2020. "In pursuit of open science, open access is not enough". *Science*, 368 (6491), p. 574-577. <https://doi.org/10.1126/science.aba3763>.
- Association for Computing Machinery. 1992. "ACM Code of Ethics and Professional Conduct". <https://www.acm.org/code-of-ethics>.
- Association of Internet Researchers. 2012. "Ethics Guide". <https://aoir.org>.
- Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, Paul Messina, David G. Messerschmitt, Jeremiah P. Ostriker et Margaret H. Wright. 2003. "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure". Washington, DC : National Science Foundation. <https://www.nsf.gov/cise/sci/reports/atkins.pdf>.
- Australian Law Reform Commission. 2014. "Defining 'Non-Consumptive' Use". <https://www.alrc.gov.au/publication/copyright-and-the-digital-economy-dp-79/4-the-case-for-fair-use-in-australia/alrcs-proposals-for-reform/>.
- Australian National Data Service. 2014. <http://www.and.s.org.au>.
- Ayers, Edward L. 2003. *In the Presence of Mine Enemies: The Civil War in the Heart of America, 1859-1863*. New York : Norton.
- Ayers, Edward L. 2007. "The Valley of the Shadow: Two Communities in the American Civil War". <http://valley.lib.virginia.edu>.
- Ayers, Edward L. et Charles M. Grisham. 2003. "Why IT Has Not Paid Off as We Hoped (Yet)". *EDUCAUSE Review*, 38 (6), p. 40-51. <https://er.educause.edu/articles/2003/1/why-it-has-not-paid-off-as-we-hoped-yet>.
- Babbie, Earl. 2013. *The Practice of Social Research*. 13^e éd. Belmont, CA : Wadsworth.
- Baca, Murtha. 1998. *Introduction to Metadata: Pathways to Digital Information*. Los Angeles : Getty Information Institute.
- Baca, Murtha. 2002. *Introduction to Art Image Access: Issues, Tools, Standards, Strategies*. Los Angeles : Getty Publications.
- Bakshy, Eytan, Jake M. Hofman, Winter A. Mason et Duncan J. Watts. 2011. "Everyone's an Influencer: Quantifying Influence on Twitter", in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, p. 65-74. New York : ACM. <https://dl.acm.org/doi/10.1145/1935826.1935845>.
- Ball, Alex. 2012. "How to License Research Data". Digital Curation Centre. <http://www.dcc.ac.uk/resources/how-guides/license-research-data>.
- Ball, James. 2013. "Verizon Court Order: Telephone Call Metadata and What it Can Show". *The Guardian*. <http://www.guardian.co.uk/world/2013/jun/06/phone-call-metadata-information-authorities>.

- Bamford, Connor. 2012. "Solving Irreproducible Science". *The Scientist*. <https://www.the-scientist.com/news-opinion/solving-irreproducible-science-40440>.
- Barbier, Geoffrey, Zhuo Feng, Pritam Gundecha et Huan Liu. 2013. "Provenance Data in Social Media". *Synthesis Lectures on Data Mining and Knowledge Discovery*, 4 (1), p. 1-84. <https://www.morganclaypool.com/doi/abs/10.2200/S00496ED1V01Y201304DMK007>.
- Bard, Jonathan. 2013. *The Living World. Unpublished manuscript*. University of Oxford.
- Barrett, Timothy H. 2008. *The Woman Who Discovered Printing*. New Haven, CT : Yale University Press.
- Basken, Paul. 2012. "NIH to Begin Enforcing Open-Access Policy on Research It Supports". *Chronicle of Higher Education*. <http://chronicle.com/article/NIH-to-Begin-Enforcing/135852/?cid=at>.
- Batalin, Maxim A., Mohammad Rahimi, Yan Yu, Duo Liu, Aman Kansal, Gaurav S. Sukhatme, William J. Kaiser *et al.* 2004. "Call and Response: Experiments in Sampling the Environment", in *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems*, p. 25-38. New York : ACM. http://cres.usc.edu/pubdb_html/files_upload/420.pdf.
- Beaulieu, Anne, Sarah de Rijcke et Bas van Heur. 2012. "Authority and Expertise in New Sites of Knowledge Production", in *Virtual Knowledge: Experimenting in the Humanities and Social Sciences*, Paul Wouters, Anne Beaulieu, Andrea Scharnhorst, et Sally Wyatt (éd.), p. 25-56. Cambridge, MA : MIT Press. <http://mitpress.mit.edu/books/virtual-knowledge-0>.
- Beaumont, Christopher N., Stella S. R. Offner, Rahul Shetty, Simon C. O. Glover et Alyssa A. Goodman. 2013. "Quantifying Observational Projection Effects Using Molecular Cloud Simulations". *Astrophysical Journal*, 777 (2), p. 173. <https://arxiv.org/abs/1310.1929>.
- Becher, Tony. 1989. *Academic Tribes and Territories: Intellectual Enquiry and the Culture of Disciplines*. Buckingham, UK : SRHE – Open University Press.
- Becher, Tony. 1994. "The Significance of Disciplinary Differences". *Studies in Higher Education*, 19 (2), p. 151-161. <https://www.tandfonline.com/doi/abs/10.1080/03075079412331382007>.
- Bechhofer, Sean, David De Roure, Matthew Gamble, Carole Goble et Iain Buchan. 2010. "Research Objects: Towards Exchange and Reuse of Digital Knowledge". *Nature Precedings*. <https://www.nature.com/articles/npre.2010.4626.1>.
- Bechhofer, Sean, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch *et al.* 2013. "Why Linked Data Is Not Enough for Scientists". *Future Generation Computer Systems*, 29 (2), p. 599-611. <https://dl.acm.org/doi/10.1016/j.future.2011.08.004>.
- Becker, Glenn, Arnold Rots, Sherry L. Winkelman, Michael McCollough, Aaron Watry et Joan Hagler. 2009. "It's Not Just for Data Anymore: The Many Faces of the Chandra Data Archive", in *Chandra's First Decade of Discovery. Proceedings of the Conference Held 22-25 September, 2009 in Boston, MA*, p. 65. <http://adsabs.harvard.edu/abs/2009cfdd.confE.65B>.
- Bell, Gordon, Tony Hey et Alex Szalay. 2009. "Beyond the Data Deluge". *Science*, 323 (5919), p. 1297-1298. <https://science.sciencemag.org/content/323/5919/1297>.
- Benkler, Yochai. 2007. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT : Yale University Press.
- Berlekamp, Elwyn. 2012. "Small Science: Radical Innovation". *Science*, 338 (6109), p. 882. <https://science.sciencemag.org/content/338/6109/882.1>.

- Berman, Francine. 2008. "Got Data? A Guide to Data Preservation in the Information Age". *Communications of the ACM*, 51, p. 50-56.
- Berman, Francine, Brian Lavoie, Paul Ayris, G. Sayeed Choudhury, Elizabeth Cohen, Paul Courant, Lee Dirks *et al.* 2010. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information". San Diego : National Science Foundation, Andrew W. Mellon Foundation, Library of Congress, Joint Information Systems Committee of the UK, Council on Library and Information Resources, National Archives and Records Administration. <http://brtf.sdsc.edu/publications.html>.
- Bhattacharjee, Yudhijit. 2012. "Pharma Firms Push for Sharing of Cancer Trial Data". *Science*, 338 (6103), p. 29. <https://science.sciencemag.org/content/338/6103/29.summary>.
- Bibliographic Services Task Force. 2005. "Rethinking How We Provide Bibliographic Services for the University of California". University of California Libraries. <https://libraries.universityofcalifornia.edu/groups/files/sopag/docs/Final.pdf>.
- Biemiller, Lawrence. 2013. "Universities Must Encourage Researchers to Share Data, Panel Says". *Chronicle of Higher Education*. <http://chronicle.com/blogs/wiredcampus/universities-must-encourage-researchers-to-share-data-panel-says/45409>.
- Bietz, Matthew J. et Charlotte P. Lee. 2009. "Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work", in *ECSCW 2009*, Ina Wagner, Hilda Tellio lu, Ellen Balka, Carla Simone et Luigina Ciolfi (éd.), p. 243-262. Londres : Springer. http://link.springer.com/chapter/10.1007/978-1-84882-854-4_15.
- Bietz, Matthew J. et Charlotte P. Lee. 2012. "Adapting Cyberinfrastructure to New Science: Tensions and Strategies", in *Proceedings of the 2012 iConference*, p. 183-190. New York : ACM. <https://dl.acm.org/doi/10.1145/2132176.2132200>.
- Biglan, Anthony. 1973. "The Characteristics of Subject Matter in Different Academic Areas". *Journal of Applied Psychology*, 57 (3), p. 195-203. https://www.researchgate.net/profile/Anthony_Biglan/publication/232441743_The_Characteristics_of_Subject_Matter_in_Different_Academic_Areas/links/552e7f700cf2acd38cb9f8a0/The-Characteristics-of-Subject-Matter-in-Different-Academic-Areas.pdf.
- Bijker, Wiebe E. 1995. *Of Bicycles, Bakelites and Bulbs: Toward a Theory of Sociotechnical Change*. Cambridge, MA : MIT Press.
- Bijker, Wiebe E., Thomas P. Hughes et Trevor Pinch. 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA : MIT Press.
- Binding, Ceri, Keith May et Douglas Tudhope. 2008. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM", in *Research and Advanced Technology for Digital Libraries*, Birte Christensen-Dalsgaard, Donatella Castelli, Bolette Ammitzbøll Jurik et Joan Lippincott (éd.), p. 280-290. Berlin : Springer.
- Birnholtz, Jeremy P. et Matthew J. Bietz. 2003. "Data at Work: Supporting Sharing in Science and Engineering", in *Proceedings of the 2003 International ACM SIGGROUP Conference*, p. 339-348. New York : ACM.

- Bishop, Ann Peterson et Susan Leigh Star. 1996. "Social Informatics of Digital Library Use and Infrastructure". *Annual Review of Information Science & Technology*, 31, p. 301-401.
- Blair, Ann M. 2010. *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven, CT : Yale University Press.¹
- Blecksmith, Emily, Stéphane Paltani, Arnold Rots et Sherry L. Winkelman. 2003. "Chandra Data Archive Download and Usage Database", in *ASP Conference Series*, 295, p. 283.
- Blocker, Alexander W. et Xiao-Li Meng. 2013. "The Potential and Perils of Preprocessing: Building New Foundations". *Bernoulli*, 19 (4), p. 1176-1211. https://projecteuclid.org/download/pdf-view_1/euclid.bj/1377612848.
- Blomberg, Jeanette et Helena Karasti. 2013. "Reflections on 25 Years of Ethnography in CSCW". *Computer Supported Cooperative Work*, 22 (4), p. 373-423. https://www.researchgate.net/publication/257387341_Reflections_on_25_Years_of_Ethnography_in_CSCW.
- Bloom, Harold. 1994. *The Western Canon: The Books and School of the Ages*. New York : Harcourt Brace.
- Bohannon, John. 2013a. "Dance Your PhD: And the Winner Is...". *Science*. <http://news.sciencemag.org/scientific-community/2013/11/dance-your-ph.d.-and-winner-...>
- Bohannon, John. 2013b. "Genealogy Databases Enable Naming of Anonymous DNA Donors". *Science*, 339 (6117), p. 262. http://www.johnbohannon.org/NewFiles/DNA_privacy.pdf.
- Bollen, Johan, Herbert Van de Sompel, Aric Hagberg et Ryan Chute. 2009. "A Principal Component Analysis of 39 Scientific Impact Measures". *PLoS ONE*, 4 (6). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006022>.
- Bollen, Johan, Huina Mao et Xiao-Jun Zeng. 2010. "Twitter Mood Predicts the Stock Market". *Journal of Computational Science*, 2 (1). <https://arxiv.org/abs/1010.3003>.
- Bollier, David. 2007. "The Growth of the Commons Paradigm", in *Understanding Knowledge as a Commons: From Theory to Practice*, Charlotte Hess et Elinor Ostrom (éd.), p. 27-40. Cambridge, MA : MIT Press.
- Borges, Jorge Luis. 1999. "The Analytical Language of John Wilkins" [Lilia Graciela Vazquez (trad.)]. *ALAMUT*. <http://www.alamut.com/subj/artifice/language/johnWilkins.html>.²
- Borgman, Christine L. 1990. *Scholarly Communication and Bibliometrics*. Newbury Park, CA : Sage.
- Borgman, Christine L. 2000. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA : MIT Press.
- Borgman, Christine L. 2003. "The Invisible Library: Paradox of the Global Information Infrastructure". *Library Trends*, 51, p. 652-674.
- Borgman, Christine L. 2006. "What Can Studies of e-Learning Teach Us About e-Research? Some Findings from Digital Library Research". *Computer Supported Cooperative Work*, 15 (4), p. 359-383. <https://link.springer.com/article/10.1007%2Fs10606-006-9024-1>.

1. Blair, Ann. 2010. *Tant de choses à savoir. Comment maîtriser l'information à l'époque moderne*. R. Chartier (préface), B. Krespine (traduction). Paris : Seuil, (L'Univers historique).

2. Borges, Jorge Luis. 1957. *Oeuvres complètes*, t. I, P. Bénichou, S. Bénichou-Roubaud J.-P. Bernès et R. Caillois (traduction). Paris : Gallimard (La Pléiade), p. 749.

- Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA : MIT Press.
- Borgman, Christine L. 2009. "The Digital Future Is Now: A Call to Action for the Humanities". *Digital Humanities Quarterly*, 3. <http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html>.
- Borgman, Christine L. 2011. "Is Data to Knowledge as the Wasp Is to the Fig Tree? Reconsidering Licklider's Intergalactic Network in the Days of Data Deluge", in *Accelerating Discovery: Human-computer Symbiosis 50 Years On*, Park City, UT : Argonne National Labs, ICIS. <https://sites.google.com/site/licklider50>.
- Borgman, Christine L. 2012a. "The Conundrum of Sharing Research Data". *Journal of the American Society for Information Science and Technology*, 63 (6), p. 1059-1078.
- Borgman, Christine L. 2012b. "Why Are the Attribution and Citation of Scientific Data Important?", in *For Attribution – Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*, p. 1-10. Washington, DC : National Academies Press. http://www.nap.edu/catalog.php?record_id=13564.
- Borgman, Christine L. 2013. "Keynote Presentation: 'ADS, Astronomy and Scholarly Infrastructure'", communication présentée au Astrophysics Data System 20th Anniversary Symposium. Harvard-Smithsonian Center for Astrophysics, Cambridge, MA. <http://conf.adsabs.harvard.edu/ADSXX/>.
- Borgman, Christine L. 2019. "The lives and after lives of data". *Harvard Data Science Review*, 1 (1). <https://doi.org/10.1162/99608f92.9a36bdb6>
- Borgman, Christine L., Dineh Moghdam et Patti K. Corbett. 1984. *Effective Online Searching: A Basic Text*. New York : Marcel Dekker.
- Borgman, Christine L. et Susan L. Siegfried. 1992. "Getty's SynonameTM and Its Cousins: A Survey of Applications of Personal Name Matching Algorithms". *Journal of the American Society for Information Science American Society for Information Science*, 43 (7), p. 459-476. [https://doi.org/10.1002/\(SICI\)1097-4571\(199208\)43:7<459::AID-ASI1>3.0.CO;2-D](https://doi.org/10.1002/(SICI)1097-4571(199208)43:7<459::AID-ASI1>3.0.CO;2-D).
- Borgman, Christine L., Andrea L. Gallagher, Sandra G. Hirsh et Virginia A. Walter. 1995. "Children's Searching Behavior on Browsing and Keyword Online Catalogs: The Science Library Catalog Project". *Journal of the American Society for Information Science American Society for Information Science*, 46, p. 663-684. [https://doi.org/10.1002/\(SICI\)1097-4571\(199510\)46:9<663::AID-ASI4>3.0.CO;2-2](https://doi.org/10.1002/(SICI)1097-4571(199510)46:9<663::AID-ASI4>3.0.CO;2-2).
- Borgman, Christine L., Anne J. Gilliland-Swetland, Gregory H. Leazer, Richard Mayer, David Gwynn, Rich Gazan et Patricia Mautone. 2000. "Evaluating Digital Libraries for Teaching and Learning in Undergraduate Education: A Case Study of the Alexandria Digital Earth ProtoType (ADEPT)". *Library Trends*, 49 (2), p. 228-250.
- Borgman, Christine L. et Jonathan Furner. 2002. "Scholarly Communication and Bibliometrics". *Annual Review of Information Science & Technology*, 36, p. 3-72.

- Borgman, Christine L., Jillian C. Wallis et Noel Enyedy. 2006. "Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology", in *10th European Conference on Digital Libraries*. "Lecture Notes in Computer Science 4172", p. 170-183. Berlin : Springer.
- Borgman, Christine L., Jillian C. Wallis, Matthew S. Mayernik et Alberto Pepe. 2007a. "Drowning in Data: Digital Library Architecture to Support Scientific Use of Embedded Sensor Networks", in *Joint Conference on Digital Libraries*, p. 269-277. New York : ACM. <http://doi.acm.org/10.1145/1255175.1255228>.
- Borgman, Christine L., Jillian C. Wallis et Noel Enyedy. 2007b. "Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries". *International Journal on Digital Libraries*, 7, p. 17-30. <https://link.springer.com/article/10.1007/s00799-007-0022-9>.
- Borgman, Christine L., Jillian C. Wallis et Matthew S. Mayernik. 2012. "Who's Got the Data? Interdependencies in Science and Technology Collaborations". *Computer Supported Cooperative Work*, 21 (6), p. 485-523. <https://link.springer.com/article/10.1007/s10606-012-9169-z>.
- Borgman, Christine L., Sharon Traweek, Peter Darch, Milena Golshan, Elaine Levia, Camille Mathieu, Ashley E. Sands et Jillian C. Wallis. 2014. "Knowledge Infrastructures Lab". Los Angeles : University of California at Los Angeles, Department of Information Studies. <http://knowledgeinfrastructures.gseis.ucla.edu/index.html>.
- Borgman, Christine L., Peter T. Darch, Ashley E. Sands et Milena Golshan. 2016. "The durability and fragility of knowledge infrastructures: Lessons learned from astronomy", in *Proceedings of the Association for Information Science and Technology*, 53, p. 1-10. <http://dx.doi.org/10.1002/pra2.2016.14505301057>.
- Borgman, Christine L., Peter T. Darch, Irene V. Pasquetto et Morgan F. Wofford. 2020. *Our knowledge of knowledge infrastructures: Lessons learned and future directions*. University of California, Los Angeles. <http://escholarship.org/uc/item/9rm6b7d4>.
- Boruch, Robert F. 1985. "Definitions, Products, Distinctions in Data Sharing", in *Sharing Research Data*, Stephen E. Fienberg, Margaret E. Martin et Miron L. Straf (éd.), p. 89-122. Washington, DC : National Academies Press. http://books.nap.edu/catalog.php?record_id=2033.
- Bos, Nathan, Ann Zimmerman, Judith Olson, Jude Yew, Jason Yerkie, Erik Dahl et Gary M. Olson. 2007. "From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories". *Journal of Computer-Mediated Communication*, 12 (2). <http://jcmc.indiana.edu/vol12/issue2/bos.html>.
- Boulton, Geoffrey. 2012. "Open Your Minds and Share Your Results". *Nature*, 486 (7404), p. 441. <https://www.nature.com/news/open-your-minds-and-share-your-results-1.10895>.
- Boulton, Geoffrey, Michael Rawlins, Patrick Vallance et Mark Walport. 2011. "Science as a Public Enterprise: The Case for Open Data". *Lancet*, 377 (9778), p. 1633-1635. [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(11\)60647-8/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)60647-8/fulltext).
- Boulton, Geoffrey, Philip Campbell, Brian Collins, Peter Elias, Wendy Hall, Graeme Laurie, Onora O'Neill et al. 2012. "Science as an Open Enterprise". The Royal Society. <http://royalsociety.org/policy/projects/science-public-enterprise/report>.

- Bouquin, Daina R., D. A. Chivvis, E. Henneken, K. Lockhart, A. Muench et J. Koch, 2020. "Credit Lost: Two Decades of Software Citation in Astronomy". *The Astrophysical Journal Supplement Series*, 249 (1), p. 8. <https://doi.org/10.3847/1538-4365/ab7be6>.
- Bourne, Philip E. 2005. "Will a Biological Database Be Different from a Biological Journal?". *PLoS Computational Biology*, 1 (3). <https://doi.org/10.1371/journal.pcbi.0010034>.
- Bourne, Philip E., Timothy Clark, Robert Dale, Anita de Waard, Eduard H. Hovy et David Shotton (éd.). 2011. "Force 11 Manifesto: Improving Future Research Communication and e-Scholarship". http://www.force11.org/white_paper.
- Bowen, G. Michael et Wolff Michael Roth. 2007. "The Practice of Field Ecology: Insights for Science Education". *Research in Science Education*, 37, p. 171-187.
- Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. Cambridge, MA : MIT Press.
- Bowker, Geoffrey C. 2013. "Data Flakes: An Afterword to 'Raw Data' Is an Oxymoron", in *"Raw Data" Is an Oxymoron*, Lisa Gitelman (éd.), p. 167-171. Cambridge, MA : MIT Press.
- Bowker, Geoffrey C. et Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA : MIT Press.
- Bowker, Geoffrey C., Karen S. Baker, Florence Millerand, David Ribes, Jeremy Hunsinger, Lisbeth Klasturp et Matthew Allen. 2010. "Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment", in *International Handbook of Internet Research*, Jeremy Hunsinger, Lisbeth Klasturp et Matthew Allen (éd.), p. 97-117. Dordrecht : Springer. http://www.springer-link.com/index/10.1007/978-1-4020-9789-8_5.
- Boyd, Danah et Kate Crawford. 2012. "Critical Questions for Big Data". *Information, Communication & Society*, 15 (5), p. 662-679. <https://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878>.
- Boyle, James et Jennifer Jenkins. 2003. "The Genius of Intellectual Property and the Need for the Public Domain", in *The Role of Scientific and Technical Data and Information in the Public Domain*, Julie M. Esanu et Paul F. Uhler (éd.), p. 10-14. Washington, DC : National Academies Press.
- Boyle, John. 2013. "Biology Must Develop Its Own Big-data Systems". *Nature*, 499 (7456), p. 7. <https://go.gale.com/ps/anonymouse?id=GALE|A336383289&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=00280836&p=AONE&sw=w>.
- Boyle, Paul. 2013. "A U.K. View on the U.S. Attack on Social Sciences". *Science*, 341 (6147), p. 719. <https://science.sciencemag.org/content/341/6147/719>.
- Brady, Henry. 2004. "Testimony to the Commission on Cyberinfrastructure for the Humanities and Social Sciences". http://www.acls.org/cyberinfrastructure/cyber_meeting_notes_august.htm-brady_summary.
- Brase, Jan, Yvonne Socha, Sarah Callaghan, Christine L. Borgman, Paul F. Uhler et Bonnie Carroll. 2014. "Data Citation", in *Managing Research Data: Practical Strategies for Information Professionals*, Joyce M. Ray (éd.), p. 167-186. Lafayette, IN : Purdue University Press.
- [The] British Library Board. 2013. "Sacred Texts: Diamond Sutra". <http://www.bl.uk/onlinegallery/sacredtexts/diamondsutra.html>.

- British Museum. 2013. "Parthenon Sculptures". http://www.britishmuseum.org/about_us/news_and_press/statements/parthenon_sculptures.aspx.
- Brooks, Frederick. 1975. *The Mythical Man-Month: Essays on Software Engineering*. Reading, MA : Addison-Wesley.
- Brown, Ian et Christopher T. Marsden. 2013. *Regulating Code: Good Governance and Better Regulation in the Information Age*. Cambridge, MA : MIT Press.
- Brown, John Seely et Paul Duguid. 1996. "The Social Life of Documents". *First Monday*, 1. <http://firstmonday.org/ojs/index.php/fm/article/view/466/387>.
- Brown, John Seely et Paul Duguid. 2000. *The Social Life of Information*. Boston : Harvard Business School Press.
- Bruckman, Amy, Kurt Luther et Casey Fiesler. À paraître. "When Should We Use Real Names in Published Accounts of Internet Research?", in *Digital Research Confidential*, Eszter Hargittai et Christian Sandvig (éd.). Cambridge, MA : MIT Press.
- Brumfiel, G. 2002. "Misconduct Finding at Bell Labs Shakes Physics Community". *Nature*, 419, p. 419-421.
- Bruns, Axel et Yuxian Eugene Liang. 2012. "Tools and Methods for Capturing Twitter Data during Natural Disasters". *First Monday*, 17 (4). <https://journals.uic.edu/ojs/index.php/fm/article/view/3937>.
- Bryson, Bill. 2008. *Bryson's Dictionary for Writers and Editors*. New York : Broadway Books.
- Buckland, Michael K. 1991. "Information as Thing". *Journal of the American Society for Information Science American Society for Information Science*, 42, p. 351-360.
- Budapest Open Access Initiative. 2002. "Budapest Declaration on Open Access". <http://www.soros.org/openaccess/read.shtml>.
- Buneman, Peter, Sanjeev Khanna et Wang-Chiew Tan. 2000. "Data Provenance: Some Basic Issues", in *Foundations of Software Technology and Theoretical Computer Science*, vol. 1974, S. Kapoor et S. Prasad (éd.), p. 87-93. Berlin : Springer.
- Buneman, Peter, Sanjeev Khanna et Wang-Chiew Tan. 2001. "Why and Where: A Characterization of Data Provenance", in *Database Theory – ICDT 2001*, vol. 1973, Jan Van den Bussche et Victor Vianu (éd.), p. 316-330. Berlin : Springer.
- Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner et Jeffrey Schnapp. 2012. *Digital Humanities*. Cambridge, MA : MIT Press.
- Burke, Peter. 2000. *A Social History of Knowledge: From Gutenberg to Diderot*. Cambridge, UK : Polity Press.
- Burke, Peter. 2012. *A Social History of Knowledge II: From the Encyclopaedia to Wikipedia*. Cambridge, UK : Polity Press.
- Burnard, Lou D. 1987. "Knowledge Base or Database? Computer Applications in Ethnology", in *Toward a Computer Ethnology*, J. Raben, S. Sugita et M. Kubo (éd.), p. 63-95. Osaka : National Museum of Ethnology. <http://cat.inist.fr/?aModele=afficheN&cpsidt=11990052>.

- Burwell, Sylvia M., Steven VanRoekel, Todd Park et Dominic J. Mancini. 2013. "Open Data Policy-Managing Information as an Asset". Executive Office of the President, Office of Management and Budget. <http://www.whitehouse.gov/.../omb/memoranda/2013/m-13-13.p>.
- Busch, Lawrence. 2013. *Standards: Recipes for Reality*. Cambridge, MA : MIT Press.
- California Digital Library. 2013. "DataUp". <http://dataup.cdlib.org>.
- Callan, Jamie et Alistair Moffat. 2012. "Panel on Use of Proprietary Data". *SIGIR Forum* 46 (2), p. 10-18. <https://sigir.org/files/forum/2012D/p010.pdf>.
- Caltech. 2013a. "Caltech Announces Open-Access Policy". <http://www.caltech.edu/content/caltech-announces-open-access-policy>.
- Caltech. 2013b. "Caltech Center for Advanced Computing Research". <http://www.cacr.caltech.edu/main/?tag=astronomy&paged=2>.
- Campbell, Eric G., Brian R. Clarridge, Manjusha Gokhale, Lauren Birenbaum, Stephen Hilgartner, Neil A. Holtzman et David Blumenthal. 2002. "Data Withholding in Academic Genetics: Evidence from a National Survey", *Journal of the American Medical Association*, 287, p. 473-480. <http://jama.ama-assn.org/cgi/content/full/287/4/473>.
- Carata, Lucian, Sherif Akoush, Nikilesh Balakrishnan, Thomas Bytheway, Ripduman Sohan, Margo Selter et Andy Hopper. 2014. "A Primer on Provenance". *Communications of the ACM*, 57 (5), p. 52-60. <https://m-cacm.acm.org/magazines/2014/5/174341-a-primer-on-provenance/fulltext?mobile=true>.
- Carpenter, Siri. 2012. "Psychology's Bold Initiative". *Science*, 335 (6076), p. 1558-1561. <https://science.sciencemag.org/content/335/6076/1558>.
- Case, Donald O. 2002. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. San Diego : Academic Press.
- Case, Donald O. 2006. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. 2^e éd. San Diego : Academic Press.
- Case, Donald O. 2012. *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior*. 3^e éd. Bingley, UK : Emerald Group Publishing.
- Cavendish, Henry. 1798. "Experiments to Determine the Density of the Earth". *Philosophical Transactions of the Royal Society of London*, 88 (1), p. 469-526. <http://www.phys.ufl.edu/~det/effsource/1798Cavendish.pdf>.
- CBETA Chinese Electronic Tripitaka Collection. 2013. http://www.cbeta.org/cbreader/help/index_e.htm
- Center for Embedded Networked Sensing. 2012. <https://escholarship.org/uc/cens>.
- Center for Embedded Networked Sensing. 2013. "CENS eScholarship Repository". <http://repositories.cdlib.org/cens>.
- Centre national de la recherche scientifique. 2012. "Aladin Sky Atlas".
- Centre national de la recherche scientifique. 2013. "CDS VizieR Service".
- Cha, Meeyoung, Hamed Haddadi, Fabrício Benevenuto et Krishna P. Gummadi. 2010. "Measuring User Influence in Twitter: The Million Follower Fallacy", in *ICWSM '10: Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, p. 10-17. Palo Alto, CA : AAAI.

- Chalmers, Iain. 2011. "Systematic Reviews and Uncertainties about the Effects of Treatments". *Cochrane Database of Systematic Reviews*. <https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.ED000004/full>.
- Chamberlain, Scott. 2013. "Consuming Article-Level Metrics: Observations and Lessons". *Information Standards Quarterly*, 25 (2), p. 4. https://www.niso.org/sites/default/files/stories/2017-08/FE_Chamberlain_Consuming_ALMs_isq_v25no2.pdf.
- Chang, Kevin, Nathan Yau, Mark Hansen et Deborah Estrin. 2006. "SensorBase.org – A Centralized Repository to Slog Sensor Network Data", in *Proceedings of the International Conference on Distributed Networks (DCOSS)/EAWMS*.
- Check Hayden, Erika. 2013. "Geneticists Push for Global Data-Sharing". *Nature*, 498 (7452), p. 16-17. <https://www.nature.com/news/geneticists-push-for-global-data-sharing-1.13133>.
- Chicago Manual of Style*. 2010. 16^e éd. Chicago : University of Chicago Press.
- Cho, Adrian. 2012. "Who Invented the Higgs Boson?". *Science*, 337 (6100), p. 1286-1289. <https://www.jstor.org/stable/23269300>.
- Claerbout, Jon. 2010. "Reproducible Computational Research: A History of Hurdles, Mostly Overcome". <http://sepwww.stanford.edu/sep/jon/reproducible.html>.
- CLARIN European Research Infrastructure Consortium. 2013. <https://www.clarin.eu/content/about-eric>.
- Cochrane Collaboration. 2013. <http://www.cochrane.org>.
- CODATA-ICSTI Task Group on Data Citation Standards and Practices³. 2013. "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data". *Data Science Journal*, 12, p. 1-75. https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article.
- Cohen, Jon. 2012. "WHO Group: H5N1 Papers Should Be Published in Full". *Science* 335 (6071), p. 899-900. <https://science.sciencemag.org/content/335/6071/899>.
- Cohen, Julie E. 2012. *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*. New Haven, CT : Yale University Press.
- Collins, Christopher E. 2011. "Twitter and Rheumatology Based Medical Education – Analysis of the First 100 Followers". *Arthritis and Rheumatism*, 63 (S10), p. 85.
- Collins, Harry M. 1975. "The Seven Sexes: A Study in the Sociology of a Phenomenon, or the Replication of Experiments in Physics". *Sociology*, 9, p. 205-224.
- Collins, Harry M. 1998. "The Meaning of Data: Open and Closed Evidential Cultures in the Search for Gravitational Waves". *American Journal of Sociology*, 104, p. 293-338.
- Collins, Harry M. et Evans Robert. 2007. *Rethinking Expertise*. Chicago : University of Chicago Press.
- Committee on Networked Systems of Embedded Computers. 2001. *Embedded, Everywhere: A Research Agenda for Networked Systems of Embedded Computers*. Washington, DC : National Academies Press. http://www.nap.edu/catalog.php?record_id=10193.
- Committee on Publication Ethics. 2013. <http://publicationethics.org>.

3. Référence traduite dans le texte : Groupe de travail sur les normes et pratiques en matière de citation de données du CODATA-ICSTI.

- ConservationSpace. 2013. <http://www.conservationspace.org/Home.html>.
- Consultative Committee for Space Data Systems. 2012. "Reference Model for an Open Archival Information System", 2. <http://public.ccsds.org/publications/RefModel.aspx>⁴.
- Data Systems. <https://public.ccsds.org/Pubs/650x0m2%28F%>.
- Contadini, Anna. 1993. "Il Grifone Di Pisa", in *Eredità dell'Islam – Arte Islamica in Italia*, Giovanni Curatola (éd.). Milan : Silvana Editoriale.
- Contadini, Anna. 2010. "Translocation and Transformation: Some Middle Eastern Objects in Europe", in *The Power of Things and the Flow of Cultural Transformations: Art and Culture between Europe and Asia*, Lieselotte E. Saurma-Jeltsch et Anja Eisenbeiss (éd.), p. 42-64. Berlin : Deutscher Kunstverlag.
- Contadini, Anna, Richard Camber et Peter Northover. 2002. "Beasts That Roared: The Pisa Griffin and the New York Lion", in *Cairo to Kabul: Afghan and Islamic Studies Presented to Ralph Pinder-Wilson*, Warwick Ball et Leonard Harrow (éd.), p. 65-83. Londres : Melisende.
- COordinated Molecular Probe Line Extinction Thermal Emission Survey of Star Forming Regions [COMPLETE]. 2011. <http://www.cfa.harvard.edu/COMPLETE>.
- Corbyn, Zoë. 2011. "Researchers Failing to Make Raw Data Public". *NATNews*, p. 9-14.
- Corporation for National Research Initiatives. 2013. "Handle System". <http://www.handle.net>.
- Costello, Anthony, Mark Maslin, Hugh Montgomery, Anne M. Johnson et Paul Ekins. 2011. "Global Health and Climate Change: Moving from Denial and Catastrophic Fatalism to Positive Action". *Philosophical Transactions of the Royal Society A*, 369, p. 1866-1882. <https://royalsocietypublishing.org/doi/10.1098/rsta.2011.0007>.
- Council of Science Editors et the Style Manual Committee. 2006. *Scientific Style and Format: The CBE Manual for Authors, Editors, and Publishers*. Reston, VA : CSE and Rockefeller University Press.
- Courant, Paul N. 2009. "The Stakes in the Google Book Search Settlement". *The Economists' Voice*, 6 (9), p. 1-6. https://econpapers.repec.org/article/bpjevoice/v_3a6_3ay_3a2009_3ai_3a9_3an_3a7.htm.
- Couzin, Jennifer et Catherine Unger. 2006. "Cleaning up the Paper Trail". *Science*, 312, p. 38-43. doi:10.1126/science.312.5770.38.
- Couzin-Frankel, Jennifer. 2010. "As Questions Grow, Duke Halts Trials, Launches Investigation". *Science*, 329 (5992), p. 614-615. <https://science.sciencemag.org/content/329/5992/614>.
- Couzin-Frankel, Jennifer. 2013a. "Return of Unexpected DNA Results Urged". *Science* 339 (6127), p. 1507-1508. <https://science.sciencemag.org/content/339/6127/1507.abstract>.
- Couzin-Frankel, Jennifer. 2013b. "Unmasking 'Invisible' Drug Trials". <http://news.sciencemag.org/scienceinsider/2013/06/unmasking-invisible-drug-trials.html>.
- Crabtree, Pam J. 1990. "Zooarchaeology and Complex Societies: Some Uses of Faunal Analysis for the Study of Trade, Social Status, and Ethnicity". *Archaeological Method and Theory*, 2, p. 155-205. <https://ucl.rl.talis.com/items/26F0B361-0890-2575-7CE3-B14CAC683D46.html>.

4. Consultative Committee for Space. 2017. « Modèle de référence pour un Système ouvert d'archivage d'information (OAIS) », 2. <https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf>

- Cragin, Melissa H. et Kalpana Shankar. 2006. "Scientific Data Collections and Distributed Collective Practice". *Computer Supported Cooperative Work*, 15, p. 185-204.
- Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson et Michael Witt. 2010. "Data Sharing, Small Science and Institutional Repositories". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368, p. 4023-4038. <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2010.0165>.
- Crane, Diana. 1970. "Nature of Scientific Communication and Influence". *International Social Science Journal*, 22, p. 28-41.
- Crane, Diana. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago : University of Chicago Press.
- Crane, Gregory R. 2006. "What Do You Do with a Million Books?". *D-Lib Magazine*, 12 (3). <http://www.dlib.org/dlib/march06/crane/03crane.html>.
- Crane, Gregory R., Alison Babeu et David Bamman. 2007. "eScience and the Humanities". *International Journal on Digital Libraries*, 7, p. 117-122. <https://link.springer.com/article/10.1007/s00799-007-0031-8>.
- Crane, Gregory R. et Amy Friedlander. 2008. "Many More Than a Million: Building the Digital Environment for the Age of Abundance. Report of a One-day Seminar on Promoting Digital Scholarship". Washington, DC : Council on Library and Information Resources. <http://www.clir.org/activities/digitalscholar/Nov28final.pdf>.
- Creative Commons. 2013. "Creative Commons License Choices". <http://creativecommons.org/choose>.
- Crombie, Alistair C. 1994. *Styles of Scientific Thinking in the European Tradition: The History of Argument and Explanation Especially in the Mathematical and Biomedical Sciences and Arts*. Londres : Duckworth.
- Cronin, Blaise. 1981. "The Need for a Theory of Citing". *Journal of Documentation*, 37 (1), p. 16-24. <https://www.emerald.com/insight/content/doi/10.1108/eb026703/full/html?skipTracking=true>.
- Cronin, Blaise. 1984. *The Citation Process: The Role and Significance of Citations in Scientific Communication*. Londres : Taylor Graham. <http://garfield.library.upenn.edu/cronin/citationprocess.pdf>.
- Cronin, Blaise. 1994. "Tiered Citation and Measures of Document Similarity". *Journal of the American Society for Information Science American Society for Information Science*, 45 (7), p. 537-538. [https://asistdl.onlinelibrary.wiley.com/doi/10.1002/\(SICI\)1097-4571\(199408\)45:7<537::AID-ASI8>3.0.CO;2-Q](https://asistdl.onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199408)45:7<537::AID-ASI8>3.0.CO;2-Q).
- Cronin, Blaise. 1995. *The Scholar's Courtesy: The Role of Acknowledgement in the Primary Communication Process*. Londres : Taylor Graham.
- Cronin, Blaise. 2005. *The Hand of Science: Academic Writing and Its Rewards*. Lanham, MD : Scarecrow Press.
- Cronin, Blaise et Sara Franks. 2006. "Trading Cultures: Resource Mobilization and Service Rendering in the Life Sciences as Revealed in the Journal Article's Paratext". *Journal of the American Society for Information Science and Technology*, 57 (14), p. 1909-1918. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20407>.

- Crosas, Mercè. 2011. "The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data". *D-Lib Magazine*, 17 (1/2). <https://dlib.org/dlib/january11/crosas/01crosas.html>.
- Crosas, Mercè, Todd Carpenter, David Shotton et Christine L. Borgman. 2013. "Amsterdam Manifesto on Data Citation Principles". <https://www.force11.org/AmsterdamManifesto>.
- CrossRef. 2009. "The Formation of CrossRef: A Short History". <http://www.crossref.org/01company/02history.html>.
- CrossRef. 2013. "FundRef". <http://www.crossref.org/fundref>.
- CrossRef. 2014. "Home Page". <http://www.crossref.org>.
- Cuff, Dana, Mark Hansen et Jerry Kang. 2008. "Urban Sensing: Out of the Woods". *Communications of the ACM*, 51, p. 24-33. <http://doi.acm.org/10.1145/1325555.1325562>.
- Cummings, Jonathon, Thomas A. Finholt, Ian Foster, Carl Kesselman et Katherine A. Lawrence. 2008. "Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations". Washington, DC : National Science Foundation. <http://www.educause.edu/Resources/BeyondBeingThereABlueprintforA/163051>.
- Cuneiform Digital Library Initiative. 2013. <http://cdli.ucla.edu>.
- Curwen, Thomas. 2013. "Capturing the Mysteries of the Sun One Drawing at a Time". *Los Angeles Times*. <http://www.latimes.com/local/columnone/la-me-c1-mt-wilson-sun-spots-20131028-dto,0,4430093.htmlstory-axzz2jlfYqvCU>.
- Dalrymple, Dana. 2003. "Scientific Knowledge as a Global Public Good: Contributions to Innovation and the Economy", in *The Role of Scientific and Technical Data and Information in the Public Domain*, Julie M. Esanu et Paul F. Uhlir (éd.), p. 35-51. Washington, DC : National Academies Press.
- Darch, Peter, Annemarie Carusi, Sharon Lloyd, Marina Jirotko, Grace de la Flor, Ralph Schroeder and Eric Meyer. 2010. "Shared Understandings in e-Science Projects". Technical Report. Oxford e-Research Centre, Oxford University. http://www.oerc.ox.ac.uk/sites/default/files/uploads/ProjectFiles/FLESSR/HiPerDNO/embedding/Shared_Understanding_30_June.pdf.
- Das, Sudeshna, Lisa Girard, Tom Green, Louis Weitzman, Alister Lewis-Bowen et Tim Clark. 2009. "Building Biomedical Web Communities Using a Semantically Aware Content Management System". *Briefings in Bioinformatics*, 10 (2), p. 129-138. <https://academic.oup.com/bib/article/10/2/129/184073>.
- Daston, Lorraine J. 1988. "The Factual Sensibility". *Isis*, 79 (3), p. 452-467. <https://www.jstor.org/stable/234675>.
- Datacitation Synthesis Group. 2014. "Joint Declaration on Data Citation Principles – Final". *Force11: The Future of Research Communications and Scholarship*. <http://www.force11.org/datacitationprinciples>.
- DataCite. 2013. <http://www.datacite.org>.
- Conservancy, Data. 2010. <http://www.dataconservancy.org/about>.
- Data Documentation Initiative. 2012. FAQ. <http://www.ddialliance.org/resources/faq.html>.
- Data Publishing. 2013. <http://datapublishing.com/about>.

- Data Seal of Approval. 2014. <http://www.datasealofapproval.org/en>.
- Davenport, Elisabeth et Blaise Cronin. 2001. "Who Dunnit? Metatags and Hyperauthorship". *Journal of the American Society for Information Science and Technology*, 52 (9), p. 770-773. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.1123>.
- Davenport, James H., James Grant, Catherine M. Jones. 2020. "Data Without Software Are Just Numbers". *Data Science Journal*, 19 (1), 3. <https://doi.org/10.5334/dsj-2020-003>.
- David, Paul A. 1985. "Clio and the Economics of QWERTY". *American Economic Review*, 75, p. 332-337.
- David, Paul A. 2003. "The Economic Logic of 'Open Science' and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer", in *The Role of the Public Domain in Scientific Data and Information*, Julie M. Esanu et Paul F. Uhlig (éd.), p. 19-34. Washington, DC : National Academies Press. <http://www.stanford.edu/group/siepr/cgi-bin/siepr/?q=system/files/shared/pubs/papers/pdf/02-30.pdf>.
- David, Paul A. 2004a. "Can 'Open Science' Be Protected from the Evolving Regime of Intellectual Property Protections". *Journal of Institutional and Theoretical Economics*, 160. <http://www-siepr.stanford.edu/papers/pdf/02-42.pdf>.
- David, Paul A. 2004b. *Towards a Cyberinfrastructure for Enhanced Scientific Collaboration: Providing Its 'Soft' Foundations May Be the Hardest Part*. Oxford : University of Oxford.
- David, Paul A. et Michael Spence. 2003. "Towards Institutional Infrastructures for e-Science: The Scope of the Challenge". Oxford : University of Oxford. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1325240.
- David, Paul A., Matthijs den Besten et Ralph Schroeder. 2010. "Will e-Science Be Open Science?", in *World Wide Research: Reshaping the Sciences and Humanities*, William H. Dutton et Paul W. Jeffreys (éd.), p. 299-316. Cambridge, MA : MIT Press.
- Day, Ronald E. 2001. *The Modern Invention of Information: Discourse, History, and Power*. Carbondale : Southern Illinois University Press.
- De Angelis, Catherine D., Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, Sheldon Kotzin et al. 2005. "Is This Clinical Trial Fully Registered? – A Statement from the International Committee of Medical Journal Editors". *New England Journal of Medicine*, 352 (23), p. 2436-2438. <https://www.nejm.org/doi/full/10.1056/NEJMe058127>.
- De Roure, David, Carole Goble et Robert Stevens. 2009. "The Design and Realisation of the Virtual Research Environment for Social Sharing of Workflows". *Future Generation Computer Systems*, 25 (5), p. 561-567. <https://www.sciencedirect.com/science/article/abs/pii/S0167739X08000939?via=ihub>.
- Declaration on Research Assessment. 2013. <http://am.ascb.org/dora>.
- DeNardis, Laura. 2011. *Opening Standards the Global Politics of Interoperability*. Cambridge, MA : MIT Press. <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=386853>.

- Den Besten, Matthijs, Arthur J. Thomas et Ralph Schroeder. 2009. "Life Science Research and Drug Discovery at the Turn of the 21st Century: The Experience of SwissBioGrid". *Journal of Biomedical Discovery and Collaboration*, 4, p. 5. https://www.researchgate.net/publication/26287884_Life_Science_Research_and_Drug_Discovery_at_the_Turn_of_the_21st_Century_The_Experience_of_SwissBioGrid.
- Deshpande, Amol, Carlos Guestrin, Samuel R. Madden, Joseph M. Hellerstein et Wei Hong. 2004. "Model-driven Data Acquisition in Sensor Networks", in *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, Mario A. Nascimento, M. T. Ozsu, Donald Kossmann, Renee J. Miller, Jose A. Blakeley et K. B. Schiefer (éd.), p. 588-599. Toronto : Morgan Kaufmann. <http://dl.acm.org/citation.cfm?id=1316689.1316741>.
- DeVorkin, David H. et Paul Routly. 1999. "The Modern Society: Changes in Demographics", in *The American Astronomical Society's First Century*, David H. DeVorkin (éd.), p. 122-136. Washington, DC : American Astronomical Society.
- Di Gennaro, Corinna et William H Dutton. 2007. "Reconfiguring Friendships: Social Relationships and the Internet". *Information Communication and Society*, 10 (5), p. 591-618. <https://www.tandfonline.com/doi/abs/10.1080/13691180701657949>.
- [The] Digital Archaeological Record. 2013. <http://www.tdar.org/about>.
- Digital Curation Centre. 2013. "Disciplinary Metadata". <http://www.dcc.ac.uk/resources/metadata-standards>.
- Digital Curation Centre. 2014. "Trustworthy Repositories". <http://www.dcc.ac.uk/resources/repository-audit-and-assessment/trustworthy-repositories>.
- Digital Public Library of America. 2013. <http://dp.la>.
- Digital Social Research. 2013. <http://www.digitalsocialresearch.net/wordpress>.
- Directory of Open Access Journals. 2013. <http://www.doaj.org>.
- Disco, Cornelis et Eda Kranakis (éd.). 2013. *Cosmopolitan Commons: Sharing Resources and Risks across Borders*. Cambridge, MA : MIT Press.
- Doorn, Peter, Ingrid Dillo et René van Horik. 2013. "Lies, Damned Lies and Research Data: Can Data Sharing Prevent Data Fraud?". *International Journal of Digital Curation*, 8 (1), p. 229-243. <http://www.ijdc.net/article/view/8.1.229>.
- Drago, Idilio, Marco Mellia, Maurizio M. Munafo, Anna Sperotto, Ramin Sadre, Aiko Pras, John Byers et Jim Kurose. 2012. "Inside Dropbox: Understanding Personal Cloud Storage Services", in *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, p. 481-494. New York : ACM. <https://dl.acm.org/doi/10.1145/2398776.2398827>.
- Drucker, Johanna. 2014. *Graphesis: Visual Forms of Knowledge Production*. Cambridge : Harvard University Press.
- Dryad. 2013. <http://datadryad.org>.
- Duderstadt, James J., Daniel E. Atkins, John Seely Brown, Marye Anne Fox, Ralph E. Gomory, Nils Hasselmo, Paul M. Horn. et al. 2002. *Preparing for the Revolution: Information Technology and the Future of the Research*. Washington, DC : National Academies Press.

- Duguid, Paul. 2005. "The Art of Knowing": Social and Tacit Dimensions of Knowledge and the Limits of the Community of Practice". *Information Society*, 21, p. 109-118. <https://www.tandfonline.com/doi/abs/10.1080/01972240590925311>.
- Duguid, Paul. 2007. "Inheritance and Loss? A Brief Survey of Google Books". *First Monday*, 12. <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847>.
- Dutton, William H., Corinna di Gennaro et A. Millwood Hargrave. 2005. *Oxford Internet Survey 2005 Report: The Internet in Britain*. Oxford : Oxford Internet Institute, University of Oxford.
- Dutton, William H. et Adrian Shepherd. 2006. "Trust in the Internet as an Experience Technology". *Information Communication and Society*, 9 (4), p. 433-451. <https://www.tandfonline.com/doi/abs/10.1080/13691180600858606>.
- Dutton, William H., Ellen J. Helsper et Monica M. Gerber. 2009. *Oxford Internet Survey 2009 Report: The Internet in Britain*. Oxford : Oxford Internet Institute, University of Oxford.
- Dutton, William H. et Paul W. Jeffreys (éd.). 2010. *World Wide Research: Reshaping the Sciences and Humanities*. Cambridge, MA : MIT Press.
- Dutton, William H. et Grant Blank. 2011. *Next Generation Internet Users: The Internet in Britain 2011*. Oxford : Oxford Internet Institute, University of Oxford.
- Dutton, William H., Grant Blank et Darja Groselj. 2013. *OxIS 2013 Report: Cultures of the Internet*. Oxford : Oxford Internet Institute, University of Oxford.
- Easterbrook, Steve M. et Timothy C. Johns. 2009. "Engineering the Software for Understanding Climate Change". *Computing in Science & Engineering*, 11 (6), p. 65-74. doi:10.1109/MCSE.2009.193. eBird. 2013. <http://ebird.org/content/ebird>.
- Edge, David O. 1979. "Quantitative Measures of Communication in Science: A Critical Review". *History of Science*, 17, p. 102-134.
- Editors. 2013. "Authors, Plagiarists, or Tradents?". *Chinese Buddhist Encyclopedia*. http://china-buddhismencyclopedia.com/en/index.php?title=Authors,_plagiarists,_or_tradents%3F.
- Edwards, Aled M. 2008a. "Bermuda Principles Meet Structural Biology". *Nature Structural & Molecular Biology*, 15 (2), p. 116. <https://www.nature.com/articles/nsmb0208-116>.
- Edwards, Aled M. 2008b. "Open-Source Science to Enable Drug Discovery". *Drug Discovery Today*, 13 (17-18), p. 731-733. doi:10.1016/j.drudis.2008.04.011.
- Edwards, Aled M., Bountra, Chas, Kerr, David J. et Willson, Timothy M. 2009. "Open Access Chemical and Clinical Probes to Support Drug Discovery". *Nature Chemical Biology*, 5 (7), p. 436-440. <https://www.sciencedirect.com/science/article/abs/pii/S1359644608001554?via=ihub>.
- Edwards, Paul N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA : MIT Press.
- Edwards, Paul N. 2013. "Predicting the Weather: An Information Commons for Europe and the World", in *Cosmopolitan Commons: Sharing Resources and Risks Across Borders*, Cornelis Disco et Eda Kranakis (éd.), p. 155-184. Cambridge, MA : MIT Press.
- Edwards, Paul N., Steven J. Jackson, Geoffrey C. Bowker et Cory P. Knobel. 2007. "Understanding Infrastructure: Dynamics, Tensions, and Design". National Science Foundation. Ann Arbor : University of Michigan.

- Edwards, Paul N., Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker et Christine L. Borgman. 2011. "Science Friction: Data, Metadata, and Collaboration". *Social Studies of Science*, 41, p. 667-690. <https://journals.sagepub.com/doi/abs/10.1177/0306312711413314>.
- Edwards, Paul N., Steven J. Jackson, Melissa K. Chalmers, Geoffrey C. Bowker, Christine L. Borgman, David Ribes, Matt Burton, et Scout Calvert. 2013. *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor : University of Michigan. <http://hdl.handle.net/2027.42/97552>.
- Ehrlich, Kate et Debra Cash. 1999. "The Invisible World of Intermediaries: A Cautionary Tale". *Computer Supported Cooperative Work*, 8 (1-2), p. 147-167.
- Eichhorn, Guenther. 1994. "An Overview of the Astrophysics Data System". *Experimental Astronomy*, 5, p. 205-220. <https://link.springer.com/article/10.1007%2F01583697>.
- Eisen, Michael. 2012. "Blinded by Big Science: The Lesson I Learned from ENCODE Is That Projects like ENCODE Are Not a Good Idea". <http://www.michaeleisen.org/blog/?p=1179>.
- Eisenstein, Elizabeth. 1979. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe*. Cambridge, UK : Cambridge University Press.
- Embedded Metadata Initiative. 2013. "Embedded Metadata Manifesto". <http://www.embedded-metadata.org/embedded-metatdata-manifesto.php>.
- EndNote. 2013. <http://endnote.com>.
- Enserink, Martin. 2006. "Avian Influenza: As H5N1 Keeps Spreading, a Call to Release More Data". *Science*, 311, p. 1224. <https://science.sciencemag.org/content/311/5765/1224/tab-article-info>.
- Enserink, Martin. 2012a. "Fraud-Detection Tool Could Shake Up Psychology". *Science Insider*. <http://news.sciencemag.org/2012/07/fraud-detection-tool-could-shake-psychology?rss=1>.
- Enserink, Martin. 2012b. "Public at Last, H5N1 Study Offers Insight into Virus's Possible Path to Pandemic". *Science*, 336 (6088), p. 1494-1497. <https://science.sciencemag.org/content/336/6077/61>.
- Enserink, Martin et Cohen, Jon. 2012. "One H5N1 Paper Finally Goes to Press; Second Greenlighted". *Science*, 336 (6081), p. 529-530. <https://science.sciencemag.org/content/336/6081/529/tab-article-info>.
- Erdos, David. 2013a. "Freedom of Expression Turned on Its Head? Academic Social Research and Journalism in the European Privacy Framework". *Public Law*. <http://papers.ssrn.com/abstract=1928177>.
- Erdos, David. 2013b. "Mustn't Ask, Mustn't Tell: Could New EU Data Laws Ban Historical and Legal Research?". UK Constitutional Law Group (February 14). <http://ukconstitutionallaw.org/2013/02/14/david-erdos-mustnt-ask-mustnt-tell-could-new-eu-data-laws-ban-historical-and-legal-research>.
- Esanu, Julie M. et Paul F. Uhlir (éd.). 2004. *Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium, March 10-11, Paris*. Washington, DC : National Academies Press.

- Estrin, Deborah, K. Mani Chandy, R. Michael Young, Larry Smarr, Andrew Odlyzko, David Clark, Viviane Reding *et al.* 2010. "Internet Predictions". *IEEE Internet Computing*, 14 (1), p. 12-42. <https://ieeexplore.ieee.org/document/5370818>.
- Estrin, Judy. 2008. *Closing the Innovation Gap: Reigniting the Spark of Creativity in a Global Economy*. New York : McGraw-Hill.
- Ettema, James S. et F. Gerald Kline. 1977. "Deficits, Differences, and Ceilings: Contingent Conditions for Understanding the Knowledge Gap". *Communication Research*, 4 (2), p. 179-202. <https://journals.sagepub.com/doi/10.1177/009365027700400204>.
- Europeana. 2013. <http://www.europeana.eu>.
- European Southern Observatory. 2013. "Common Pipeline Library". <http://www.eso.org/sci/software/cpl>.
- European Union Publications Office. 2018. *Turning FAIR data into reality: Final report and action plan from the European Commission expert group on FAIR data*. <https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF>
- Eysenbach, Gunther. 2011. "Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact". *Journal of Medical Internet Research*, 13 (4). <https://www.jmir.org/2011/4/e123/>.
- Falk, Harry. 2011. "The 'Split' Collection of Kharosthi Texts", in *Annual Report of the International Research Institute for Advanced Buddhism at Soka University for the Academic Year 2010*, p. 13-23. Tokyo : The International Research Institute for Advanced Buddhism, Soka University.
- Faniel, Ixchel M. et Trond E. Jacobsen. 2010. "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data". *Computer Supported Cooperative Work*, 19, p. 355-375. <https://link.springer.com/article/10.1007/s10606-010-9117-8>.
- Faoláin, Simon Ó. et Peter J. Northover. 1998. "The Technology of Late Bronze Age Sword Production in Ireland". *Journal of Irish Archaeology*, 9, p. 69-88.
- Fenner, Martin. 2013. "Letter from the Guest Content Editor: Altmetrics Have Come of Age". *Information Standards Quarterly*, 25 (2), p. 3. <https://www.niso.org/niso-io/isq/2013/06/information-standards-quarterly-summer-2013>.
- Field, Dawn, Sansone, Susanna-Assunta Sansone, Amanda Collis, Tim Booth, Peter Dukes, Susan K. Gregurick, Karen Kennedy *et al.* 2009. "Omics Data Sharing". *Science*, 326 (5950), p. 234-236. <https://science.sciencemag.org/content/326/5950/234.abstract>.
- Fienberg, Stephen E., Margaret E. Martin et Miron L. Straf. 1985. *Sharing Research Data*. Washington, DC : National Academies Press. http://books.nap.edu/catalog.php?record_id=2033.
- Finch, Janet. 2012. "Accessibility, Sustainability, Excellence: How to Expand Access to Research Publications". Report of the Working Group on Expanding Access to Published Research Findings. Londres : Research Information Network. <http://www.researchinfonet.org/publish/finch>.
- Finkbeiner, Ann K. 2010. *A Grand and Bold Thing: The Extraordinary New Map of the Universe Ushering in a New Era of Discovery*. New York : Free Press.

- Fischman, Josh. 2012. "Fake Peer Reviews, the Latest Form of Scientific Fraud, Fool Journals". *Chronicle of Higher Education*. <http://chronicle.com/article/Fake-Peer-Reviews-the-Latest/134784>.
- Fisher, Celia B. 2006. "Clinical Trials Databases: Unanswered Questions". *Science*, 311 (5758), p. 180-181. <https://science.sciencemag.org/content/311/5758/180/tab-article-info>.
- Fitzpatrick, Kathleen. 2011. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York : New York University Press.
- Flor, Grace de la, Marina Jirotko, Paul Luff, John Pybus et Ruth Kirkham. 2010. "Transforming Scholarly Practice: Embedding Technological Interventions to Support the Collaborative Analysis of Ancient Texts". *Computer Supported Cooperative Work*, 19 (3-4), p. 309-334. <https://link.springer.com/article/10.1007%2Fs10606-010-9111-1>.
- Forbes, Duncan. 2008. "So You Want to Be a Professional Astronomer". *Mercury Magazine*. <http://www.astronomynotes.com/careers/Mercury-career.pdf>.
- Foster, Ian, Christine L. Borgman, P. Bryan Heidorn, William Howe et Carl Kesselman. 2013. "Empowering Long Tail Research". <https://sites.google.com/site/ieltrconcept>.
- Foster, Ian et Luc Moreau. 2006. *Provenance and Annotation of Data*. Heidelberg : Springer. http://www.w3.org/2011/prov/wiki/Connection_Task_Force_Informal_Report.
- Foster, Jonathan B. et Alyssa A. Goodman. 2006. "Cloudshine: New Light on Dark Clouds". *Astrophysical Journal Letters*, 636 (2), L105.
- Foucault, Michel. 1994. *The Order of Things: An Archaeology of the Human Sciences*. New York : Vintage Books.
- Fouchier, Ron A. M., Sander Herfst et Albert D. M. E. Osterhaus. 2012. "Restricted Data on Influenza H5N1 Virus Transmission". *Science*, 335 (6069), p. 662-663. <https://science.sciencemag.org/content/335/6069/662/tab-article-info>.
- Fox, Peter et Ray Harris. 2013. "ICSU and the Challenges of Data and Information Management for International Science". *Data Science Journal*, 12, WDS1-WDS12. https://www.jstage.jst.go.jp/article/dsj/12/0/12_WDS-001/_article.
- Frankel, Henry. 1976. "Alfred Wegener and the Specialists". *Centaurus*, 20 (4), p. 305- 324. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0498.1976.tb00937.x>.
- Freeman, Linton C. 2004. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver, BC : Empirical Press.
- Friedlander, Amy. 2009. "Asking Questions and Building a Research Agenda for Digital Scholarship", in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*, Kathlin Smith et Brian Leney (éd.), p. 1-15. Washington, DC : Council on Library and Information Resources. <http://www.clir.org/pubs/resources/pubs/reports/pub145/pub145.pdf>.
- Frischer, Bernard, Philip Stinson, Neil A. Silberman et Dirk Callebaut. 2002. "Scientific Verification and Model-making Methodology: Case Studies of the Virtual Reality Models of the House of Augustus (Rome) and Villa of the Mysteries (Pompeii)", in *Interpreting the Past: Heritage, New Technologies & Local Development*. Bruxelles : Flemish Heritage Institute.

- Furnas, Alexander et Devin Gaffney. 2012. "Statistical Probability That Mitt Romney's New Twitter Followers Are Just Normal Users: 0%". *The Atlantic*. <http://www.theatlantic.com/technology/archive/2012/07/statistical-probability-that-mitt-romneys-new-twitter-followers-are-just-normal-users-0/260539>.
- Furner, Jonathan. 2003a. "Little Book, Big Book: Before and after Little Science, Big Science: A Review Article, Part I". *Journal of Librarianship and Information Science*, 35, p. 115-125. <https://journals.sagepub.com/doi/10.1177/0961000603352006>.
- Furner, Jonathan. 2003b. "Little Book, Big Book: Before and after Little Science, Big Science: A Review Article, Part II". *Journal of Librarianship and Information Science*, 35, p. 189-201. <https://journals.sagepub.com/doi/abs/10.1177/0961000603353006>.
- Furner, Jonathan. 2004a. "Conceptual Analysis: A Method for Understanding Information as Evidence, and Evidence as Information". *Archival Science*, 4, p. 233-265. <https://link.springer.com/article/10.1007/s10502-005-2594-8>.
- Furner, Jonathan. 2004b. "Information Studies Without Information". *Library Trends*, 52, p. 427-446.
- Furner, Jonathan. 2010. "Philosophy and Information Studies". *Annual Review of Information Science & Technology*, 44 (1), p. 159-200. <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.2010.1440440111>.
- Gale Cengage Learning. 2013. "Eighteenth Century Collections Online". <http://gale.cengage.co.uk/product-highlights/history/eighteenth-century-collections-online.aspx>.
- Galilei, Galileo. 1610. *Sidereus Nuncius*. Tommaso Baglioni et Herbert M. Evans (éd.). Venetiis : Apud Thomam Baglionum. <http://archive.org/details/Sidereusnuncius00Gali>.
- Gallagher, Ryan. 2013. "NSA Phone Spying: EPIC, Privacy International File Lawsuits to Halt Government Surveillance". *Slate*. http://www.slate.com/blogs/future_tense/2013/07/08/nsa_phone_spying_epic_privacy_international_file_lawsuits_to_halt_government.html.
- Gamazon, Eric R. 2012. "Small Science: High Stakes". *Science*, 338 (6109), p. 883. <https://science.sciencemag.org/content/338/6109/883.1>.
- Gamble, Matthew et Carole Goble. 2011. "Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model". *ACM WebSci'11*, p. 1-8. http://www.websci11.org/fileadmin/websci/Papers/177_paper.pdf.
- Garfield, Eugene. 1955. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas". *Science*, 122 (3159), p. 108-111. <https://science.sciencemag.org/content/122/3159/108>.
- Garfinkel, H. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, NJ : Prentice Hall.
- Geertz, Clifford. 1973. *The Interpretation of Cultures*. New York : Basic Books.
- General Social Survey. 2013. <http://www3.norc.ox.ac.uk/gss+website>.
- Genova, Françoise. 2013. "Strasbourg Astronomical Data Center (CDS)". *Data Science Journal*, 12, WDS56-WDS60. <https://datascience.codata.org/articles/abstract/10.2481/dsj.WDS-007/>.

- Getty Research Institute. 2013. "Getty Vocabularies". <http://www.getty.edu/research/tools/vocabularies>.
- Getty Trust. 2013. "Open Content Program". <http://www.getty.edu/about/open-content.html>.
- Gibbons, Ann. 2012. "A Crystal-Clear View of an Extinct Girl's Genome". *Science*, 337 (6098), p. 1028-1029. <https://science.sciencemag.org/content/337/6098/1028/tab-article-info>.
- Gil, Yolanda, James Cheney, Paul Groth, Olaf Hartig, Simon Miles, Luc Moreau et Paulo Pinheiro da Silva. 2010. "Provenance XG Final Report". W3C Incubator Group. <http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214>.
- Gilliland, Anne J. 2008. "Setting the Stage", in *Introduction to Metadata*, Murtha Baca (éd.). 3^e éd. Los Angeles : Getty Research Institute. http://www.getty.edu/research/publications/electronic_publications/intrometadata.
- Gilliland-Swetland, Anne J. 1998. "Defining Metadata", in *Introduction to Metadata: Pathways to Digital Information*, Murtha Baca (éd.), p. 1-8. Los Angeles : Getty Research Institute.
- Ginsparg, Paul. 1994. "First Steps towards Electronic Research Communication". *Computers in Physics*, 8 (4), p. 390-396. <http://dl.acm.org/citation.cfm?id=187178.187185>.
- Ginsparg, Paul. 2001. "Creating a Global Knowledge Network", in *Second Joint ICSU Press-UNESCO Expert Conference on Electronic Publishing in Science*. Paris : UNESCO. <http://people.ccmr.cornell.edu/~ginsparg/blurp/pg01unesco.html>.
- Gitelman, Lisa (éd.). 2013. *"Raw Data" Is an Oxymoron*. Cambridge, MA : MIT Press.
- Gladwell, Malcolm. 2002. *The Tipping Point: How Little Things Can Make a Big Difference*. New York : Back Bay Books.
- Glaser, Barney G. et Anselm L. Strauss. 1967. *The Discovery of Grounded Theory; Strategies for Qualitative Research*. Chicago : Aldine Publishing.
- Gleck, P. H. 2011. "Climate Change and the Integrity of Science (Letter to Editor; 255 Signatories)". *Science*, 328 (5979), p. 689-690. <https://science.sciencemag.org/content/328/5979/689>.
- Gnip. 2013a. "Gnip Twitter Activity Streams Format". <http://support.gnip.com/customer/portal/articles/477765-twitter-activity-streams-format>.
- Gnip. 2013b. <http://gnip.com>.
- Goble, Carole et David De Roure. 2009. "The Impact of Workflow Tools on Dataintensive Research", in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Tony Hey, Stewart Tansley et Kristin Tolle (éd.), p. 137-146. Redmond, WA : Microsoft.
- Goble, Carole, David De Roure et Sean Bechhofer. 2013. "Accelerating Scientists' Knowledge Turns", in *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Ana Fred, Jan L. G. Dietz, Kecheng Liu et Joaquim Filipe (éd.), p. 3-25. Berlin : Springer.
- Goenka, S. N., Myungsoo Kim, Lewis Lancaster, John R. McRae, Charles Muller, Min Bahadur Shakya, Morten Schlutter et Christian Wittern. 2008. "CBETA 10 Years". Chinese Buddhist Electronic Text Association. <http://www.cbeta.org/data/cbeta10y/friends.htm>.
- Goldacre, Ben. 2008. *Bad Science*. Londres : Fourth Estate.
- Goldacre, Ben. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. Londres : Fourth Estate.

- Goldsmith, Jack L. et Tim Wu. 2006. *Who Controls the Internet? Illusions of a Borderless World*. Oxford : Oxford University Press.
- Goodman, Alyssa A. 2012. "Principles of High-dimensional Data Visualization in Astronomy". *Astronomische Nachrichten*, 333, p. 505. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asna.201211705>.
- Goodman, Alyssa A., Erik W. Rosolowsky, Michelle A. Borkin, Jonathan B. Foster, Michael Halle, Jens Kauffmann et Jaime E. Pineda. 2009a. "A Role for Self-gravity at Multiple Length Scales in the Process of Star Formation". *Nature*, 457 (7225), p. 63-66. <https://www.nature.com/articles/nature07609>.
- Goodman, Alyssa A. et Curtis G. Wong. 2009b. "Bringing the Night Sky Closer: Discoveries in the Data Deluge", in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Tony Hey, Stewart Tansley et Kristin Tolle (éd.), p. 39-44. Redmond, WA : Microsoft.
- Goodman, Alyssa A., Jaime E. Pineda et Scott L. Schnee. 2009c. "The 'True' Column Density Distribution in Star-Forming Molecular Clouds". *Astrophysical Journal*, 692, p. 91-103. <https://dash.harvard.edu/handle/1/4268294>.
- Goodman, Alyssa A., August Muench et Alicia Soderberg. 2012. "Introducing the Astronomy Dataverse (theastrodata.org)". Presentation and Panel Discussion (April 2). Cambridge, MA : Harvard-Smithsonian Center for Astrophysics. <http://thedata.org/presentations/introducing-astronomy-dataverse-theastrodataorg>.
- Goodman, Alyssa A., Jonathan Fay, August Muench, Alberto Pepe, Patricia Udomprasert et Curtis Wong. 2012. "WorldWide Telescope in Research and Education". *arXiv:1201.1285*. <http://arxiv.org/abs/1201.1285>.
- Goodman, Alyssa A., Joao Alves, Chris Beaumont, Tom Dame, James Jackson, Jens Kauffmann, Thomas Robitaille *et al.* 2013. "The Bones of the Milky Way". *Astrophysical Journal*. https://www.authorea.com/users/23/articles/249/_show_article.
- Goodman, Alyssa A., Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano *et al.* 2014. "10 Simple Rules for the Care and Feeding of Scientific Data". *PLoS Computational Biology*, 10 (4). <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542>.
- Grafton, Anthony. 2007. "Future Reading". *The New Yorker*. http://www.newyorker.com/reporting/2007/11/05/071105fa_fact_grafton?currentPage=all.
- Gray, Jim, David T. Liu, Maria Nieto-Santisteban, Alexander Szalay, David DeWitt et Gerd Heber. 2005. "Scientific Data Management in the Coming Decade". *CT Watch Quarterly*, 1. <http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management>.
- Gray, Jim et Alexander Szalay. 2002. "The World-wide Telescope". *Communications of the ACM*, 45, p. 51-55.
- Groth, Paul, Yolanda Gil, James Cheney et Simon Miles. 2012. "Requirements for Provenance on the Web". *International Journal of Digital Curation*, 7 (1), p. 39-56. <http://www.ijdc.net/article/view/203>.
- Groth, Paul et Luc Moreau (éd.). 2013. "PROV-Overview". W3C. <http://www.w3.org/TR/prov-overview>.

- Guibault, Lucie. 2013. "Licensing Research Data under Open Access Conditions", in *Information and Knowledge: 21st Century Challenges in Intellectual Property and Knowledge Governance*, Dana Beldiman (éd.). Cheltenham : Edward Elgar.
- Gutmann, Myron, Mark Abrahamson, Margaret Adams, Micah Altman, Caroline R. Arms et Gary King. 2009. "From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data". *Library Trends*, 57, p. 315-337. <https://muse.jhu.edu/issue/14067>.
- Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin et Yaniv Erlich. 2013. "Identifying Personal Genomes by Surname Inference". *Science*, 339 (6117), p. 321-324. <https://science.sciencemag.org/content/339/6117/321.abstract>.
- Haak, Laurel L., David Baker, Donna K. Ginther, Gregg J. Gordon, Matthew A. Probus, Nirmala Kannankutty et Bruce A. Weinberg. 2012. "Standards and Infrastructure for Innovation Data Exchange". *Science*, 338 (6104), p. 196-197. <https://science.sciencemag.org/content/338/6104/196/tab-article-info>.
- Hackett, Edward J., Olga Amsterdamska, Michael Lynch et Judy Wajcman. 2007. *The Handbook of Science and Technology Studies*. 3^e éd. Cambridge, MA : MIT Press.
- Hamilton, David P. 1990. "Information Decontrol Urged". *Science*, 248 (4958), p. 957-958. <https://science.sciencemag.org/content/248/4958/957/tab-article-info>.
- Hamilton, Michael P., Eric A. Graham, Philip W. Rundel, Michael F. Allen, William Kaiser, Mark H. Hansen et Deborah L. Estrin. 2007. "New Approaches in Embedded Networked Sensing for Terrestrial Ecological Observatories". *Environmental Engineering Science*, 24 (2), p. 149-150.
- Hanisch, Robert J. 2013. "The Future of the Virtual Observatory – US Virtual Astronomical Observatory". <http://www.usvao.org/2013/08/12/the-future-of-the-virtual-observatory>.
- Hanisch, Robert J., A. Farris, E. W. Greisen, W. D. Pence, B. M. Schlesinger, P. J. Teuben, R. W. Thompson et A. Warnock. 2001. "Definition of the Flexible Image Transport System (FITS)". *Astronomy and Astrophysics*, 376 (1), p. 359-380.
- Hanisch, Robert J. et Peter J. Quinn. 2002. "The International Virtual Observatory". <http://www.ivoa.net/about/TheIVOA.pdf>.
- Hanson, Karen, Alisa Surkis et Karen Yacobucci. 2012. *Data Sharing and Management Snafu in 3 Short Acts*. Film and Animation. <http://www.youtube.com/watch?v=N2zK3sAtr-4&sns=em>.
- Hardin, Garrett. 1968. "The Tragedy of the Commons". *Science*, 162 (3859), p. 1243-1248. <https://science.sciencemag.org/content/162/3859/1243>.
- Harnad, Stevan. 1998. "The Invisible Hand of Peer Review", *Nature*, 5. <http://www.nature.com/nature/webmatters/invisible/invisible.html>.
- Harpring, Patricia. 2010. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. Los Angeles : Getty Publications.
- Hart, Michael S. 1992. "History and Philosophy of Project Gutenberg". <http://www.gutenberg.org/about/history>.
- Hart, Michael S. 2013. "Project Gutenberg". <http://www.gutenberg.org>.

- Harvard Law Review Association. 2005. *The Bluebook: A Uniform System of Citation*. Cambridge, MA : Author.
- Harvard-Smithsonian Astrophysical Observatory. 2013a. "Chandra X-ray Observatory". <http://chandra.harvard.edu>.
- Harvard-Smithsonian Astrophysical Observatory. 2013b. "Digitizing the Harvard College Observatory Astronomical Plate Stacks". <http://tdc-www.harvard.edu/plates>.
- Harvard-Smithsonian Astrophysical Observatory. 2013c. "The SAO/NASA Astrophysics Data System". <http://adswww.harvard.edu>.
- Harvard University. 2010. "Open Access Policies". Harvard University Library, Office for Scholarly Communication. <https://osc.hul.harvard.edu/policies>.
- Harvard University and Wellcome Trust. 2012. International Workshop on Contributorship and Scholarly Attribution. http://projects.iq.harvard.edu/files/attribution_workshop/files/iwcsa_report_final_18sept12.pdf.
- Hedstrom, Margaret, Lee Dirks, Nicholas Economides, Peter Fox, Michael F. Goodchild, Heather Joseph, Ronald L. Larsen *et al.* 2014. "Future Career Opportunities and Educational Requirements for Digital Curation". http://sites.nationalacademies.org/PGA/brdi/PGA_069853.
- Heidorn, Bryan. 2008. "Shedding Light on the Dark Data in the Long Tail of Science". *Library Trends*, 57 (2), p. 280-299. <https://muse.jhu.edu/article/262029>.
- Hess, Charlotte et Elinor Ostrom. 2007a. "Introduction: An Overview of the Knowledge Commons", in *Understanding Knowledge as a Commons: From Theory to Practice*, Charlotte Hess et Elinor Ostrom (éd.), p. 3-26. Cambridge, MA : MIT Press.
- Hess, Charlotte et Elinor Ostrom. 2007b. *Understanding Knowledge as a Commons: From Theory to Practice*. Cambridge, MA : MIT Press.
- Hettne, Kristina, Katy Wolstencroft, Khalid Belhajjame, Carole Goble, Eleni Mina et Harish Dharuri. 2012. "Best Practices for Workflow Design: How to Prevent Workflow Decay", in *Proceedings of the ESWC 2012 Workshop on the Future of Scholarly Communication in the Semantic Web*, Frank Van Harmelen, Alexander G. Castro, Christoph Lange et Benjamin Good (éd.). Grèce : Sepublica. http://ceur-ws.org/Vol-952/paper_23.pdf.
- Hey, Tony, Stewart Tansley et Kristin Tolle. 2009. "Jim Gray on eScience: A Transformed Scientific Method", in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Tony Hey, Stewart Tansley et Kristin Tolle (éd.), p. xix-xxxiii. Redmond, WA : Microsoft.
- Hey, Tony et Anne Trefethen. 2005. "Cyberinfrastructure and e-Science". *Science*, 308, p. 818-821.
- Higgins, D., C. Berkley et M. B. Jones. 2002. "Managing Heterogeneous Ecological Data Using Morpho", in *Proceedings 14th International Conference on Scientific and Statistical Database Management*, p. 69-76. Édimbourg, UK : IEEE Computer Society.
- Hilgartner, Stephen. 1997. "Access to Data and Intellectual Property: Scientific Exchange in Genome Research", in *Intellectual Property Rights and the Dissemination of Research Tools in Molecular Biology. Summary of a Workshop Held at the National Academy of Science, February 15-16, 1996*, p. 28-39. Washington, DC : National Academies Press.

- Hilgartner, Stephen et Sherry I. Brandt-Rauf. 1994. "Data Access, Ownership and Control: Toward Empirical Studies of Access Practices". *Knowledge*, 15, p. 355-372.
- Hine, Christine. 2000. *Virtual Ethnography*. Londres : Sage.
- Hine, Christine. 2008. *Systematics as Cyberscience: Computers, Change, and Continuity in Science*. Cambridge, MA : MIT Press.
- Hirsh, Sandra G. 1996. *The Effect of Domain Knowledge on Elementary School Children's Information Retrieval Behavior on an Automated Library Catalog*. Los Angeles : UCLA.
- Hirtle, Peter B. 2011. "Introduction to Intellectual Property Rights in Data Management". <https://confluence.cornell.edu/display/rdmsgweb/introduction-intellectual-property-rights-data-management>.
- Hirtle, Peter B. 2012. "When Is 1923 Going to Arrive and Other Complications of the U.S. Public Domain". *Searcher*, 20 (6). <http://www.infotoday.com/searcher/sep12/Hirtle-When-Is-1923-Going-to-Arrive-and-Other-Complications-of-the-U.S.-Public-Domain.shtml>.
- Hirtle, Peter B., Emily Hudson et Andrew T. Kenyon. 2009. *Copyright and Cultural Institutions: Guidelines for U.S. Libraries, Archives, and Museums*. <http://hdl.handle.net/1813/14142>.
- H-Net. 2013a. "H-Buddhism Discussion Network". <https://www.h-net.org/~buddhism>.
- H-Net. 2013b. "What Is H-Net?". <http://www.h-net.org/about>.
- Hockey, Susan. 1999. "Making Technology Work for Scholarship: Investing in the Data", in *Technology and Scholarly Communication*, Richard Ekman et Richard E. Quandt (éd.). Berkeley : University of California Press ; publié en association avec The Andrew K. Mellon Foundation. <http://ark.cdlib.org/ark:/13030/ft5w10074r>.
- Hogg, David W. et Dustin Lang. 2008. "Astronomical Imaging: The Theory of Everything". *arXiv:0810.3851*. <https://aip.scitation.org/doi/abs/10.1063/1.3059072>.
- Holdren, John P. 2013a. "Increasing Access to the Results of Federally Funded Scientific Research". Executive Office of the President, Office of Science and Technology Policy. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
- Holdren, John P. 2013b. "Memorandum for the Heads of Executive Departments and Agencies". Executive Office of the President, Office of Science and Technology Policy. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
- Hollan, Jim et Scott Stornetta. 1992. "Beyond Being There", in *CHI '92*, Penny Bauersfeld, John Bennett et Gene Lynch (éd.), p. 119-125. New York : ACM.
- Hollinger, David A. 2013. "The Wedge Driving Academe's Two Families Apart". *Chronicle of Higher Education*. http://chronicle.com/article/Why-Cant-the-Sciencethe/142239/?cid=cr&utm_source=cr&utm_medium=en.
- Howard, Jennifer. 2013a. "Posting Your Latest Article? You Might Have to Take It Down". *Chronicle of Higher Education [Blog]*. <http://chronicle.com/blogs/wiredcampus/posting-your-latest-article-you-might-have-to-take-it-down/48865>.
- Howard, Jennifer. 2013b. "White House Delivers New Open-Access Policy That Has Activists Cheering". *Chronicle of Higher Education*. http://chronicle.com/article/White-House-Delivers-New/137549/?cid=at&utm_source=at&utm_medium=en.

- Howard, Jennifer. 2014. "Born Digital, Projects Need Attention to Survive". *Chronicle of Higher Education*. <http://chronicle.com/article/Born-Digital-Projects-Need/143799>.
- Howe, Doug, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill *et al.* 2008. "Big Data: The Future of Biocuration". *Nature*, 455 (7209), p. 47-50. <https://www.nature.com/articles/455047a>.
- Hrynaskiewicz, Iain et Douglas G. Altman. 2009. "Towards Agreement on Best Practice for Publishing Raw Clinical Trial Data". *Trials*, 10, p. 17. <https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-10-17>.
- HubbleSite. 2013a. <http://hubblesite.org>.
- HubbleSite. 2013b. "The Telescope: Hubble Essentials". http://hubblesite.org/the_telescope/hubble_essentials.
- Hughes, Thomas P. 1989. "The Evolution of Large Technological Systems", in *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, Wiebe E. Bijker, Thomas P. Hughes et Trevor J. Pinch (éd.), p. 51-82. Cambridge, MA : MIT Press.
- Hughes, Thomas P. 2004. *Human-Built World: How to Think About Technology and Culture*. Chicago : University of Chicago Press.
- Hugo, W. 2013. "A Maturity Model for Digital Data Centers". *Data Science Journal*, 12, WDS189-WDS192.
- Hunter, Jane. 2009. "Collaborative Semantic Tagging and Annotation Systems". *Annual Review of Information Science & Technology*, 43, p. 187-239.
- Hvistendahl, Mara. 2013. "China's Publication Bazaar". *Science*, 342 (6162), p. 1035-1039. <https://science.sciencemag.org/content/342/6162/1035.summary>.
- IAU Working Group Libraries. 2013. "Best Practices for Creating a Telescope Bibliography". *IAU-Commission5 – WG Libraries*. <http://iau-commission5.wikispaces.com/WG+Libraries>.
- Incorporated Research Institutions for Seismology. 2013. <http://www.iris.edu/hq>.
- Ingwersen, Peter. 1998. "The Calculation of Web Impact Factors". *Journal of Documentation*, 54, p. 236-243.
- Ingwersen, Peter et Kalervo Jarvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Dordrecht : Springer.
- Institute for Quantitative Social Sciences. 2011. "Data Citation Principles Workshop". Harvard University. http://projects.iq.harvard.edu/datacitation_workshop.
- Institute for Quantitative Social Sciences. 2013. Harvard University. <http://www.iq.harvard.edu>.
- Institute of Electrical and Electronics Engineers. 2013. "IEEE Code of Ethics". <http://www.ieee.org/about/corporate/governance/p7-8.html>.
- International Astronomical Union. 2013. "About the International Astronomical Union". <http://www.iau.org/about>.

- International Committee of Medical Journal Editors. 2013. "Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals"⁵. <http://www.icmje.org/recommendations>.
- International Council of Museums. 2013. "CIDOC Conceptual Reference Model". http://www.cidoc-crm.org/who_we_are.html.
- International Dunhuang Project. 2013. <http://idp.bl.uk>.
- International Social Survey Programme. 2013. <http://www.issp.org>.
- International Standard Book Number Agency. 2013. <http://www.isbn.org>.
- International Standard Name Identifier International Agency. 2013. <http://www.isni.org>.
- International Standard Serial Number International Centre. 2013. "Understanding the ISSN: What Is an ISSN?". <http://www.issn.org/2-22636-All-about-ISSN.php>.
- International Virtual Observatory Alliance. 2013a. <http://www.ivoa.net>.
- International Virtual Observatory Alliance. 2013b. "Virtual Observatory Applications for Astronomers". <http://www.ivoa.net/astronomers/applications.html>.
- Inter-University Consortium for Political and Social Research. 2012. "Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle". 5^e éd. Ann Arbor, MI : ICPSR. <https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/index.html>.
- Inter-University Consortium for Political and Social Research. 2013. <http://www.icpsr.umich.edu/icpsrweb/landing.jsp>.
- Ioannidis, John P. A. et Muin J. Khoury. 2011. "Improving Validation Practices in 'Omics' Research". *Science*, 334 (6060), p. 1230-1232. doi:10.1126/science.1211811.
- IRIS Data Management Center. 2013. <http://www.iris.edu/data>.
- ISO 10646. 2013. "Ideograph Characters". http://glyph.iso10646hk.net/english/icharacters_1.jsp.
- Jackson, Brian A., Tora K. Bikson et Patrick P. Gunn. 2013. "Human Subjects Protection and Research on Terrorism and Conflict". *Science*, 340 (6131), p. 434-435. <https://science.sciencemag.org/content/340/6131/434/tab-article-info>.
- Jackson, Steven J. 2006. "Water Models and Water Politics: Deliberative Design and Virtual Accountability", in *Proceedings of the 7th Annual International Conference on Digital Government Research*, Jose A. B. Fortes et Ann MacIntosh (éd.), p. 95-104. San Diego : Digital Government Research Center.
- Jackson, Steven J., David Ribes, Ayse Buyuktur et Geoffrey C. Bowker. 2011. "Collaborative Rhythm: Temporal Dissonance and Alignment in Collaborative Scientific Work", in *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, p. 245-254. New York : ACM. <https://dl.acm.org/doi/10.1145/1958824.1958861>.
- Jackson, Steven J. et Ayse Buyuktur. 2014. "Who Killed WATERS? Mess, Method, and Forensic Explanation in the Making and Unmaking of Large-Scale Science Networks". *Science, Technology & Human Values*, 39 (2), p. 285-308. <https://journals.sagepub.com/doi/abs/10.1177/0162243913516013>.

5. Référence française : International Committee of Medical Journal Editors. 2013. « Recommandations pour la conduite, la présentation, la rédaction et la publication des travaux de recherche soumis à des revues médicales ». <http://www.icmje.org/recommendations/translations/french2019.pdf>

- Jacobs, Neil. 2006. *Open Access: Key Strategic, Technical and Economic Aspects*. Oxford : Chandos.
- James San Jacinto Mountains Reserve. 2013. "Data Resources". http://www.jamesreserve.edu/data_arch.html.
- Janee, Greg et James Frew. 2002. "The ADEPT Digital Library Architecture", in *Second ACM/IEEE-CS Joint Conference on Digital Libraries*, p. 342-350. New York : ACM.
- Jantz, Ronald et Michael J. Giarlo. 2005. "Digital Preservation: Architecture and Technology for Trusted Digital Repositories". *D-Lib Magazine*, 11. <http://www.dlib.org/dlib/june05/jantz/06jantz.html>.
- Jasny, Barbara R., Gilbert Chin, Lisa Chong et Sacha Vignieri. 2011. "Again, and Again, and Again...". *Science*, 334 (6060), p. 1225. <https://science.sciencemag.org/content/334/6060/1225>.
- JavaScript Object Notation. 2013. <http://www.json.org>.
- Jenkins, Henry, Ravi Purushotma, Margaret Weigel et Katie Clinton. 2009. *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*. Cambridge, MA : MIT Press.
- Jirotko, Marina, Rob Procter, Tom Rodden et Geoffrey C. Bowker. 2006. "Special Issue: Collaboration in e-Research". *Computer Supported Cooperative Work*, 15, p. 251-255.
- Johnson, George. 2007. "A Trip Back in Time and Space". *New York Times*. <http://www.nytimes.com/2007/07/10/science/10astro.html>.
- Johnston, E. H. 1938. "The Gop Ipur Bricks". *Journal of the Royal Asiatic Society*, 70 (4), p. 547-553. <https://www.cambridge.org/core/journals/journal-of-the-royal-asiatic-society/article/gopalpur-bricks/5E055F2336F51D9FF491DCF8B1DB1AAC>.
- Jones, James Howard. 1981. *Bad Blood: The Tuskegee Syphilis Experiment*. New York : Free Press.
- Journal of Open Archaeology Data*. 2013. <http://openarchaeologydata.metajnl.com>.
- JoVE: Peer Reviewed Scientific Video Journal*. 2013. <http://www.jove.com>.
- Kahin, Brian et Dominique Foray. 2006. *Advancing Knowledge and the Knowledge Economy*. Cambridge, MA : MIT Press.
- Kaiser, Jocelyn. 2008. "Making Clinical Data Widely Available". *Science*, 322 (5899), p. 217-218. <https://science.sciencemag.org/content/322/5899/217/tab-article-info>.
- Kaiser, Jocelyn. 2013. "Rare Cancer Successes Spawn 'Exceptional' Research Efforts". *Science*, 340 (6130), p. 263. <https://science.sciencemag.org/content/340/6130/263.summary>.
- Kanfer, Alana G., Caroline Haythornthwaite, B. C. Bruce, Geoffrey C. Bowker, N. C. Burbules, J. F. Porac et J. Wade. 2000. "Modelling Distributed Knowledge Processes in Next Generation Multidisciplinary Alliances". *Information Systems Frontiers*, 2 (3-4), p. 317-331. <https://ieeexplore.ieee.org/document/843277>.
- Kansa, Eric C. 2012. "Openness and Archaeology's Information Ecosystem". *World Archaeology*, 44 (4), p. 498-520. <https://www.tandfonline.com/doi/abs/10.1080/00438243.2012.737575>.
- Kansa, Sarah W., Eric C. Kansa et J. M. Schultz. 2007. "An Open Context for Near Eastern Archaeology". *Near Eastern Archaeology*, 70 (4), p. 188-194.

- Karabag, Solmaz Filiz et Christian Berggren. 2012. "Retraction, Dishonesty and Plagiarism: Analysis of a Crucial Issue for Academic Publishing, and the Inadequate Responses from Leading Journals in Economics and Management Disciplines". *Journal of Applied Economics and Business Research*, 4 (2). <http://www.aebrjournal.org/volume-2-issue-4.html>.
- Karasti, Helena, Karen S. Baker et Eija Halkola. 2006. "Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network". *Computer Supported Cooperative Work*, 15 (4), p. 321-358. <https://link.springer.com/article/10.1007/s10606-006-9023-2>.
- Kelty, Christopher M. 2008. *Two Bits: The Cultural Significance of Free Software*. Durham, NC : Duke University Press.
- Kelty, Christopher M. 2012. "This Is Not an Article: Model Organism Newsletters and the Question of 'Open Science'". *Biosocieties*, 7 (2), p. 140-168. <https://link.springer.com/article/10.1057/biosoc.2012.8>.
- Kessler, Elizabeth A. 2012. *Picturing the Cosmos: Hubble Space Telescope Images and the Astronomical Sublime*. Minneapolis : University of Minnesota Press.
- King, C. Judson, Diane Harley, Sarah Earl-Novell, Jennifer Arter, Shannon Lawrence, et Irene Perciali. 2006. "Scholarly Communication: Academic Values and Sustainable Models". Berkeley : Center for Studies in Higher Education. <http://cshe.berkeley.edu/publications/scholarly-communication-academic-values-and-sustainable-models>.
- King, Christopher. 2013. "Single-Author Papers: A Waning Share of Output, but Still Providing the Tools for Progress". *ScienceWatch*. <http://sciencewatch.com/articles/single-author-papers-waning-share-output-still-providing-tools-progress>.
- King, Gary, Jennifer Pan et Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression". *American Political Science Review*, 107 (2), p. 326-343. <https://www.cambridge.org/core/journals/american-political-science-review/article/how-censorship-in-china-allows-government-criticism-but-silences-collective-expression/C7EF4A9C9D59425C2D09D83742C1FE00>.
- Kintisch, Eli. 2010. "Embattled U.K. Scientist Defends Track Record of Climate Center". *Science*, 327 (5968), p. 934. <https://science.sciencemag.org/content/327/5968/934/tab-article-info>.
- Klump, Jens, Roland Bertelmann, Jan Brase, Michael Diepenbroek, Hannes Grobe, Heinke Höck, Michael Lautenschlager, Uwe Schindler, Irina Sens et Joachim Wächter. 2006. "Data Publication in the Open Access Initiative". *Data Science Journal*, 5, p. 79-83. <https://datascience.codata.org/articles/abstract/10.2481/dsj.5.79/>.
- Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA : Harvard University Press.
- Knowledge Network for Biocomplexity. 2010. "Ecological Metadata Language". <http://knb.ecoinformatics.org/software/eml>.
- Knowledge Network for Biocomplexity. 2013. "Ecological Metadata Language (EML) Specification". <http://knb.ecoinformatics.org/software/eml/eml-2.1.1/index.html>.

- Knox, Keith T. 2008. "Enhancement of Overwritten Text in the Archimedes Palimpsest", in *Computer Image Analysis in the Study of Art*, David G. Stork et Jim Coddington (éd.), vol. 6810, 681004. SPIE. <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/6810/681004/Enhancement-of-overwritten-text-in-the-Archimedes-Palimpsest/10.1117/12.766679.short?SSO=1>.
- Kolb, David A. 1981. "Learning Styles and Disciplinary Differences", in *The Modern American College*, Arthur W. Chickering (éd.). San Francisco : Jossey-Bass.
- Kolowich, Steve. 2014. "Hazards of the Cloud: Data-Storage Service's Crash Sets Back Researchers". *The Chronicle of Higher Education [Blog]*. <http://chronicle.com/blogs/wiredcampus/hazards-of-the-cloud-data-storage-services-crash-sets-back-researchers/52571>.
- Korn, Naomi et Charles Oppenheim. 2011. "Licensing Open Data: A Practical Guide". JISC. http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf.
- Korsmo, Fae L. 2010. "The Origins and Principles of the World Data Center System". *Data Science Journal*, 8, p. 55-65. https://datascience.codata.org/articles/abstract/10.2481/dsj.SS_IGY-011/.
- Koshland, Daniel. 1987. "Sequencing the Human Genome". *Science*, 236 (4801), p. 505. <https://science.sciencemag.org/content/236/4801/505/tab-article-info>.
- Kouw, Matthijs, Charles Van den Heuvel et Andrea Scharnhorst. 2013. "Exploring Uncertainty in Knowledge Representations: Classifications, Simulations, and Models of the World", in *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences*, Paul Wouters, Anne Beaulieu, Andrea Scharnhorst et Sally Wyatt (éd.), p. 127-149. Cambridge, MA : MIT Press.
- Kranich, Nancy. 2004. "The Information Commons: A Public Policy Report". New York : The Free Expression Policy Project, Brennan Center for Justice, NYU School of Law. <http://www.fepproject.org/policyreports/InformationCommons.pdf>.
- Kranich, Nancy. 2007. "Countering Enclosure: Reclaiming the Knowledge Commons", in *Understanding Knowledge as a Commons: From Theory to Practice*, Charlotte Hess et Elinor Ostrom (éd.), p. 85-122. Cambridge, MA : MIT Press.
- Kraut, Robert, Sara Kiesler, Bonka Boneva, Jonathon Cummings, Vicki Helgeson et Anne Crawford. 2002. "Internet Paradox Revisited". *Journal of Social Issues*, 58 (1), p. 49-74. <https://spssi.onlinelibrary.wiley.com/doi/10.1111/1540-4560.00248>.
- Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago : University of Chicago Press.
- Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions*. 2^e éd. Chicago : University of Chicago Press.
- Kurtz, Donna et T. S. Lipinski. 1995. "Telecommunications for the Arts in Archive, Museum and Gallery: The Beazley Archive and Cast Gallery, Ashmolean Museum, Oxford", in *Networking in the Humanities: Proceedings of the Second Conference on Scholarship and Technology in the Humanities*, Stephanie Kenna et Seamus Ross (éd.), p. 97-109. Londres : Bowker-Saur.

- Kurtz, Donna, Greg Parker, David Shotton, Graham Klyne, Florian Schroff, Andrew Zisserman et Yorick Wilks. 2009. "CLAROS – Bringing Classical Art to a Global Public", in *Proceedings of the 2009 Fifth IEEE International Conference on e-Science*, p. 20-27. Oxford : IEEE. https://www.researchgate.net/profile/Greg_Parker3/publication/221187968_CLAROS_-_Bringing_classical_art_to_a_global_public/links/0deec519dd4652c954000000/CLAROS-Bringing-classical-art-to-a-global-public.pdf.
- Kurtz, Michael J. et Johan Bollen. 2010. "Usage Bibliometrics", in *Annual Review of Information Science and Technology*, vol. 44, Blaise Cronin (éd.). Medford, NJ : Information Today.
- Kurtz, Michael J., Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Markus Demleitner et Stephen S. Murray. 2005. "Worldwide Use and Impact of the NASA Astrophysics Data System Digital Library". *Journal of the American Society for Information Science and Technology*, 56 (1), p. 36-45. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20095>.
- Kurtz, Michael J., Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray et Joyce M. Watson. 2000. "The NASA Astrophysics Data System: Overview". *Astronomy & Astrophysics. Supplement Series*, 143 (1), p. 41-59. <https://ui.adsabs.harvard.edu/abs/2000A%26AS..143...41K/abstract>.
- Kwa, Chunglin. 2005. "Local Ecologies and Global Science: Discourses and Strategies of the International Geosphere-Biosphere Programme". *Social Studies of Science*, 35, p. 923-950.
- Kwa, Chunglin. 2011. *Styles of Knowing*. Pittsburgh : University of Pittsburgh Press.
- Kwa, Chunglin et Rene Rector. 2010. "A Data Bias in Interdisciplinary Cooperation in the Sciences: Ecology in Climate Change Research", in *Collaboration in the New Life Sciences*, John N. Parker, Niki Vermeulen et Bart Penders (éd.), p. 161-176. Farnheim, UK : Ashgate.
- Kwak, Haewoon, Changhyun Lee, Hosung Park et Sue Moon. 2010. "What Is Twitter, a Social Network or a News Media?", in *Proceedings of the 19th International Conference on World Wide Web*, p. 591-600. New York : ACM. <https://dl.acm.org/doi/10.1145/1772690.1772751>.
- Kwon, Diana. 2018. "Plan S: The Ambitious Initiative to End the Reign of Paywalls". *The Scientist Magazine*®. <https://www.the-scientist.com/news-opinion/plan-s-the-ambitious-initiative-to-end-the-reign-of-paywalls-65231>.
- Laakso, Mikael et Bo-Christer Björk. 2013. "Delayed Open Access: An Overlooked High-impact Category of Openly Available Scientific Literature". *Journal of the American Society for Information Science and Technology*, 64 (7), p. 1323-1329. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22856>.
- Laakso, Mikael, Patrik Welling, Helena Bukvova, Linus Nyman, Bo-Christer Björk et Turid Hedlund. 2011. "The Development of Open Access Journal Publishing from 1993 to 2009". *PLoS ONE*, 6 (6). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020961>.
- Lagerstrom, Jill, Sherry L. Winkelman, Uta Grothkopf et Marsha Bishop. 2012. "Observatory Bibliographies: Current Practices", in *Observatory Operations: Strategies, Processes, and Systems IV*, vol. 8448, Alison B. Peck, Robert L. Seaman et Fernando Comerón (éd.). Amsterdam : SPIE.
- Lagoze, Carl et Theresa Velden. 2009a. "Communicating Chemistry". *Nature Chemistry*, 1, p. 673-678. <https://www.nature.com/articles/nchem.448>.

- Lagoze, Carl et Theresa Velden. 2009b. "The Value of New Scientific Communication Models for Chemistry". <http://ecommons.cornell.edu/handle/1813/14150>.
- Laine, Christine, Richard Horton, Catherine D. DeAngelis, Jeffrey M. Drazen, Frank A. Frizelle, Fiona Godlee, Charlotte Haug *et al.* 2007. "Clinical Trial Registration – Looking Back and Moving Ahead". *New England Journal of Medicine*, 356 (26), p. 2734-2736. <https://www.nejm.org/doi/full/10.1056/nejme078110>.
- Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago : University of Chicago Press.
- Lampland, Martha et Susan Leigh Star (éd.). 2009. *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Ithaca, NY : Cornell University Press.
- Landsberger, Henry A. 1958. *Hawthorne Revisited: Management and the Worker: Its Critics, and Developments in Human Relations in Industry*. Ithaca, NY : Cornell University Press.
- Lane, Julia I., Ian Mulvaney et Paco Nathan (éd.). 2020. *Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure*. New York : Sage. <https://us.sagepub.com/en-us/nam/rich-search-and-discovery-for-research-datasets/book270223>.
- Laney, Doug. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety". <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
- Lang, Dustin, David W. Hogg, Keir Mierle, Michael Blanton et Sam Roweis. 2009. "Astrometry. net: Blind Astrometric Calibration of Arbitrary Astronomical Images". *arXiv:0910.2233*. <https://iopscience.iop.org/article/10.1088/0004-6256/139/5/1782/meta>.
- Large Synoptic Survey Telescope Corporation. 2010. <http://www.lsst.org/lsst>.
- Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA : Harvard University Press.
- Latour, Bruno. 1988. "Drawing Things Together", in *Representation in Scientific Practice*, Michael E. Lynch et Steve Woolgar (éd.), p. 19-68. Cambridge, MA : MIT Press.
- Latour, Bruno. 1993. *We Have Never Been Modern*. Cambridge, MA : Harvard University Press.
- Latour, Bruno. 2004. *Politics of Nature: How to Bring Sciences into Democracy*. Cambridge, MA : Harvard University Press.
- Latour, Bruno et Steve Woolgar. 1979. *Laboratory Life: The Construction of Scientific Facts*. Beverly Hills, CA : Sage.
- Latour, Bruno et Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ : Princeton University Press.
- Lave, Jean et Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK : Cambridge University Press.
- Lawrence, Bryan, Catherine Jones, Brian Matthews, Sam Pepler et Sarah Callaghan. 2011. "Citation and Peer Review of Data: Moving towards Formal Data Publication". *International Journal of Digital Curation*, 6 (2), p. 4-37. <http://www.ijdc.net/article/view/181>.

- Lee, Charlotte P., Paul Dourish et Gloria Mark. 2006. "The Human Infrastructure of Cyberinfrastructure", in *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, p. 483-492. New York : ACM.
- Lehman, Richard et Elizabeth Loder. 2012. "Missing Clinical Trial Data: A Threat to the Integrity of Evidence Based Medicine". *British Medical Journal*, 344. <https://www.bmj.com/content/344/bmj.d8158.abstract>.
- Leland, John. 2013. "Online Battle over Ancient Scrolls Spawns Real-World Consequences". *New York Times*. <http://www.nytimes.com/2013/02/17/nyregion/online-battle-over-ancient-scrolls-spawns-real-world-consequences.html>.
- Leonelli, Sabina. 2019a. "Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science". *Harvard Data Science Review*, 1 (1). <https://doi.org/10.1162/99608f92.17405bb6>.
- Leonelli, Sabina. 2019b. "Data – From objects to assets". *Nature*, 574 (7778), p. 317-320. <https://doi.org/10.1038/d41586-019-03062-w>
- Leptin, Maria. 2012. "Open Access – Pass the Buck". *Science*, 335 (6074), p. 1279. <https://science.sciencemag.org/content/335/6074/1279>.
- Lessig, Lawrence. 1999. *Code and Other Laws of Cyberspace*. New York : Basic Books.
- Lessig, Lawrence. 2001. *The Future of Ideas: The Fate of the Commons in a Connected World*. New York : Random House.
- Lessig, Lawrence. 2004. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York : Penguin.
- Lévi-Strauss, Claude. 1966. *The Savage Mind*. Chicago : University of Chicago Press.
- Levien, Roger. S., Robert Austein, Christine L. Borgman, Timothy Casey, Hugh Dubberly, Patrik Faltstrom, Per-Kristian Halvorsen et al. 2005. *Signposts in Cyberspace: The Domain Name System and Internet Navigation*. Washington, DC : National Academies Press.
- Lewis, Anthony, Paul N. Courant, Laine Farley, Paula Kaufman et John Leslie King. 2010. "Google & the Future of Books: An Exchange". *New York Review of Books*. <http://www.nybooks.com/articles/archives/2010/jan/14/google-the-future-of-books-an-exchange>.
- Libicki, Martin C. 1995. "Standards: The Rough Road to the Common Byte", in *Standards Policy for Information Infrastructure*, Brian Kahin et Janet Abbate (éd.), p. 35-78. Cambridge, MA : MIT Press.
- Licklider, J. C. R. 1960. "Man-Computer Symbiosis". *IRE Transactions on Human Factors in Electronics*, 1, p. 4-11. <http://groups.csail.mit.edu/medg/people/psz/Licklider.html>.
- Lide, David R. et Gordon H. Wood. 2012. *CODATA @ 45 Years: 1966 to 2010. The Story of the ICSU Committee on Data for Science and Technology (CODATA) from 1966 to 2010*. Paris : CODATA. <http://www.codata.org/about/CODATA@45years.pdf>.
- Lievrouw, Leah A. 2010. "Social Media and the Production of Knowledge: A Return to Little Science?". *Social Epistemology*, 24 (3), p. 219-237. <https://www.tandfonline.com/doi/abs/10.1080/02691728.2010.499177?scroll=top&needAccess=true&journalCode=tsep20>.
- Lievrouw, Leah A. et Sonia Livingstone. 2002. *The Handbook of New Media*. Londres : Sage.

- Lin, Rong Gong, Rosanna Xia et Doug Smith. 2014a. "In Reversal, Quake Researchers to Turn over List of Concrete Buildings". *Los Angeles Times*. <http://www.latimes.com/local/la-me-01-18-concrete-building-quake-20140118,0,371340.story-axzz2tcxSukHC>.
- Lin, Rong Gong, Rosanna Xia et Doug Smith. 2014b. "UC Releases List of 1,500 Buildings; Big Step for L.A. Quake Safety". *Los Angeles Times*. <http://www.latimes.com/local/lanow/la-me-ln-concrete-buildings-list-20140125,0,425,6501.story-axzz2tcxSukHC>.
- Lipetz, Ben-Ami. 1965. "Improvement of the Selectivity of Citation Indexes to Science Literature through Inclusion of Citation Relationship Indicators". *American Documentation*, 16 (2), p. 81-90.
- Liu, Alan. 2004. *The Laws of Cool: Knowledge Work and the Culture of Information*. Chicago : University of Chicago Press.
- Lofland, John, David Snow, Leon Anderson et Lyn H. Lofland. 2006. *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont, CA : Wadsworth/Thomson Learning.
- Lozano, George A., Vincent Lariviere et Yves Gingras. 2012. "The Weakening Relationship between the Impact Factor and Papers' Citations in the Digital Age". *Journal of the American Society for Information Science and Technology*, 63 (11), p. 2140-2145. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22731>.
- Lyman, Peter. 1996. "What Is a Digital Library? Technology, Intellectual Property, and the Public Interest". *Daedalus: Proceedings of the American Academy of Arts and Sciences*, 125, p. 1-33.
- Lynch, Clifford A. 2009. "Jim Gray's Fourth Paradigm and the Construction of the Scientific Record", in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Tony Hey, Stewart Tansley et Kristin Tolle (éd.), p. 177-184. Redmond, WA : Microsoft.
- Lynch, Clifford A. 2013. "The Next Generation of Challenges in the Curation of Scholarly Data", in *Research Data Management: Practical Strategies for Information Professionals*, Joyce M. Ray (éd.). West Lafayette, IL : Purdue University Press.
- Lynch, Michael E. et Steve Woolgar. 1988a. "Introduction: Sociological Orientations to Representational Practice in Science", in *Representation in Scientific Practice*, Michael E. Lynch et Steve Woolgar (éd.), p. 1-19. Cambridge, MA : MIT Press.
- Lynch, Michael E. et Steve Woolgar (éd.). 1988b. *Representation in Scientific Practice*. Cambridge, MA : MIT Press.
- Machlup, Fritz et Una Mansfield. 1983. *The Study of Information: Interdisciplinary Messages*. New York : Wiley.
- MacLean, Don. 2004. *Internet Governance: A Grand Collaboration*. New York: United Nations ICT Task Force.
- MacRoberts, Michael H. et Barbara R. MacRoberts. 1989. "Problems of Citation Analysis: A Critical Review". *Journal of the American Society for Information Science American Society for Information Science*, 40 (5), p. 342-349.
- MacRoberts, Michael H. et Barbara R. MacRoberts. 2010. "Problems of Citation Analysis: A Study of Uncited and Seldom-cited Influences". *Journal of the American Society for Information Science and Technology*, 61 (1), p. 1-12. <https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.21228>.

- Maisonneuve, Hervé. 2019. « Recommandations pour la conduite, la présentation, la rédaction et la publication des travaux de recherche soumis à des revues médicales », <http://www.icmje.org/recommendations/translations/french2019.pdf>
- Makice, Kevin. 2009. *Twitter API: Up and Running*. Sebastopol, CA : O'Reilly Media.
- Mandell, Rachel Alyson. 2012. "Researchers' Attitudes towards Data Discovery: Implications for a UCLA Data Registry". <http://escholarship.org/uc/item/5bv8j7g3>.
- Manyika, James, Michael Chui, Diana Farrell, Steve Van Kuiken, Peter Groves et Elizabeth Almasi Doshi. 2013. "Open Data: Unlocking Innovation and Performance with Liquid Information". McKinsey & Company. http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.
- Marchionini, Gary et Gregory R. Crane. 1994. "Evaluating Hypermedia and Learning: Methods and Results from the Perseus Project". *ACM Transactions on Information Systems*, 12, p. 5-34.
- Marcus, Adam. 2013. "Influential Reinhart-Rogoff Economics Paper Suffers Spreadsheet Error". *Retraction Watch*. <http://retractionwatch.wordpress.com/2013/04/18/influential-reinhart-rogooff-economics-paper-suffers-database-error>.
- Marcus, George E. 1995. "Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography". *Annual Review of Anthropology*, 24 (1), p. 95-117. <https://www.annualreviews.org/doi/abs/10.1146/annurev.an.24.100195.000523>.
- Markus, M. Lynne et Mark Kell. 1994. "If We Build It, They Will Come: Designing Information Systems That People Want to Use". *MIT Sloan Management Review*, 35 (4), p. 11-25.
- Marshall, Catherine C. 2008a. "Rethinking Personal Digital Archiving, Part 1". *D-Lib Magazine*, 14 (3/4). <http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html>.
- Marshall, Catherine C. 2008b. "Rethinking Personal Digital Archiving, Part 2". *D-Lib Magazine*, 14 (3/4). <http://www.dlib.org/dlib/march08/marshall/03marshall-pt2.html>.
- Marshall, Catherine C. 2009. "No Bull, No Spin: a Comparison of Tags with Other Forms of User Metadata", in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, p. 241-250. New York : ACM. <https://dl.acm.org/doi/10.1145/1555400.1555438>.
- Marshall, Eliot. 2011. "Unseen World of Clinical Trials Emerges From U.S. Database". *Science*, 333 (6039), p. 145. <https://science.sciencemag.org/content/333/6039/145/tab-article-info>.
- Mathae, Kathie Bailey et Paul F. Uhler (éd.). 2012. *The Case for International Sharing of Scientific Data: A Focus on Developing Countries; Proceedings of a Symposium*. Washington, DC : National Academies Press.
- Maunsell, John. 2010. "Announcement Regarding Supplemental Material". *Journal of Neuroscience* 30 (32), p. 10599-10600. <http://www.jneurosci.org/content/30/32/10599>.
- Mayer, Rob. 2010. "Authors, Plagiarists, or Tradents?". *Kili Kīlaya*. <http://blogs.orient.ox.ac.uk/kila/2010/10/09/authors-plagiarists-or-tradents>.
- Mayer-Schonberger, Viktor et Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston : Houghton Mifflin Harcourt.

- Mayernik, Matthew S. 2011. "Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators". PhD diss., Los Angeles : University of California at Los Angeles. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2042653.
- Mayernik, Matthew S. 2015. "Research Data and Metadata Curation as Institutional Issues". *Journal of the Association for Information Science and Technology*.
- Mayernik, Matthew S., Jillian C. Wallis et Christine L. Borgman. 2007. "Adding Context to Content: The CENS Deployment Center", in *Proceedings of the American Society for Information Science & Technology*, vol. 44. Milwaukee, WI : Information Today.
- Mayernik, Matthew S., Archer L. Batcheller et Christine L. Borgman. 2011. "How Institutional Factors Influence the Creation of Scientific Metadata", in *Proceedings of the 2011 iConference*, p. 417-425. New York : ACM. <https://dl.acm.org/doi/10.1145/1940761.1940818>.
- Mayernik, Matthew S., Jillian C. Wallis et Christine L. Borgman. 2012. "Unearthing the Infrastructure: Humans and Sensors in Field-based Research". *Computer Supported Cooperative Work*, 22 (1), p. 65-101. <https://link.springer.com/article/10.1007/s10606-012-9178-y>.
- Mayernik, Matthew S., David L. Hart, Keith E. Maull, Nicholas M. Weber. 2017. "Assessing and tracing the outcomes and impact of research infrastructures". *Journal of the Association for Information Science and Technology*, 68 (6), p. 1341-1359. https://www.researchgate.net/publication/305821104_Assessing_and_tracing_the_outcomes_and_impact_of_research_infrastructures.
- McCain, Katherine W. 2012. "Assessing Obliteration by Incorporation: Issues and Caveats". *Journal of the American Society for Information Science and Technology*, 63 (11), p. 2129-2139. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22719>.
- McCray, W. Patrick. 2000. "Large Telescopes and the Moral Economy of Recent Astronomy". *Social Studies of Science*, 30 (5), p. 685-711. <https://www.jstor.org/stable/i212661>.
- McCray, W. Patrick. 2001. "What Makes a Failure? Designing a New National Telescope, 1975-1984". *Technology and Culture*, 42 (2), p. 265-291. https://www.history.ucsb.edu/wp-content/uploads/histpublications/files/00228-2001_mccray_what_makes_a_failure.pdf.
- McCray, W. Patrick. 2003. "The Contentious Role of a National Observatory". *Physics Today*, 56 (10), p. 55-61. <https://physicstoday.scitation.org/doi/10.1063/1.1629005>.
- McCray, W. Patrick. 2004. *Giant Telescopes: Astronomical Ambition and the Promise of Technology*. Cambridge, MA : Harvard University Press.
- McCray, W. Patrick. 2014. "How Astronomers Digitized the Sky". *Technology & Culture*.
- McGuire, Amy L., Steven Joffe, Barbara A. Koenig, Barbara B. Biesecker, Laurence B. McCullough, Jennifer S. Blumenthal-Barby, Timothy Caulfield, Sharon F. Terry et Robert C. Green. 2013. "Ethics and Genomic Incidental Findings". *Science*, 340 (6136), p. 1047-1048. <https://science.sciencemag.org/content/early/2013/05/15/science.1240156.abstract>.
- McLaughlin, Jamie, Michael Meredith, Michael Pidd et Katherine Rogers. 2010. "A Review of the AHRC Technical Appendix and Recommendations for a Technical Plan". Sheffield, UK : Humanities Research Institute, University of Sheffield. http://digital.humanities.ox.ac.uk/Support/technicalappendix_final3.pdf.

- Meadows, A. J. 1974. *Communication in Science*. Londres : Butterworths.
- Meadows, A. J. 1998. *Communicating Research*. San Diego : Academic Press.
- Meadows, Jack. 2001. *Understanding Information*. Munich : K. G. Saur.
- Mele, Salvatore. 2013. "Higgs Boson Discovery at CERN: Physics and Publishing". *The Oxford Internet Institute, Innovation and Digital Scholarship Lecture Series Events*. <http://www.oii.ox.ac.uk/events/?id=598>.
- Mendeley. 2013. <http://www.mendeley.com/features>.
- Meng, Xiao-Li. 2011. "Multi-party Inference and Uncongeniality", in *International Encyclopedia of Statistical Science*, Miodrag Lovric (éd.), p. 884-888. Berlin : Springer.
- Merton, Robert K. 1963a. "The Ambivalence of Scientists". *Bulletin of the Johns Hopkins Hospital*, 112, p. 77-97.
- Merton, Robert K. 1963b. "The Mosaic of the Behavioral Sciences", in *The Behavioral Sciences Today*, Bernard Berelson (éd.), p. 247-272. New York : Basic Books.
- Merton, Robert K. 1968. "The Matthew Effect in Science". *Science*, 159, p. 56-63. <https://science.sciencemag.org/content/159/3810/56>.
- Merton, Robert K. 1970. "Behavior Patterns of Scientists". *Leonardo*, 3, p. 213-220.
- Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. Norman W. Storer (éd.). Chicago : University of Chicago Press.
- Merton, Robert K. 1988. "The Matthew Effect in Science II: Cumulative Advantage and the Symbolism of Intellectual Property". *Isis*, 79, p. 606-623.
- Merton, Robert K. 1995. "The Thomas Theorem and the Matthew Effect". *Social Forces*, 74 (2), p. 379-422. <https://www.jstor.org/stable/2580486?origin=crossref>.
- Meyer, Eric T. 2009. "Moving from Small Science to Big Science: Social and Organizational Impediments to Large Scale Data Sharing", in *e-Research: Transformations in Scholarly Practice*, Nicholas W. Jankowski (éd.), p. 147-159. New York : Routledge.
- Meyer, Eric T. et Ralph Schroeder. 2014. *Digital Transformations of Research*. Cambridge, MA : MIT Press.
- Meyer, Eric T., Ralph Schroeder et Linnet Taylor. 2013. "Big Data in the Study of Twitter, Facebook and Wikipedia: On the Uses and Disadvantages of Scientificity for Social Research". Presented at the Annual Meeting of the American Sociological Association Annual Meeting, Hilton New York and Sheraton New York.
- Miles, Alistair, Jun Zhao, Graham Klyne, Helen White-Cooper et David Shotton. 2010. "OpenFlyData: An Exemplar Data Web Integrating Gene Expression Data on the Fruit Fly *Drosophila Melanogaster*". *Journal of Biomedical Informatics*, 43 (5), p. 752-761. <https://www.sciencedirect.com/science/article/pii/S1532046410000511>.
- Milgram, Stanley. 1974. *Obedience to Authority: An Experimental View*. New York: Harper & Row.
- Miller, Mary K. 2007. "Reading between the Lines". *Smithsonian Magazine*. <http://www.smithsonianmag.com/science-nature/archimedes.html>.
- Millerand, Florence et Geoffrey C. Bowker. 2009. "Metadata Standards: Trajectories and Enactment in the Life of an Ontology", in *Standards and Their Stories*, Martha Lampland et Susan Leigh Star (éd.), p. 149-165. Ithaca, NY : Cornell University Press.

- Mimno, David, Gregory Crane et Alison Jones. 2005. "Hierarchical Catalog Records: Implementing a FRBR Catalog". *D-Lib Magazine*, 11 (10). <http://www.dlib.org/dlib/october05/crane/10crane.html>.
- Ministry of Education, Republic of China. 2000. "Dictionary of Chinese Character Variants". <http://dict.variants.moe.edu.tw>.
- MIT Libraries. 2009. "MIT Faculty Open Access Policy". Scholarly Publishing @ MIT Libraries. <http://libraries.mit.edu/scholarly/mit-open-access/open-access-at-mit/mit-open-access-policy>.
- Modern Language Association of America. 2009. *MLA Handbook for Writers of Research Papers*. New York : The Modern Language Association of America.
- Moffett, Jonathan. 1992. "The Beazley Archive: Making a Humanities Database Accessible to the World". *Bulletin of the John Rylands University of Manchester*, 74 (3), p. 39-52.
- Monaghan, Peter. 2013. "They Said at First That They Hadn't Made a Spreadsheet Error, When They Had". *Chronicle of Higher Education*. <http://chronicle.com/article/UMass-Graduate-Student-Talks/138763>.
- Moreau, Luc. 2010. "The Foundations for Provenance on the Web". *Foundations and Trends® in Web Science*, 2 (2-3), p. 99-241. <https://www.nowpublishers.com/article/Details/WEB-010>.
- Moreau, Luc, Paul Groth, Simon Miles, Javier Vazquez-Salceda, John Ibbotson, Sheng Jiang, Steve Munroe et al. 2008. "The Provenance of Electronic Data". *Communications of the ACM*, 51 (4), p. 52-58. <https://dl.acm.org/doi/10.1145/1330311.1330323>.
- Mullaney, Thomas S. 2012. "The Moveable Typewriter: How Chinese Typists Developed Predictive Text During the Height of Maoism". *Technology and Culture*, 53 (4), p. 777-814. doi:10.1353/tech.2012.0132.
- Munns, David P. D. 2012. *A Single Sky: How an International Community Forged the Science of Radio Astronomy*. Cambridge, MA : MIT Press.
- Murphy, Fiona. 2013. "The Now and the Future of Data Publishing". *Wiley Exchanges*. <http://exchanges.wiley.com/blog/2013/07/05/the-now-and-the-future-of-data-publishing>.
- Murray-Rust, Peter et Henry S. Rzepa. 2004. "The Next Big Thing: From Hypermedia to Datuments". *Journal of Digital Information* 5 (1). <http://journals.tdl.org/jodi/article/view/130>.
- Murray-Rust, Peter, Cameron Neylon, Rufus Pollock et John Wilbanks. 2010. "Panton Principles". <http://pantonprinciples.org>.
- Murray-Rust, Peter, Henry S. Rzepa, Simon Tyrrell et Yong Zhang. 2004. "Representation and Use of Chemistry in the Global Electronic Age". *Organic & Biomolecular Chemistry*, 2, p. 3192-3203.
- Murthy, Dhiraj. 2011. "Twitter: Microphone for the Masses?". *Media Culture & Society* 33 (5), p. 779-789. <https://journals.sagepub.com/doi/10.1177/0163443711404744>.
- Naik, Gautam. 2011. "Scientists' Elusive Goal: Reproducing Study Results". *Wall Street Journal*. <http://online.wsj.com/news/articles/SB10001424052970203764804577059841672541590>.
- NASA's Earth Observing System Data and Information System. 2013. "Processing Levels". *EOS DIS*. <http://earthdata.nasa.gov/data/standards-and-references/processing-levels>.
- NASA Spitzer Space Telescope. 2013. <http://www.spitzer.caltech.edu>.

- [The] National Academies. 2010. *Astro2010: The Astronomy and Astrophysics Decadal Survey*. Washington, DC : National Academies Press. http://sites.nationalacademies.org/bpa/BPA_049810.
- National Aeronautics and Space Administration, Goddard Space Flight Center. 2014. "BSC5P – Bright Star Catalog". <http://heasarc.gsfc.nasa.gov/W3Browse/star-catalog/bsc5p.html>.
- National Aeronautics and Space Administration, Goddard Space Flight Center. 2013a. "FITS Documentation". http://fits.gsfc.nasa.gov/fits_documentation.html.
- National Aeronautics and Space Administration, Goddard Space Flight Center. 2013b. "FITS World Coordinate System". http://fits.gsfc.nasa.gov/fits_wcs.html.
- National Aeronautics and Space Administration, Infrared Processing and Analysis Center. 2014a. "NASA Extragalactic Database (NED)". <http://ned.ipac.caltech.edu>.
- National Aeronautics and Space Administration, Infrared Processing and Analysis Center. 2014b. "NASA Exoplanet Archive". <http://www.ipac.caltech.edu/project/25>.
- National Aeronautics and Space Administration, Infrared Processing and Analysis Center. 2014c. "Two Micron All Sky Survey". <http://www.ipac.caltech.edu/2mass>.
- National Aeronautics and Space Administration, Jet Propulsion Laboratory. 2014. "NASA Planetary Data System". <http://pds.jpl.nasa.gov>.
- National Aeronautics and Space Administration, Mikulski Archive for Space Telescopes. 2013. "About MAST". <http://archive.stsci.edu/aboutmast.html>.
- [The] National Archives. 2013. "Domesday Book". <http://www.nationalarchives.gov.uk/domesday>.
- National Center for Biotechnology Information. 2013. "Taxonomy Database". <http://www.ncbi.nlm.nih.gov/taxonomy>.
- National Health and Medical Research Council. 2007. "Australian Code for the Responsible Conduct of Research". <http://www.nhmrc.gov.au/guidelines/publications/r39>.
- National Information Standards Organization. 2004. *Understanding Metadata*. Bethesda, MD : NISO Press.
- National Information Standards Organization. 2013. "Recommended Practice for Online Supplemental Journal Article Materials". <http://www.niso.org/publications/rp/rp-15-2013>.
- National Institutes of Health. 2003. "NIH Data Sharing Policy". http://grants2.nih.gov/grants/policy/data_sharing.
- National Institutes of Health. 2013. "NIH Public Access Policy". http://publicaccess.nih.gov/submit_process.htm.
- National Optical Astronomy Observatory. 2003. "NOAO Policies for the Allocation of Observing Time". <http://www.noao.edu/noaoprop/help/policies.html> - dr.
- National Optical Astronomy Observatory. 2013a. "AURA/NOAO Data Rights Policy". <http://www.noao.edu/noaoprop/help/datarights.html>.
- National Optical Astronomy Observatory. 2013b. "Data". <http://ast.noao.edu/data>.
- National Research Council. 1997. *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC : National Academies Press.

- National Research Council. 1999. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington, DC : National Academies Press.
- National Research Council. 2001. *The Internet's Coming of Age*. Washington, DC : National Academies Press.
- National Research Council. 2013. *Proposed Revisions to the Common Rule: Perspectives of Social and Behavioral Scientists: Workshop Summary*. Washington, DC : National Academies Press.
- National Science Board. 2005. "Long-Lived Digital Data Collections". <http://www.nsf.gov/pubs/2005/nsb0540>.
- National Science Foundation. 2010a. "NSF Data Management Plans". Washington, DC : NSF. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp-dmp.
- National Science Foundation. 2010b. "NSF Data Sharing Policy". Washington, DC : NSF. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp-VID4.
- Natural Reserve System, University of California. 2013. <http://nrs.ucop.edu>.
- Nature Staff. 2012. "Data Standards Urged". *Nature*, 492 (7427), p. 145. <http://www.nature.com/naturejobs/science/articles/10.1038/nj7427-145a>.
- Naylor, Bernard et Marilyn Geller. 1995. "A Prehistory of Electronic Journals: The EIES and BLEND Projects", in *Advances in Serials Management*, Marcia Tuttle et Karen D. Darling (éd.), p. 27-47. Greenwich, CT : JAI Press.
- Nelson, Bryn. 2009. "Data Sharing: Empty Archives". *NATNews*, 461 (7261), p. 160-163. <https://www.nature.com/news/2009/090909/full/461160a.html>.
- Nexleaf. 2013. <http://nexleaf.org/about-us-0>.
- Nguyễn, Tánh Tr nti n Khanh et Hi n Tr n Ti n Huy n Nguyễn. 2006. "Computer Translation of the Chinese Taisho Tripitaka". <http://www.daitangvietnam.com/Computer Translation of the Chinese Tripitaka.pdf>.
- Nielsen, Michael. 2011. *Reinventing Discovery: The New Era of Networked Science*. Princeton, NJ : Princeton University Press.
- Nisbet, Miriam M. 2005. "Library Copyright Alliance, Orphan Works Notice of Inquiry". <http://www.copyright.gov/orphan/comments/OW0658-LCA.pdf>.
- Normile, Dennis, Gretchen Vogel et Jennifer Couzin. 2006. "Cloning – South Korean Team's Remaining Human Stem Cell Claim Demolished". *Science*, 311, p. 156-157. <https://science.sciencemag.org/content/311/5758/156.abstract>.
- Norris, Ray, Heinz Andernach, Guenther Eichhorn, Françoise Genova, Elizabeth Griffin, Robert J. Hanisch, Ajit Kembhavi, Robert Kennicutt et Anita Richards. 2006. "Astronomical Data Management", in *Highlights of Astronomy, XXVth IAU General Assembly*, K. A. van der Hucht (éd.). Cambridge, UK : Cambridge University Press.
- Northover, J. Peter et Shirley M. Northover. 2012. "Applications of Electron Backscatter Diffraction (EBSD) in Archaeology", in *Historical Technology, Materials and Conversation: SEM and Micro-analysis*, Nigel Meeks (éd.). Londres : Archetype.
- [The] NSA files. 2013. *The Guardian*. <http://www.guardian.co.uk/world/the-nsa-files>.

- Nunberg, Geoffrey. 2009. "Google's Book Search: A Disaster for Scholars". *Chronicle of Higher Education*. <http://chronicle.com/article/Googles-Book-Search-A/48245>.
- Nunberg, Geoffrey. 2010. "Counting on Google Books". *Chronicle Review*. <https://chronicle.com/article/Counting-on-Google-Books/125735>.
- NVivo 10. 2013. "NVivo 10 Research Software for Analysis and Insight". http://www.qsrinternational.com/products_nvivo.aspx.
- O'Brien, Danny. 2004. "How to Mend a Broken Internet". *New Scientist*. <http://www.newscientist.com/article/mg18424736.100-how-to-mend-a-broken-internet.html>.
- Odlyzko, Andrew M. 2000. "The Internet and Other Networks: Utilization Rates and Their Implications". *Information Economics and Policy*, 12 (4), p. 341-365.
- Office of Scholarly Communication. 2013. "UC Open Access Policy". University of California. <http://osc.universityofcalifornia.edu/openaccesspolicy>.
- Office of Science and Technology Policy. 2013. "Expanding Public Access to the Results of Federally Funded Research". The White House. <http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>.
- Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization". *UCLA Law Review*, 57, p. 1701. <http://ssrn.com/abstract=1450006>.
- Olsen, Mark, Russell Horton et Glenn Roe. 2011. "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections". *Digital Studies / Le Champ Numérique*, 2 (1). http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/190.
- Olson, Gary M. et Judith S. Olson. 2000. "Distance Matters". *Human-Computer Interaction*, 15 (2), p. 139-178.
- Olson, Gary M., Ann Zimmerman et Nathan Bos. 2008. *Scientific Collaboration on the Internet*. Cambridge, MA : MIT Press.
- Open Access Infrastructure for Research in Europe (OpenAIRE). 2014. Frequently Asked Questions. <http://www.openaire.eu/en/support/faq>.
- Open Bibliography and Open Bibliographic Data. 2013. <http://openbiblio.net>.
- Open Biological, Open et Biomedical Ontology Foundry. 2013. <http://www.obofoundry.org>.
- Open Context. 2013. <http://opencontext.org>.
- Open Geospatial Consortium. 2014. <http://www.opengeospatial.org>.
- Open Data Commons. 2013. <http://opendatacommons.org>.
- Open Geospatial Consortium. 2014. <http://www.opengeospatial.org>.
- Open Researcher et Contributor ID. 2011. <http://www.orcid.org>.
- Oransky, Ivan. 2012. "Why Aren't There More Retractions in Business and Economics Journals?". *Retraction Watch*. <http://retractionwatch.wordpress.com/2012/12/12/why-arent-there-more-retractions-in-business-and-economics-journals>.

- Organisation for Economic Co-operation and Development. 2007. "OECD Principles and Guidelines for Access to Research Data from Public Funding"⁶. <http://www.oecd.org/dataoecd/9/61/38500813.pdf>.
- Osterlund, Carsten et Paul Carlile. 2005. "Relations in Practice: Sorting through Practice Theories on Knowledge Sharing in Complex Organizations". *Information Society* 21, p. 91-107.
- Ostrom, Elinor et Charlotte Hess. 2007. "A Framework for Analyzing the Knowledge Commons", in *Understanding Knowledge as a Commons: From Theory to Practice*, Charlotte Hess et Elinor Ostrom (éd.), p. 41-81. Cambridge, MA : MIT Press.
- Ostrom, Vincent et Elinor Ostrom. 1977. "Public Goods and Public Choices", in *Alternatives for Delivering Public Services: Toward Improved Performance*, E. S. Savas (éd.), p. 7-49. Boulder, CO : Westview Press.
- Overpeck, Jonathan T., Gerald A. Meehl, Sandrine Bony et David R. Easterling. 2011. "Climate Data Challenges in the 21st Century". *Science*, 331, p. 700-702. <https://science.sciencemag.org/content/331/6018/700>.
- Owen, Whitney J. 2004. "In Defense of the Least Publishable Unit". *Chronicle of Higher Education*. <http://chronicle.com/article/In-Defense-of-the-Least/44761>.
- Oxford English Dictionary. 2014. <http://www.oed.com>.
- Ozsoy, Selami. 2011. "Use of New Media by Turkish Fans in Sport Communication: Facebook and Twitter". *Journal of Human Kinetics*, 28, p. 165-176. https://www.researchgate.net/publication/236043755_Use_of_New_Media_by_Turkish_Fans_in_Sport_Communication_Facebook_and_Twitter.
- Paisley, William J. 1980. "Information and Work", in *Progress in the Communication Sciences*. vol. 2., Brenda Dervin et Melvin J. Voigt (éd.), p. 114-165. Norwood, NJ : Ablex.
- Paisley, William J. 1990. "The Future of Bibliometrics", in *Scholarly Communication and Bibliometrics*, Christine L. Borgman (éd.), p. 281-299. Newbury Park, CA : Sage.
- Palfrey, John G et Urs Gasser. 2012. *Interop: The Promise and Perils of Highly Interconnected Systems*. New York : Basic Books.
- PANGAEA. Data Publisher for Earth & Environmental Science. 2013. <http://www.pangaea.de>.
- Pan-STARRS. 2012. "Pan-STARRS Data Release of PS1 Surveys". http://ps1sc.org/Data_Release.shtml.
- Pan-STARRS. 2013a. "Camera Design – Pan-Starrs – Panoramic Survey Telescope & Rapid Response System". <http://pan-starrs.ifa.hawaii.edu/public/design-features/cameras.html>.
- Pan-STARRS. 2013b. <http://pan-starrs.ifa.hawaii.edu/public>.
- Parry, Marc. 2014. "How the Humanities Compute in the Classroom". *Chronicle of Higher Education*. <http://chronicle.com/article/How-the-Humanities-Compute-in/143809>.
- Parsons, M. A. et P. A. Fox. 2013. "Is Data Publication the Right Metaphor?". *Data Science Journal*, 12, WDS32-WDS46. <https://datascience.codata.org/articles/abstract/10.2481/dsj.WDS-042/>.

6. Organisation de coopération et de développement économiques (OCDE). 2007. « Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics ». <http://www.oecd.org/fr/science/inno/38500823.pdf>

- Paskin, Norman. 1997. "Information Identifiers". *Learned Publishing*, 10, p. 135-156.
- Paskin, Norman. 1999. "Toward Unique Identifiers". *Proceedings of the IEEE*, 87, p. 1208-1227.
- Pasquetto, Irene V., Christine L. Borgman et Morgan F. Wofford. 2019. "Uses and Reuses of Scientific Data: The Data Creators' Advantage". *Harvard Data Science Review*, 1 (2). <https://doi.org/10.1162/99608f92.fc14bf2d>.
- Pasquetto, Irene V., Bernadette M. Randles et Christine L. Borgman. 2017. "On the Reuse of Scientific Data". *Data Science Journal*, 16. <https://doi.org/10.5334/dsj-2017-008>.
- Pearson, Sarah Hinchcliff. 2012. "Three Legal Mechanisms for Sharing Data", in *For Attribution – Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*, Paul F. Uhler (éd.), p. 71-76. Washington, DC : National Academies Press. http://www.nap.edu/openbook.php?record_id=13564&page=71.
- Peng, Roger D. 2011. "Reproducible Research in Computational Science". *Science*, 334 (6060), p. 1226-1227. <https://science.sciencemag.org/content/334/6060/1226>.
- Pepe, Alberto. 2010. "Structure and Evolution of Scientific Collaboration Networks in a Modern Research Collaboratory". Los Angeles : University of California at Los Angeles. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1616935.
- Pepe, Alberto. 2011. "The Relationship between Acquaintanceship and Coauthorship in Scientific Collaboration Networks". *Journal of the American Society for Information Science and Technology*, 62 (11), p. 2121-2132. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21629>.
- Pepe, Alberto, Christine L. Borgman, Jillian C. Wallis et Matthew S. Mayernik. 2007. "Knitting a Fabric of Sensor Data and Literature", in *Information Processing in Sensor Networks*. New York : ACM/IEEE.
- Pepe, Alberto, Matthew S. Mayernik, Christine L. Borgman et Herbert Van de Sompel. 2010. "From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web". *Journal of the American Society for Information Science and Technology*, 61, p. 567-582. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21263>.
- Pepe, Alberto, Alyssa A. Goodman et August Muench. 2011. "The ADS All-Sky Survey". *arXiv:1111.3983*. <http://arxiv.org/abs/1111.3983>.
- Pepe, Alberto, Alyssa A. Goodman, August Muench, Mercè Crosas et Christopher Erdmann. À paraître. "Sharing, Archiving, and Citing Data in Astronomy". *PLoS ONE*. https://www.authorea.com/users/3/articles/288/_show_article.
- Perseus Digital Library. 2009. <http://www.perseus.tufts.edu/hopper>.
- Petersen, Arthur C. 2012. *Simulating Nature: a Philosophical Study of Computer-simulation Uncertainties and Their Role in Climate Science and Policy Advice*. Boca Raton, F L: CRC Press.
- Phelps, Thomas A. et Robert Wilensky. 1997. "Multivalent Annotations", in *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, Rachel Heery et Liz Lyon (éd.), p. 287-303. Londres : Springer-Verlag.
- Phelps, Thomas A. et Robert Wilensky. 2000. "Multivalent Documents". *Communications of the ACM*, 43, p. 83-90.

- Pienta, Amy M., George C. Alter et Jared A. Lyle. 2010. "The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data". http://deepblue.lib.umich.edu/bitstream/handle/2027.42/78307/pienta_alter_lyle_100331.pdf?sequence=1.
- Pine, Kathleen, Christine T. Wolf et Melissa Mazmanian. 2014. "The Work of Reuse: Quality Measurement in Healthcare Organizations". Paper presented at Workshop on Sharing, Re-Use, and Circulation of Resources in Cooperative Scientific Work, Baltimore, February.
- Pineda, Jaime E., Erik W. Rosolowsky et Alyssa A. Goodman. 2009. "The Perils of Clumpfind: The Mass Spectrum of Sub-structures in Molecular Clouds". *Astrophysical Journal*, 699 (2), L134-L138. <https://dash.harvard.edu/handle/1/4268284>.
- Pinter, Frances. 2012. "Open Access for Scholarly Books?". *Publishing Research Quarterly*, 28 (3), p. 183-191. <https://link.springer.com/article/10.1007/s12109-012-9285-0>.
- [The] Pisa Griffin Project. 2013. <http://vcg.isti.cnr.it/griffin>.
- Plan S: How important is open access publishing? 2019. Times Higher Education (THE). <https://www.timeshighereducation.com/features/plan-s-how-important-open-access-publishing>.
- Planck Collaboration. P.A.R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela *et al.* 2013. "Planck 2013 Results: Overview of Products and Scientific Results". *arXiv:1303.5062*. <http://arxiv.org/abs/1303.5062>.
- PLoS Medicine Editors. 2006. "The Impact Factor Game". *PLoS Medicine*, 3 (6). <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0030291>.
- Polanyi, Michael. 1966. *The Tacit Dimension*. Garden City, NY : Doubleday.
- Poovey, Mary. 1998. *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. Chicago : University of Chicago Press.
- Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ : Princeton University Press.
- Prayle, Andrew, Matthew N. Hurley et Alan R. Smyth. 2012. "Compliance with Mandatory Reporting of Clinical Trial Results on ClinicalTrials.gov: Cross Sectional Study". *British Medical Journal*, 344. <https://www.bmj.com/content/344/bmj.d7373>.
- Presner, Todd Samuel. 2010. "HyperCities: Building a Web 2.0 Learning Platform", in *Teaching Literature at a Distance*, Anastasia Natsina et Takis Tagialis (éd.). Londres : Continuum Books.
- Prewitt, Kenneth. 2013. "Is Any Science Safe?". *Science*, 340 (6132), p. 525. <https://science.sciencemag.org/content/340/6132/525>.
- Price, Derek J. de Solla. 1963. *Little Science, Big Science*. New York : Columbia University Press.
- Price, Derek J. de Solla. 1975. *Science since Babylon*. New Haven, CT : Yale University Press.
- Priem, Jason, Dario Taraborelli, Paul Groth et Cameron Neylon. 2010. "Altmetrics: A Manifesto – Altmetrics.org". <http://altmetrics.org/manifesto>.
- Pritchard, Sarah M., Larry Carver et Smiti Anand. 2004. "Collaboration for Knowledge Management and Campus Informatics". Santa Barbara : University of California, Santa Barbara. http://www.immagic.com/eLibrary/ARCHIVES/GENERAL/UCSB_US/S040823P.pdf.
- Project Bamboo. 2013. <http://projectbamboo.org>.

- Protein Data Bank. 2011. <http://www.rcsb.org/pdb>.
- PubMed Central. 2009. <http://www.ncbi.nlm.nih.gov/pmc>.
- Rabesandratana, Tania. 2013. "DrugWatchdog Ponders How to Open Clinical Trial Data Vault". *Science*, 339 (6126), p. 1369-1370. <https://science.sciencemag.org/content/339/6126/1369/tab-article-info>.
- Rahtz, Sebastian, Alexander Dutton, Donna Kurtz, Graham Klyne, Andrew Zisserman et Relja Arandjelovic. 2011. "CLAROS – Collaborating on Delivering the Future of the Past". <http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-224.xml;query=&brand=default>.
- Raymond, Eric S. 2001. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Cambridge, MA : O'Reilly.
- reCAPTCHA. 2013. "What Is reCAPTCHA?". <http://www.google.com/recaptcha/intro/index.html>.
- Reichman, Jerome H., Tom Dedeurwaerdere et Paul F. Uhler. 2009. *Designing the Microbial Research Commons: Strategies for Accessing, Managing, and Using Essential Public Knowledge Assets*. Washington, DC : National Academies Press.
- Reichman, O. J., Matthew B. Jones et Mark P. Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology". *Science*, 331, p. 703-705. <https://science.sciencemag.org/content/331/6018/703>.
- Reimer, Thorsten, Lorna Hughes et David Robey. 2007. "After the AHDS: The End of National Support?". *Arts-Humanities.net: Digital Humanities and Arts*. http://www.arts-humanities.net/forumtopic/after_ahds_end_national_support.
- Renear, Allen H. et Carole L. Palmer. 2009. "Strategic Reading, Ontologies, and the Future of Scientific Publishing". *Science*, 325, p. 828-832. <https://science.sciencemag.org/content/325/5942/828/tab-article-info>.
- Renear, Allen H., Simone Sacchi et Karen M. Wickett. 2010. "Definitions of Dataset in the Scientific and Technical Literature". *Proceedings of the 73rd ASIS&T Annual Meeting: Navigating Streams in an Information Ecosystem*, 47 (1), p. 1-4. <https://dl.acm.org/doi/10.5555/1920331.1920447>.
- Research Acumen. 2013. <http://research-acumen.eu>.
- Research Councils UK. 2011. "RCUK Common Principles on Data Policy". <http://www.rcuk.ac.uk/research/datapolicy>.
- Research Councils UK. 2012a. "Guidance for the RCUK Policy on Access to Research Output".
- Research Councils UK. 2012b. "RCUK Policy on Open Access and Supporting Guidance". <http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf>.
- http://roarmap.eprints.org/671/1/RCUK%20_Policy_on_Access_to_Research_Outputs.pdf.
- Research Councils UK. 2012c. "Research Councils UK Announces New Open Access Policy". <http://www.rcuk.ac.uk/media/news/120716>.
- Research Councils UK. 2013. "RCUK Policy on Open Access: Frequently Asked Questions". <http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/OpenaccessFAQs.pdf>.
- Research Data Alliance. 2013. <https://rd-alliance.org/node>.

- Ribes, David, Karen S. Baker, Florence Millerand et Geoffrey C. Bowker. 2005. "Comparative Interoperability Project: Configurations of Community, Technology, Organization". *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, p. 65-66. <https://dl.acm.org/doi/10.1145/1065385.1065399>.
- Ribes, David et Thomas Finholt. 2009. "The Long Now of Technology Infrastructure: Articulating Tensions in Development". *Journal of the Association for Information Systems*, 10 (5). <http://aisel.aisnet.org/jais/vol10/iss5/5>.
- Ribes, David et Steven J. Jackson. 2013. "Data Bite Man: The Work of Sustaining a Long-term Study", in *"Raw Data" Is an Oxymoron*, Lisa Gitelman (éd.), p. 147-166. Cambridge, MA : MIT Press.
- Ridge, Naomi A., James Di Francesco, Helen Kirk, Di Li, Alyssa A. Goodman, João F. Alves, Héctor G. Arce et al. 2006. "The COMPLETE Survey of Star-Forming Regions: Phase I Data". *Astronomical Journal*, 131 (6), p. 2921. <https://www.researchhub.com/paper/678813/the-complete-survey-of-star-forming-regions-phase-i-data>.
- Roberts, L. 1987. "Who Owns the Human Genome?". *Science*, 237 (4813), p. 358-361. <https://science.sciencemag.org/content/237/4813/358>.
- Roberts, L. 1990. "Genome Project: An Experiment in Sharing". *Science*, 248 (4958), p. 953. <https://science.sciencemag.org/content/248/4958/953>.
- Robertson, C. Martin. 1976. "The Beazley Archive". *American Journal of Archaeology*, 80 (4), p. 445. <https://www.jstor.org/stable/i221994>.
- Robey, David. 2007. "Consequences of the Withdrawal of AHDS Funding". Swindon, UK : Arts & Humanities Research Council.
- Robey, David. 2011. "Sustainability and Related Issues for Publicly Funded Data Resources", in *Evaluating & Measuring the Value, Use and Impact of Digital Collections*, Lorna Hughes (éd.). Londres : Facet Publishing.
- Robey, David. 2013. "Infrastructure Needs for the Digital Humanities". *e-Research South*. <http://www.eresearchsouth.ac.uk/news/infrastructure-needs-for-the-digital-humanities>.
- Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals". *American Sociological Review*, 15 (3), p. 351-357. <https://www.jstor.org/stable/2087176>.
- Rodriguez, Laura L., Lisa D. Brooks, Judith H. Greenberg et Eric D. Green. 2013. "The Complexities of Genomic Identifiability". *Science*, 339 (6117), p. 275-276. <https://science.sciencemag.org/content/339/6117/275.summary>.
- Rodriguez, Marko A., Johan Bollen et Herbert Van de Sompel. 2007. "A Practical Ontology for the Large-scale Modeling of Scholarly Artifacts and Their Usage", in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, p. 278-287. New York : ACM. <http://portal.acm.org/citation.cfm?id=1255229>.
- Rogers, Everett M. 1962. *Diffusion of Innovations*. New York : Free Press of Glencoe.
- Rosenberg, Daniel. 2013. "Data before the Fact", in *"Raw Data" Is an Oxymoron*, Lisa Gitelman (éd.), p. 15-40. Cambridge, MA : MIT Press.
- Rosenberg, Daniel. 2018. "Data as Word". *Historical Studies in the Natural Sciences*, 48 (5), p. 557-567. <https://doi.org/10.1525/hsns.2018.48.5.557>

- Rosenthal, David S. H. 2010. "Stepping Twice into the Same River". JCDL 2010 Keynote Address, Queensland, Australia. <http://blog.dshr.org/2010/06/jcdl-2010-keynot.html>.
- Rosenthal, David S. H. 2014. "Storage Will Be Much Less Free Than It Used to Be". Presented at the Seagate Corporation. <http://blog.dshr.org/2014/05/talk-at-seagate.html>.
- Rosolowsky, E. W., J. E. Pineda, J. B. Foster, M. A. Borkin, J. Kauffmann, P. Caselli, P. C. Myers et Alyssa Goodman. 2008. "An Ammonia Spectral Atlas of Dense Cores in Perseus". *Astrophysical Journal. Supplement Series*, 175 (2), p. 509-521. [https://www.research.manchester.ac.uk/portal/en/publications/an-ammonia-spectral-atlas-of-dense-cores-in-perseus\(8df74752-c8c5-4463-90d2-9037081cf652\)/export.html](https://www.research.manchester.ac.uk/portal/en/publications/an-ammonia-spectral-atlas-of-dense-cores-in-perseus(8df74752-c8c5-4463-90d2-9037081cf652)/export.html).
- Ross, Joseph S., Tony Tse, Deborah A. Zarin, Hui Xu, Lei Zhou et Harlan M. Krumholz. 2012. "Publication of NIH-funded Trials Registered in ClinicalTrials.gov: Cross Sectional Analysis". *British Medical Journal*, 344. <https://www.bmj.com/content/344/bmj.d7292>.
- Roth, Wendy D. et Jal D. Mehta. 2002. "The Rashomon Effect: Combining Positivist and Interpretivist Approaches in the Analysis of Contested Events". *Sociological Methods & Research*, 31 (2), p. 131-173. <https://journals.sagepub.com/doi/10.1177/0049124102031002002>.
- Rots, Arnold H. et Sherry L. Winkelman. 2013. "Observatory Bibliographies as Research Tools". Paper presented at American Astronomical Society, AAS Meeting #221. <http://adsabs.harvard.edu/abs/2013AAS...22115603R>.
- Rots, Arnold H., Sherry L. Winkelman et Glenn E. Becker. 2012. "Meaningful Metrics for Observatory Publication Statistics", in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 8448. <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8448/84480J/Meaningful-metrics-for-observatory-publication-statistics/10.1117/12.927134.short?SSO=1>.
- Ruhleder, Karen. 1994. "Rich and Lean Representations of Information for Knowledge Work: The Role of Computing Packages in the Work of Classical Scholars". *ACM Transactions on Information Systems*, 12 (2), p. 208-230. <https://dl.acm.org/doi/10.1145/196734.196746>.
- Ruixin, Yang. 2002. "Managing Scientific Metadata Using XML". *IEEE Internet Computing*, 6, p. 52-59.
- Ryan, Michael J. 2011. "Replication in Field Biology: The Case of the Frog-Eating Bat". *Science*, 334, p. 1229-1230. <https://science.sciencemag.org/content/334/6060/1229>.
- Sage Bionetworks. 2013. <http://sagebase.org>.
- Salaheldeen, Hany M. et Michael L. Nelson. 2013. "Resurrecting My Revolution", in *Research and Advanced Technology for Digital Libraries*, Trond Aalberg, Christos Papatheodorou, Milena Dobрева, Giannis Tsakonas et Charles J. Farrugia (éd.), p. 333-345. Berlin : Springer.
- Salerno, Emanuele, Anna Tonazzini et Luigi Bedini. 2007. "Digital Image Analysis to Enhance Underwritten Text in the Archimedes Palimpsest". *International Journal of Document Analysis and Recognition*, 9, p. 79-87. <https://link.springer.com/article/10.1007/s10032-006-0028-7>.
- Salk, Jonas. 1986. "Foreword", in *Laboratory Life: The Construction of Scientific Facts*, Bruno Latour et Steve Woolgar (éd.). 2^e éd. Princeton, NJ : Princeton University Press.

- Samuelson, Pamela. 2009a. "Legally Speaking: The Dead Souls of the Google Booksearch Settlement". <http://toc.oreilly.com/2009/04/legally-speaking-the-dead-soul.html>.
- Samuelson, Pamela. 2009b. "The Dead Souls of the Google Book Search Settlement". *Communications of the ACM*, 52, p. 28-30. <https://dl.acm.org/doi/10.1145/1538788.1538800>.
- Samuelson, Pamela. 2010. "Should the Google Book Settlement Be Approved?". *Communications of the ACM*, 53, p. 32-34.
- Sands, Ashley, Christine L. Borgman, Laura Wynholds et Sharon Traweek. 2014. "We're Working on It: Transferring the Sloan Digital Sky Survey from Laboratory to Library". Paper presented at International Conference on Digital Curation, San Francisco. <http://www.dcc.ac.uk/events/idcc14/programme>.
- Santer, B. D., T. M. L. Wigley et K. E. Taylor. 2011. "The Reproducibility of Observational Estimates of Surface and Atmospheric Temperature Change". *Science* 334, p. 1232-1233. <https://science.sciencemag.org/content/334/6060/1232.abstract>.
- Savage, Mike et Roger Burrows. 2007. "The Coming Crisis of Empirical Sociology". *Sociology*, 41, p. 885-899.
- Savage, Mike et Roger Burrows. 2009. "Some Further Reflections on the Coming Crisis of Empirical Sociology". *Sociology*, 43 (4), p. 762-772.
- Sawyer, Steve. 2008. "Data Wealth, Data Poverty, Science and Cyberinfrastructure". *Prometheus* 26 (4), p. 355-371. <https://www.tandfonline.com/doi/abs/10.1080/08109020802459348?journalCode=cpro20>.
- Schiffman, Lawrence H. 2002. "The Many Battles of the Scrolls". *Journal of Religious History*, 26, p. 157-178.
- Schiller, Dan. 2007. *How to Think About Information*. Urbana : University of Illinois Press.
- Schirrwagen, Jochen, Paolo Manghi, Natalia Manola, Lukasz Bolikowski, Najla Rettberg et Birgit Schmidt. 2013. "Data Curation in the OpenAIRE Scholarly Communication Infrastructure". *Information Standards Quarterly*, 25 (3), p. 13. <https://www.niso.org/niso-io/2013/09/data-curation-openaire-scholarly-communication-infrastructure>.
- Schmidt, Kjeld. 2012. "The Trouble with 'Tacit Knowledge'". *Computer Supported Cooperative Work*, 21 (2-3), p. 163-225. <https://link.springer.com/article/10.1007/s10606-012-9160-8>.
- Schnee, Scott, Thomas Bethell et Alyssa A. Goodman. 2006. "Estimating the Column Density in Molecular Clouds with Far-Infrared and Submillimeter Emission Maps". *Astrophysical Journal Letters* 640 (1), L47. <https://iopscience.iop.org/article/10.1086/50329292>.
- Schrier, Bill. 2011. "Bright Shiny Objects". *Chief Seattle Geek Blog*. <http://schrier.wordpress.com/tag/bright-shiny-objects>.
- Schroeder, Ralph. 2007. *Rethinking Science, Technology, and Social Change*. Stanford, CA : Stanford University Press.
- Schroeder, Ralph. 2014. "Big Data: Towards a More Scientific Social Science and Humanities?", in *Society and the Internet: How Information and Social Networks Are Changing Our Lives*, Mark Graham et William H. Dutton (éd.). Oxford : Oxford University Press.
- SciDrive. 2014. <http://www.scidrive.org>.

- Science Exchange Network. 2012. "The Reproducibility Initiative". <http://validation.scienceexchange.com/#/reproducibility-initiative>.
- Science Staff. 1987. "Ownership of the Human Genome". *Science*, 237 (4822), p. 1555. <https://science.sciencemag.org/content/237/4822/1555.1/tab-article-info>.
- Science Staff. 2011. "Challenges and Opportunities". *Science*, 331 (6018), p. 692-693. <https://science.sciencemag.org/content/331/6018/692>.
- Science Staff. 2012. "The Runners-Up". *Science*, 338 (6114), p. 1525-1532. <https://science.sciencemag.org/content/338/6114/1525.abstract>.
- Science Staff. 2013. "Random Sample: Was the Downfall of Louis XVI in His DNA?". *Science*, 340 (6135), p. 906-907. <https://science.sciencemag.org/content/340/6135/906.3/tab-article-info>.
- Scotchmer, Suzanne. 2003. "Intellectual property – When Is It the Best Incentive Mechanism for S&T Data and Information?", in *The Role of Scientific and Technical Data and Information in the Public Domain*, Julie M. Esanu et Paul F. Uhler (éd.), p. 15-18. Washington, DC : National Academies Press.
- Scroggins, Michael J. et Irene V. Pasquetto. 2020. "Labor Out of Place: On the Varieties and Valences of (In)visible Labor in Data-Intensive Science". *Engaging Science, Technology, and Society*, 6 (0), p. 111-132. <https://doi.org/10.17351/ests2020.341>.
- Scroggins, Michael J., Irene V. Pasquetto, R. Stuart Geiger, Bernadette M. Boscoe, Peter T. Darch, Charlotte Cabasse-Mazel, Cheryl Thompson, Milena S. Golshan et Christine L. Borgman. 2020. "Thorny problems in data (-intensive) science". *Communications of the ACM*, 63 (8), p. 30-32. <https://doi.org/10.1145/3408047>.
- Segal, Judith. 2005. "When Software Engineers Met Research Scientists: A Case Study". *Empirical Software Engineering*, 10, p. 517-536. <https://link.springer.com/article/10.1007/s10664-005-3865-y>.
- Segal, Judith. 2009. "Software Development Cultures and Cooperation Problems: A Field Study of the Early Stages of Development of Software for a Scientific Community". *Computer Supported Cooperative Work*, 18 (5-6), p. 1-26. <https://link.springer.com/article/10.1007/s10660-009-9096-9>.
- Sensor Modeling Language. 2010. <http://vast.uah.edu/SensorML>.
- Servick, Kelly. 2013. "House Subpoena Revives Battle over Air Pollution Studies". *Science*, 341 (6146), p. 604. <https://science.sciencemag.org/content/341/6146/604/tab-article-info>.
- Shadish, William R., Thomas D. Cook et Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston : Houghton Mifflin.
- Shankar, Kalpana. 2003. "Scientific Data Archiving: The State of the Art in Information, Data, and Metadata Management". <http://works.bepress.com/borgman/234>.
- Shankland, Stephen. 2013. "Big Blue, Big Bang, Big Data: Telescope Funds Computing R&D". *CNET*. http://news.cnet.com/8301-11386_3-57572519-76/big-blue-big-bang-big-data-telescope-funds-computing-r-d.

- Shapley, Alan H. et Pembroke J. Hart. 1982. "World Data Centers". *Eos, Transactions, American Geophysical Union*, 63 (30), p. 585.
- Shatz, David. 2004. *Peer Review: A Critical Inquiry*. Lanham, MD : Rowman & Littlefield.
- Shea, Christopher. 2011. "Fraud Scandal Fuels Debate over Practices of Social Psychology". *Chronicle of Higher Education*. http://chronicle.com/article/As-Dutch-Research-Scandal/129746/?sid=wb&utm_source=wb&utm_medium=en.
- SHERPA/RoMEO. 2014. "Publisher Copyright Policies & Self-archiving". [http:// www.sherpa.ac.uk/romeo](http://www.sherpa.ac.uk/romeo).
- Shilton, Katie. 2011. "Building Values into the Design of Pervasive Mobile Technologies". Los Angeles : University of California, Los Angeles. <http://ssrn.com/paper=1866783>.
- Shotton, David. 2011. "Why Researchers Don't Publish Data". *Semantic Publishing*. <http://semanticpublishing.wordpress.com/2011/08/04/why-researchers-dont-publish-data>.
- Shotton, David. 2013. "SCoRO, the Scholarly Contributions and Roles Ontology". <http://www.essepuntato.it/lode/http://purl.org/spar/scoro>.
- Shotton, David, Katie Portwin, Graham Klyne et Alistair Miles. 2009. "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article". *PLoS Computational Biology*, 5 (4). <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000361>.
- Shuai, Xin, Alberto Pepe et Johan Bollen. 2012. "How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations". *PLoS ONE*, 7 (11). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0047523>.
- SIMBAD Astronomical Database. 2013. <http://simbad.u-strasbg.fr/simbad>.
- Siminovitch, Lou. 2012. "Small Science: Big Science Will Prevail". *Science*, 338 (6109), p. 882-883. <https://science.sciencemag.org/content/338/6109/882.3/tab-article-info>.
- Simonite, Tom. 2013. "Chinese Censorship of Twitter-Style Social Networks Weibo, Tencent, and Sina Offers Clues to Government Plans | MIT Technology Review". *MIT Technology Review*. <http://www.technologyreview.com/news/511011/social-media-censorship-offers-clues-to-chinas-plans>.
- Sloan Digital Sky Survey. 2013a. <http://www.sdss.org>.
- Sloan Digital Sky Survey. 2013b. "The Ninth Sloan Digital Sky Survey Data Release (DR9)". <http://www.sdss3.org/dr9>.
- Smith, Arfon M., Daniel S. Katz, Kyle E. Niemeyer, "FORCE11 Software Citation Working Group. 2016. Software citation principles". *PeerJ Computer Science*, 2. <https://doi.org/10.7717/peerj-cs.86>.
- Smith, David A., Anne Mahoney et Gregory R. Crane. 2002. "Integrating Harvesting into Digital Library Content", in *2nd ACM IEEE-CS Joint Conference on Digital Libraries*, p. 183-184. New York : ACM. <http://www.perseus.tufts.edu/Articles/oaishort.pdf>.
- Smith, Doug, Rosanna Xia et Rong Gong Lin. 2013. "Earthquake Risk: L.A. Formally Requests List of Concrete Buildings". *Los Angeles Times*. <http://articles.latimes.com/2013/oct/24/local/la-me-ln-earthquake-concrete-list-20131024>.

- Smith, Michael E., Gary M. Feinman, Robert D. Drennan, Timothy Earle et Ian Morris. 2012. "Archaeology as a Social Science". *Proceedings of the National Academy of Sciences of the United States of America*, 109 (20), p. 7617-7621. <https://www.pnas.org/content/109/20/7617>.
- Smith, T. R. et M. Zheng. 2002. "ADEPT Knowledge Organization Team. Structured Models of Scientific Concepts as a Basis for Organizing, Accessing and Using Learning Materials", in *Joint Conference on Digital Libraries*, 202. New York : ACM.
- Snow, Charles P. 1956. "The Two Cultures". *New Statesman*. <http://www.newstatesman.com/cultural-capital/2013/01/c-p-snow-two-cultures>.
- Sobel, Dava. 2007. *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. New York : Walker.
- Social Media Research Foundation. 2013. <http://smrfoundation.com>.
- Social Science Data Archive. 2014. <http://www.sscnet.ucla.edu/issr/da>.
- Social Science Research Network. 2014. <http://www.ssrn.com/en>.
- Society of American Archivists. 2009. "Orphan Works: Statement of Best Practices". <http://www.archivists.org/standards/OWBP-V4.pdf>.
- Software Heritage Foundation. 2019. <https://www.softwareheritage.org/>.
- Spitzer Science Center. 2013. "Spitzer Observing Rules". <http://ssc.spitzer.caltech.edu/warmmission/propkit/sor/15>.
- Square Kilometre Array. 2013. <http://www.skatelescope.org>.
- Srinivasan, Ramesh. 2013. "Re-thinking the Cultural Codes of New Media: The Question Concerning Ontology". *New Media & Society*, 15 (2), p. 203-223. <https://journals.sagepub.com/doi/10.1177/1461444812450686>.
- Srinivasan, Ramesh, Robin Boast, Katherine M. Becvar et Jonathan Furner. 2009. "Blobjects: Digital Museum Catalogs and Diverse User Communities". *Journal of the American Society for Information Science and Technology*, 60 (4), p. 666-678. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21027>.
- Srinivasan, Ramesh, Robin Boast, Jonathan Furner, and Katherine M. Becvar. 2009. "Digital Museums and Diverse Cultural Knowledges: Moving Past the Traditional Catalog". *Information Society*, 25 (4), p. 265-278. <https://www.tandfonline.com/doi/abs/10.1080/01972240903028714>.
- Stallman, Richard M. 2002. *Free Software, Free Society: Selected Essays of Richard M. Stallman*, Joshua Gay (éd.). Boston : The Free Software Foundation.
- Stanley, Barbara et Michael Stanley. 1988. "Data Sharing: The Primary Researcher's Perspective". *Law and Human Behavior*, 12 (1), p. 173-180.
- Star, Susan Leigh. 1983. "Simplification in Scientific Work: An Example from Neuroscience Research". *Social Studies of Science*, 13, p. 205-228.
- Star, Susan Leigh. 1989. "The Structure of Ill-structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving", in *Distributed Artificial Intelligence*, Les Gasser et M. Huhns (éd.), vol. 2, p. 37-54. San Mateo, CA : Morgan Kaufmann.

- Star, Susan Leigh. 1991. "Power, Technologies, and the Phenomenology of Standards: On Being Allergic to Onions", in *A Sociology of Monsters? Power, Technology and the Modern World*, John Law (éd.), p. 27-57. Oxford : Basic Blackwell.
- Star, Susan Leigh. 1995. "The Politics of Formal Representations: Wizards, Gurus and Organizational Complexity", in *Ecologies of Knowledge: Work and Politics in Science and Technology*, Susan Leigh Star (éd.), p. 88-118. Albany : State University of New York Press.
- Star, Susan Leigh. 1999. "The Ethnography of Infrastructure". *American Behavioral Scientist*, 43 (3), p. 377-391. <https://journals.sagepub.com/doi/10.1177/00027649921955326>.
- Star, Susan Leigh et James R. Griesemer. 1989. "Institutional Ecology, 'Translations,' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-1939". *Social Studies of Science*, 19, p. 387-420.
- Star, Susan Leigh et Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces". *Information Systems Research*, 7 (1), p. 111-134. <https://doi.org/10.1287/isre.7.1.111>.
- Star, Susan Leigh et Anselm Strauss. 1999. "Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work". *Computer Supported Cooperative Work*, 8 (1-2), p. 9-30. <https://link.springer.com/article/10.1023/A:1008651105359>.
- Star, Susan Leigh et Geoffrey C. Bowker. 2002. "How to Infrastructure", in *Handbook of New Media*, Leah A. Lievrouw et Sonia Livingstone (éd.), p. 151-162. Londres : Sage.
- Star, Susan Leigh, Geoffrey C. Bowker et Laura J. Neumann. 2003. "Transparency beyond the Individual Level of Scale: Convergence between Information Artifacts and Communities of Practice", in *Digital Library Use: Social Practice in Design and Evaluation*, Ann Peterson Bishop, Nancy Van House et Barbara P. Battenfield (éd.), p. 241-270. Cambridge, MA : MIT Press.
- Starke, Marcus. 2013. "The Bright Shiny Object Syndrome of Marketing: Avoid It or Assess It?". *Business Innovation from SAP*. <http://blogs.sap.com/innovation/sales-marketing/the-bright-shiny-object-syndrome-of-marketing-avoid-it-or-assess-it-024063>.
- Stata Data Analysis and Statistical Software (STATA). 2013. <http://www.stata.com>.
- Steve: The Museum Social Tagging Project. 2013. "Welcome to the Steve Project". <http://www.steve.museum>.
- Structural Genomics Consortium. 2013. "SGC Mission and Philosophy". http://www.thesgc.org/about/what_is_the_sgc.
- Stodden, Victoria C. 2009a. "Enabling Reproducible Research: Open Licensing for Scientific Innovation". *International Journal of Communications Law and Policy*, 13, p. 1-55. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1362040.
- Stodden, Victoria C. 2009b. "The Legal Framework for Reproducible Scientific Research: Licensing and Copyright". *Computing in Science & Engineering*, 11, p. 35-40. doi:10.1109/MCSE.2009.19.
- Stodden, Victoria C. 2010. "Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science". *Computing in Science & Engineering*, 12 (5), p. 8-12. <https://ieeexplore.ieee.org/document/5562471>.

- Stokes, Donald. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington, DC : Brookings Institution Press.
- Storer, Norman W. 1973. "The Normative Structure of Science", in *The Sociology of Science: Theoretical and Empirical Investigations*, Robert K. Merton (éd.), p. 267-278. Chicago : University of Chicago Press.
- Suber, Peter. 2012a. *Open Access*. Cambridge, MA: MIT Press.
- Suber, Peter. 2016. *Qu'est-ce que l'accès ouvert ?*, trad. M. Lebert, Marseille, OpenEdition Press.
- Suber, Peter. 2012b. "SPARC Open Access Newsletter". <http://www.earlham.edu/~peters/fos/newsletter/09-02-12.htm#uk-ec>.
- Sullivan, Woodruff T., Dan Werthimer, Stuart Bowyer, Jeff Cobb, David Gedye et David Anderson. 1997. "A New Major SETI Project Based on Project SERENDIP Data and 100,000 Personal Computers", in *IAU Colloq. 161: Astronomical and Biochemical Origins and the Search for Life in the Universe*, 1, p. 729. <http://adsabs.harvard.edu/abs/1997abos.conf.729S>.
- Summers, Lawrence. 2013. "The Buck Does Not Stop with Reinhart and Rogoff". *Financial Times*. <http://www.ft.com/cms/s/2/41d14954-b317-11e2-b5a5-00144feabdc0.html#axzz2chzkbMZX>.
- Svenonius, Elaine. 2000. *The Intellectual Foundation of Information Organization*. Cambridge, MA : MIT Press.
- Sweeney, Latanya. 2002. "K-anonymity: a Model for Protecting Privacy". *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), p. 557-570. <https://www.worldscientific.com/doi/abs/10.1142/S0218488502001648>.
- Szalay, Alexander. 2008. "Jim Gray, Astronomer". *Communications of the ACM*, 51, p. 59-65.
- Szalay, Alexander. 2011. "Cosmology: Science in an Exponential World". *TEDxCaltech*. <http://www.youtube.com/watch?v=hB92o4H46hc&NR=1>.
- Takakusu, Junjir et Kaikyoku Watanabe. 1932. *Taish Shinsh Daiz ky (The Buddhist Canon, New Compilation of the Taish [1912-1925] Era)*, vol. 1-85. Tokyo : Taish Issaiky Kank kai.
- Taper, Mark L. et Subhash R. Lele (éd.). 2004. "Models of Scientific Inquiry and Statistical Practice: Implications for the Structure of Scientific Knowledge", in *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, p. 17-50. Chicago : University of Chicago Press.
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff et Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions". *PLoS ONE*, 6 (6). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101>.
- Text Encoding Initiative. 2013. <http://www.tei-c.org>.
- Thelwall, Mike, Stefanie Haustein, Vincent Larivière et Cassidy R. Sugimoto. 2013. "Do Altmetrics Work? Twitter and Ten Other Social Web Services". *PLoS ONE*, 8 (5). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064841>.
- Thelwall, Mike, Liwen Vaughan et Lennart Bjorneborn. 2005. "Webometrics". *Annual Review of Information Science & Technology*, 39, p. 81-135.
- Thomas, Katie. 2013. "Breaking the Seal on Drug Research". *New York Times*. <http://www.nytimes.com/2013/06/30/business/breaking-the-seal-on-drug-research.html>.

- Thomson Reuters. 2013. "Unlocking the Value of Research Data: A Report from the Thomson Reuters Industry Forum". <http://researchanalytics.thomsonreuters.com/m/pdfs/1003903-1.pdf>.
- Tichenor, Philip J., George A. Donohue et Clarice N. Olien. 1970. "Mass Media and Differential Growth in Knowledge". *Public Opinion Quarterly*, 34, p. 158-170.
- Tomasello, Michael et Josep Call. 2011. "Methodological Challenges in the Study of Primate Cognition". *Science*, 334, p. 1227-1228. <https://science.sciencemag.org/content/334/6060/1227/tab-article-info>.
- Udomprasert, Patricia et Alyssa A. Goodman. 2012. "WWT Ambassadors: WorldWide Telescope for Interactive Learning", in *Annual Meeting of the American Astronomical Society*. Austin, TX : AAS.
- Uhlir, Paul F. 2006. "The Emerging Role of Open Repositories as a Fundamental Component of the Public Research Infrastructure", in *Open Access: Open Problems*, Giandomenico Sica (éd.). Monza : Polimettrica.
- Uhlir, Paul F. 2007. "Open Data for Global Science: A Review of Recent Developments in National and International Scientific Data Policies and Related Proposals". *Data Science Journal*, 6, p. 1-3. <http://www.codata.org/dsj/special-open-data.html>.
- Uhlir, Paul F. (éd.). 2012. *For Attribution – Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Board on Research Data and Information. Washington, DC : National Academies Press.
- Uhlir, Paul F. et Peter Schröder. 2007. "Open Data for Global Science". *Data Science Journal*, 6, p. 36-53. <http://www.codata.org/dsj/special-open-data.html>.
- Uhlir, Paul F. et Daniel Cohen. 2011. "Internal Document". Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences.
- UK Data Archive. 2014. <http://www.data-archive.ac.uk>.
- Unicode, Inc. 2013. "Chronology of Unicode Version 1.0". <http://www.unicode.org/history/versionone.html>.
- United Nations Educational, Scientific and Cultural Organization. 1970. "Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property". http://portal.unesco.org/en/ev.php-URL_ID=13039&URL_DO=DO_TOPIC&URL_SECTION=201.html.
- United Nations Educational, Scientific and Cultural Organization. 2013. "UNESCO Database of National Cultural Heritage Laws". <http://www.unesco.org/culture/natlaws>.
- United States Copyright Office. 2006. "Report on Orphan Works". Washington, DC : Register of Copyrights. <http://www.copyright.gov/orphan/orphan-report.pdf>.
- University of Arizona Science Mirror Lab. 2013. "Casting for GMT3 to Start!". *Steward Observatory Mirror Lab*. http://mirrorlab.as.arizona.edu/castings/projects/gmt/gmt3_casting.
- University of Maryland University Libraries. 2013. "Primary, Secondary and Tertiary Sources". <http://www.lib.umd.edu/ues/guides/primary-sources>.
- University of Oxford. 2013a. "Bodleian Libraries Launch Shakespeare's First Folio Online". *Oxford Thinking*. http://www.campaign.ox.ac.uk/news/first_folio.html.

- University of Oxford. 2013b. "CLAROS: The World of Art on the Semantic Web". <http://www.clarosnet.org/XDB/ASP/clarosHome/index.html>.
- University of Oxford Podcasting Service. 2011. *CLAROS – A Virtual Art Collection*. <http://podcasts.ox.ac.uk/claros-virtual-art-collection-video>.
- Unsworth, John, Paul Courant, Sarah Fraser, Mike Goodchild, Margaret Hedstrom, Charles Henry, Peter B. Kaufman, Jerome McGann, Roy Rosenzweig et Bruce Zuckerman. 2006. "Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for Humanities and Social Sciences". New York : American Council of Learned Societies. <http://www.acls.org/cyberinfrastructure/cyber.htm>.
- Urbina, Ian. 2013. "I Flirt and Tweet. Follow Me at #Socialbot". *New York Times*. <http://www.nytimes.com/2013/08/11/sunday-review/i-flirt-and-tweet-follow-me-at-social-bot.html>.
- US Department of Health and Human Services. 1979. *The Belmont Report*. Washington, DC : U.S. Government Printing Office. <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>.
- Van de Sompel, Herbert. 2013. "From the Version of Record to a Version of the Record" presented at the Coalition for Networked Information. http://www.youtube.com/watch?v=fhrGS-QbNVA&feature=youtube_gdata_player.
- Van de Sompel, Herbert et Carl Lagoze. 2009. "All Aboard: Toward a MachineFriendly Scholarly Communication System", in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Tony Hey, Stewart Tansley et Kristin Tolle (éd.), p. 1-8. Redmond, WA : Microsoft.
- Van de Sompel, Herbert, Robert Sanderson, Martin Klein, Michael L. Nelson, Bernhard Haslhofer, Simeon Warner et Carl Lagoze. 2012. "A Perspective on Resource Synchronization". *D-Lib Magazine*, 18 (9 / 10).
- Van House, Nancy A. 2004. "Science and Technology Studies and Information Studies". *Annual Review of Information Science & Technology*, 38, p. 3-86.
- Van Houweling, Molly Shaffer. 2009. "Author Autonomy and Atomism in Copyright Law". *Virginia Law Review*, 96 (549), p. 549-642.
- Van Noorden, Richard. 2013a. "Half of 2011 Papers Now Free to Read". *Nature*, 500 (7463), p. 386-387. <https://www.nature.com/news/half-of-2011-papers-now-free-to-read-1.13577>.
- Van Noorden, Richard. 2013b. "Open Access: The True Cost of Science Publishing". *Nature*, 495 (7442), p. 426-429. <https://www.nature.com/news/open-access-the-true-cost-of-science-publishing-1.12676>.
- Van Noorden, Richard. 2020. "Open-access Plan S to allow publishing in any journal". *Nature*. <https://doi.org/10.1038/d41586-020-02134-6>.
- Van Raan, Anthony F. J. 1988. *Handbook of Quantitative Studies of Science and Technology*. Amsterdam : Elsevier.
- Vandewalle, Patrick, Jelena Kovacevic et Martin Vetterli. 2009. "Reproducible Research in Signal Processing". *IEEE Signal Processing Magazine*, 26, p. 37-47. <https://ieeexplore.ieee.org/document/4815541>.

- Vardigan, Mary, Pascal Heus et Wendy Thomas. 2008. "Data Documentation Initiative: Toward a Standard for the Social Sciences". *International Journal of Digital Curation* 3 (1). <http://www.ijdc.net/article/view/66>.
- Vargas, Rodrigo, Alisha Glass, Mike Taggart, Kuni Kitajima, Michael Hamilton et Michael Allen. 2006. "Linking Minirhizotron Images to Soil Physical Properties and Microbial Diversity (TER 2)". *Center for Embedded Network Sensing*. <http://escholarship.org/uc/item/4312j473>.
- Vertesi, Janet et Paul Dourish. 2011. "The Value of Data: Considering the Context of Production in Data Economies", in *Computer Supported Cooperative Work*, p. 533-542. New York : ACM. <https://dl.acm.org/doi/10.1145/1958824.1958906>.
- Virtual International Authority File. 2013. <http://viaf.org>.
- Visual Resources Association. 2012. "Statement on the Fair Use of Images for Teaching, Research, and Study". <http://www.vraweb.org/organization/pdf/VRAFairUseGuidelinesFinal.pdf>.
- Vita, Silvio. 2003. "Printings of the Buddhist 'Canon' in Modern Japan", in *Buddhism Asia 1: Papers from the First Conference of Buddhist Studies Held in Naples*, Giovanni Verard et Silvio Vita (éd.), p. 217-245. Kyoto : ISEAS.
- Vogel, Gretchen et Jennifer Couzin-Frankel. 2013. "Europe Debates Ethics Reviews, Data Release". *Science*, 339 (6123), p. 1024. <https://science.sciencemag.org/content/339/6123/1024/tab-article-info>.
- Voltaire. 1759. *Candide: Or, All for the Best*. Londres : Printed for J. Nourse⁷.
- Von Ahn, Luis, Benjamin Maurer, Colin McMillen, David Abraham et Manuel Blum. 2008. "reCAPTCHA: Human-Based Character Recognition via Web Security Measures". *Science*, 321 (5895), p. 1465-1468. <https://science.sciencemag.org/content/321/5895/1465.abstract>.
- Waldrop, M. Mitchell. 2001. *The Dream Machine: J.C.R. Licklider and the Revolution That Made Computing Personal*. New York : Viking Penguin.
- Wallis, Jillian C. 2012. "The Distribution of Data Management Responsibility within Scientific Research Groups". University of California, Los Angeles. <http://gradworks.umi.com/35/13/3513859.html>.
- Wallis, Jillian C., Stasa Milojevic, Christine L. Borgman et William A. Sandoval. 2006. "The Special Case of Scientific Data Sharing with Education", in *Proceedings 69th Annual Meeting of the American Society for Information Science and Technology*. Medford, NJ : Information Today.
- Wallis, Jillian C., Christine L. Borgman, Matthew S. Mayernik, Alberto Pepe, Nithya Ramanathan et Mark Hansen. 2007. "Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries", in *Research and Advanced Technology for Digital Libraries*, Laszlo Kovacs, Norbert Fuhr et Carlo Meghini (éd.), vol. 4675, p. 380-391. Berlin : Springer.
- Wallis, Jillian C., Christine L. Borgman, Matthew S. Mayernik et Alberto Pepe. 2008. "Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research". *International Journal of Digital Curation*, 3. <http://www.ijdc.net/article/view/67>.

7. Voltaire. 1759. *Candide ou l'optimisme*. Genève

- Wallis, Jillian C., Matthew S. Mayernik, Christine L. Borgman et Alberto Pepe. 2010a. "Digital Libraries for Scientific Data Discovery and Reuse: From Vision to Practical Reality", in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, p. 333-340. New York : ACM. <https://dl.acm.org/doi/10.1145/1816123.1816173>.
- Wallis, Jillian C., Christine L. Borgman et Matthew S. Mayernik. 2010b. "Who Is Responsible for Data? A Case Study Exploring Data Authorship, Ownership, and Responsibility and Their Implications for Data Curation", in *6th International Digital Curation Conference*. Chicago : Digital Curation Center. <http://www.dcc.ac.uk/events/conferences/6th-international-digital-curation-conference>.
- Wallis, Jillian C. et Christine L. Borgman. 2011. "Who Is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility", *Proceedings of the American Society for Information Science and Technology*, 48 (1), p. 1-10.
- Wallis, Jillian C., Laura A. Wynholds, Christine L. Borgman, Ashley E. Sands et Sharon Traweek. 2012. "Data, Data Use, and Inquiry: A New Point of View on Data Curation". Full paper (long version) submitted to the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. <http://works.bepress.com/borgman/280>.
- Wallis, Jillian C., Elizabeth Rolando et Christine L. Borgman. 2013. "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology". *PLoS ONE*, 8 (7). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332>.
- Watson, J. 1990. "The Human Genome Project: Past, Present, and Future". *Science*, 248 (4951), p. 44-49. <https://science.sciencemag.org/content/248/4951/44/tab-article-info>.
- Weigelt, Johan. 2009. "The Case for Open-Access Chemical Biology. A Strategy for Pre-competitive Medicinal Chemistry to Promote Drug Discovery". *EMBO Reports*, 10 (9), p. 941-945. <https://www.embopress.org/doi/full/10.1038/embor.2009.193>.
- Weinberg, Alvin M. 1961. "Impact of Large-Scale Science on the United States". *Science* 134 (3473), p. 161-164. <https://science.sciencemag.org/content/134/3473/161>.
- Weinberger, David. 2012. *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room*. New York : Basic Books.
- Weller, Ann C. 2001. *Editorial Peer Review: Its Strengths and Weaknesses*. Medford, NJ : Information Today.
- Wellman, Barry et Caroline Haythornthwaite. 2002. *The Internet in Everyday Life*. Oxford : Blackwell.
- Wells, Don C., Eric W. Greisen et R. H. Harten. 1981. "FITS - A Flexible Image Transport System". *Astronomy & Astrophysics. Supplement Series*, 44, p. 363. <http://adsabs.harvard.edu/abs/1981A%26AS...44.363W>.
- Wenger, Etienne. 1998. *Communities of Practice: Learning, Meaning, and Identity*. New York : Cambridge University Press.
- Whalen, Maureen. 2009. "What's Wrong with This Picture? An Examination of Art Historians' Attitudes about Electronic Publishing Opportunities and the Consequences of Their Continuing Love Affair with Print". *Art Documentation* 28, p. 13-22.

- What's Invasive! 2010. <http://whatsinvasive.com>.
- White, Howard D. 1990. "Author Co-citation Analysis: Overview and Defense", in *Scholarly Communication and Bibliometrics*, Christine L. Borgman (éd.), p. 84-106. Newbury Park, CA : Sage.
- White, Richard L., Alberto Accomazzi, G. Bruce Berriman, Giuseppina Fabbiano, Barry F. Madore, Joseph M. Mazzarella, Arnold Rots, Alan P. Smale, Lombardi Storrie et Sherry L. Winkelman. 2009. "The High Impact of Astronomical Data Archives". *Astro2010: The Astronomy and Astrophysics Decadal Survey*.
- Whitley, Richard. 2000. *The Intellectual and Social Organization of the Sciences*. Oxford : Oxford University Press.
- Whitlock, Michael C. 2011. "Data Archiving in Ecology and Evolution: Best Practices". *Trends in Ecology & Evolution*, 26 (2), p. 61-65. <https://www.sciencedirect.com/science/article/abs/pii/S0169534710002697>.
- Whitlock, Michael C., Mark A. McPeck, Mark D. Rausher, Loren Rieseberg et Allen J. Moore. 2010. "Data Archiving". *American Naturalist*, 175, E45-E146. <https://www.journals.uchicago.edu/doi/full/10.1086/650340>.
- Wickham, Chris et Nigel Vincent. 2013. "Debating Open Access: Introduction". *Debating Open Access*, p. 4-12. http://issuu.com/thebritishacademy/docs/debating_open_access-introduction__.
- Wiener, Jon. 2002. "How the Critics Shot up Michael Bellesiles's Book Arming America". *Nation*, 275 (15), p. 28-32.
- Wieseler, Biete, Michaela F. Kerekes, Volker Vervoeelgyi, Natalie McGauran et Thomas Kaiser. 2012. "Impact of Document Type on Reporting Quality of Clinical Drug Trials: a Comparison of Registry Reports, Clinical Study Reports, and Journal Publications". *British Medical Journal*, 344. <https://www.bmj.com/content/344/bmj.d8141>.
- Wiesenthal, Joe. 2013. "REINHART AND ROGOFF: 'Full Stop,' We Made a Microsoft Excel Blunder in Our Debt Study, and It Makes a Difference". *Business Insider*. <http://www.businessinsider.com/reinhart-and-rogooff-admit-excel-blunder-2013-4>.
- Wilbanks, John. 2006. "Another Reason for Opening Access to Research". *British Medical Journal*, 333 (7582), p. 1306-1308. <https://www.bmj.com/content/333/7582/1306/rapid-responses>.
- Wilbanks, John. 2009. "I Have Seen the Paradigm Shift and It Is Us", in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Tony Hey, Stewart Tansle et Kristin Tolle (éd.), p. 209-214. Redmond, WA : Microsoft.
- Wilbanks, John. 2011. "Openness as Infrastructure". *Journal of Cheminformatics*, 3 (36). <https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-36>.
- Wilbanks, John. 2013. "Licence Restrictions: A Fool's Errand". *Nature*, 495 (7442), p. 440-441. <https://www.nature.com/articles/495440a>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*, 3, 160018. <http://dx.doi.org/10.1038/sdata.2016.18>.

- Williams, Antony J., John Wilbanks et Sean Ekins. 2012. "Why Open Drug Discovery Needs Four Simple Rules for Licensing Data and Models". *PLoS Computational Biology*, 8 (9). <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002706>.
- Willinsky, John. 2006. *The Access Principle: The Case for Open Access to Research and Scholarship*. Cambridge, MA : MIT Press.
- Willis, Craig, Jane Greenberg et Hollie White. 2012. "Analysis and Synthesis of Metadata Goals for Scientific Data". *Journal of the American Society for Information Science and Technology*, 63 (8), p. 1505-1520. <https://onlinelibrary.wiley.com/doi/10.1002/asi.22683>.
- Winkelman, Sherry L., Arnold Rots, Michael McCollough, Glenn Becker, Aaron Watry et Joan Hagler. 2009. "The Chandra Bibliography Cataloging System: A Scientific Research Aid", in *Chandra's First Decade of Discovery, Proceedings of the Conference Held 22-25 September, 2009 in Boston, MA*. <http://adsabs.harvard.edu/abs/2009cfdd.confE.207W>.
- Winkelman, Sherry L. et Arnold Rots. 2012a. "Observatory Bibliographies: Not Just for Statistics Anymore", in *Observatory Operations: Strategies, Processes, and Systems IV. Proceedings of the SPIE*, Alison B. Peck, Robert L. Seaman et Fernando Comerón (éd.), vol. 8448. Amsterdam : SPIE. <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8448/844829/Observatory-bibliographies-not-just-for-statistics-anymore/> 10.1117/12.925207.short?SSO=1.
- Winkelman, Sherry L. et Arnold Rots. 2012b. "The Chandra Observational Ontology: Tying the Threads Together", in *Astronomical Data Analysis Software and Systems XXI*, P. Ballester, D. Egret et N.P.F. Lorente (éd.), p. 241. San Francisco : Astronomical Society of the Pacific.
- Wiseman, James. 1964. "Archaeology and the Humanities". *Arion*, 3 (2), p. 131-142.
- Wofford, Morgan F., Bernadette M. Boscoe, Christine L. Borgman, Irene V. Pasquetto et Milena S. Golshan. 2020. "Jupyter notebooks as discovery mechanisms for open science: Citation practices in the astronomy community". *Computing in Science Engineering*, 22 (1), p. 5-15. <https://doi.org/10.1109/MCSE.2019.2932067>.
- Wohn, D. Yvette et Dennis Normile. 2006. "Korean Cloning Scandal: Prosecutors Allege Elaborate Deception and Missing Funds". *Science*, 312, p. 980-981.
- Wolf, Susan M., George J. Annas et Sherman Elias. 2013. "Patient Autonomy and Incidental Findings in Clinical Genomics". *Science*, 340 (6136), p. 1049-1050. <https://science.sciencemag.org/content/early/2013/05/15/science.1239119/tab-article-info?versioned=true>.
- Wood, John, Thomas Andersson, Achim Bachem, Christopher Best, Françoise Genova, Diego R. Lopez, Wouter Los et al. 2010. *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*. Final report of the High Level Expert Group on Scientific Data. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>.
- Woolgar, Steve. 1988. *Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge*. Londres : Sage.
- World Internet Project. 2013. <http://www.worldinternetproject.net>.
- WorldWide Telescope. 2012. <http://www.worldwidetelescope.org/Home.aspx>.
- Wouters, Paul, Anne Beaulieu, Andrea Scharnhorst et Sally Wyatt (éd.). 2012. *Virtual Knowledge: Experimenting in the Humanities and Social Sciences*. Cambridge, MA : MIT Press.

- Wuchty, Stefan, Benjamin F. Jones et Brian Uzzi. 2007. "The Increasing Dominance of Teams in the Production of Knowledge". *Science*, 316 (5827), p. 1036-1039.
- Wynholds, Laura A., David S. Fearon, Christine L. Borgman et Sharon Traweek. 2011. "Awash in Stardust: Data Practices in Astronomy", in *iConference: Proceedings of the 2011 iConference*, p. 802-804. New York : ACM. <https://dl.acm.org/doi/10.1145/1940761.1940912>.
- Wynholds, Laura A., Jillian C. Wallis, Christine L. Borgman, Ashley E. Sands et Sharon Traweek. 2012. "Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices", in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, p. 19-22. New York : ACM. <https://dl.acm.org/doi/10.1145/2232817.2232822>.
- Xia, Rosanna, Doug Smith et Michael Finnegan. 2013. "UC Quake Researchers Refuse to Share Building Data with L.A". *Los Angeles Times*. <http://www.latimes.com/local/la-me-concrete-quake-20131019,0,1097898.story#axzz2IPPa1ctb>.
- Yan, Koon-Kiu et Mark Gerstein. 2011. "The Spread of Scientific Information: Insights from the Web Usage Statistics in PLoS Article-Level Metrics". *PLoS ONE*, 6 (5). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019917>.
- Yoffee, Norman et Severin Fowles. 2011. "Archaeology in the Humanities". *Diogenes*, 58 (1-2), p. 35-52. <https://journals.sagepub.com/doi/10.1177/0392192112441906>.
- Younis, Mohamed et Kemal Akkaya. 2008. "Strategies and Techniques for Node Placement in Wireless Sensor Networks: A Survey". *Ad Hoc Networks*, 6 (4), p. 621-655. <https://www.sciencedirect.com/science/article/abs/pii/S1570870507000984>.
- Zacchetti, Stefano. 2002. "An Early Chinese Translation Corresponding to Chapter 6 of the Petakopadesa: An Shigao's Yin Chi Ru Jing T 603 and Its Indian Original: a Preliminary Survey". *Bulletin of the School of Oriental and African Studies, University of London*, 65 (1), p. 74-98. <https://www.cambridge.org/core/journals/bulletin-of-the-school-of-oriental-and-african-studies/article/abs/an-early-chinese-translation-corresponding-to-chapter-6-of-the-petakopadesa-an-shigaos-yin-chi-ru-jing-t-603-and-its-indian-original-a-preliminary-survey/4AFF7A8884AD8B97EE6F23D0F97AF9B8>.
- Zacchetti, Stefano. 2005. *In Praise of the Light: A Critical Synoptic Edition with an Annotated Translation of Chapters 1-3 of Dharmarakṣa's Guang Zan Jing, Being the Earliest Chinese Translation of the Larger Prajñāpāramitā*. Open Research Centre Project. Tokyo : The International Research Institute for Advanced Buddhism, Soka University. http://iriab.soka.ac.jp/orc/Publications/BPPB/index_BPPB.html.
- Zappavigna, Michele. 2011. "Ambient Affiliation: A Linguistic Perspective on Twitter". *New Media & Society*, 13 (5), p. 788-806. <https://journals.sagepub.com/doi/10.1177/1461444810385097>.
- ZENODO. 2013. Frequently Asked Questions. <http://zenodo.org/faq>.
- Zerhouni, Elias A. 2006. "Report on the NIH Public Access Policy". Department of Health and Human Services, National Institutes of Health. http://publicaccess.nih.gov/Final_Report_20060201.pdf.

- Zhang, Guo, Ying Ding et Staša Milojević. 2013. "Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content". *Journal of the American Society for Information Science and Technology*, 64 (7), p. 1490-1503. <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22850>.
- Zimmer, Michael. 2010. "“But the Data Is Already Public”: On the Ethics of Research in Facebook". *Ethics and Information Technology*, 12 (4), p. 313-325. <https://link.springer.com/article/10.1007/s10676-010-9227-5>.
- Zimmerman, Ann S. 2003. "Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists". PhD diss., Ann Arbor, MI : University of Michigan.
- Zimmerman, Ann S. 2007. "Not by Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse". *International Journal on Digital Libraries*, 7, p. 5-16. <https://link.springer.com/article/10.1007/s00799-007-0015-8>.
- Zittrain, Jonathan. 2005. *The Future of the Internet – and How to Stop It*. Cambridge, MA : MIT Press.
- Zooniverse. 2014. <https://www.zooniverse.org/projects>.
- Zorich, Diane. 2008. "A Survey of Digital Humanities Centers in the United States". Washington, DC : Council on Library and Information Resources. <http://www.clir.org/pubs/reports/pub143/contents.html>.
- Zotero. 2013. <http://www.zotero.org>.
- Zotero. 2014. "Zotero Style Repository". <https://www.zotero.org/styles>.

Remerciements

Il faut tout un village pour écrire un livre, surtout lorsqu'il couvre des disciplines et des sujets aussi variés que le présent ouvrage. Mes écrits s'appuient sur le travail d'un grand village de collègues d'horizons divers, un « collège invisible » comme on l'appelle dans la communication savante. Les scientifiques se soucient plus que tout de leurs données et ont consacré un temps généreux à débattre, à participer à des séminaires et ateliers et à lire de nombreux brouillons de mes chapitres.

La genèse de ce projet d'ouvrage remonte à trop loin pour lister toutes celles et tous ceux qui ont alimenté mes réflexions. C'est pourquoi je ne peux ici remercier, au mieux, que celles et ceux qui ont touché, d'une manière ou d'une autre, les mots sur les pages du présent volume. Bien d'autres sont cités dans l'abondante bibliographie. J'ai certainement omis de mentionner plusieurs d'entre vous avec qui j'ai partagé des discussions mémorables sur les sujets qui m'occupent ici.

J'ai commencé mes recherches sur les pratiques scientifiques en matière de données à la fin des années 1990, en m'appuyant sur des travaux antérieurs sur les bibliothèques numériques, le comportement informationnel, l'interaction personne-machine, la recherche d'information et la communication informationnelle. J'ai conduit ces recherches avec un formidable ensemble de collaborateurs et collaboratrices, dont les contributions prolifiques à ma réflexion sont trop empreintes de connaissances tacites pour que je les explicite ici. Nos travaux en commun sont cités tout au long du présent ouvrage. Beaucoup de membres du corps professoral, d'étudiantes et étudiants et de chercheurs et chercheuses postdoctoraux ont participé à plusieurs projets; c'est pourquoi ils sont énumérés dans une liste alphabétique unique. Parmi les projets de recherche sur les pratiques en matière de données, on compte l'Alexandria Digital Earth Prototype Project (ADEPT); le Center for Embedded Networked Sensing (CENS) ; la Cyberlearning Task Force; Monitoring, Modeling, and Memory ; Data Conservancy ; Knowledge Infrastructures ; et Long-Tail Research.

Parmi les professeures et professeurs ayant collaboré à ces projets, on compte Daniel Atkins, Geoffrey Bowker, Sayeed Choudhury, Paul Davis, Tim DiLauro, George Djorgovski, Paul Edwards, Noel Enyedy, Deborah Estrin, Thomas Finholt, Ian Foster, James Frew, Jonathan Furner, Anne Gilliland, Michael Goodchild, Alyssa Goodman, Mark Hansen, Thomas Harmon, Bryan Heidorn, William Howe, Steven Jackson, Carl Kesselman, Carl Lagoze, Gregory Leazer, Mary Marlino, Richard Mayer, Carole Palmer, Roy Pea, Gregory Pottie, Allen Renear, David Ribes, William

Sandoval, Terence Smith, Susan Leigh Star, Alex Szalay, Charles Taylor et Sharon Traweek. Parmi les étudiantes et étudiants, les chercheurs et chercheuses post-doctoraux et les personnels de recherche, on compte Rebekah Cummings, Peter Darch, David Fearon, Rich Gazan, Milena Golshan, Eric Graham, David Gwynn, Greg Janee, Elaine Levia, Rachel Mandell, Matthew Mayernik, Staša Milojević, Alberto Pepe, Elizabeth Rolando, Ashley Sands, Katie Shilton, Jillian Wallis et Laura Wynholds.

L'essentiel de ce livre a été imaginé et écrit lors de mon année sabbatique à l'université d'Oxford en 2012-2013. Mes collègues d'Oxford étaient des puits de science et d'idées nouvelles qui m'ont répondu avec enthousiasme lorsque je leur ai demandé : « en quoi consistent vos données ? ». Balliol College m'a généreusement accueillie en tant que chercheuse invitée à la chaire Oliver Smithies ; en parallèle, j'ai occupé des postes de chercheuse invitée à l'Oxford Internet Institute et à l'Oxford eResearch Centre. Les discussions que j'y ai eues avec des pontes comme avec des étudiants ont éclairé mes réflexions sur les données et tout ce qui s'ensuit : bouddhisme, cosmologie, Dante, génomique, chiralité, nanotechnologies, éducation, économie, humanités, philosophie, mathématiques, médecine, langues et littératures, informatique et bien plus encore. Le système des collèges d'Oxford rassemble autour d'une même table des gens qui, autrement, ne se seraient jamais rencontrés et auraient encore moins entrepris des recherches par-delà les frontières disciplinaires. Je serai éternellement reconnaissante à mes hôtes, Sir Drummond Bone, directeur de Balliol, et Nicola Trott, Senior Tutor ; William Dutton de l'Oxford Internet Institute ; David de Roure de l'Oxford eResearch Centre ; Sarah Thomas, bibliothécaire de Bodley. Kofi Agawu, Martin Burton, George et Carmella Edwards, Panagis Filippakopoulos, Marina Jirotko, Will Jones, Elena Lombardi, Eric Meyer, Concepcion Naval, Peter et Shirley Northover, Ralph Schroeder, Anne Trefethen et Stefano Zacchetti ont été mes compagnons au quotidien et une source d'inspiration.

Parmi les gens d'Oxford qui ont éclairé mes réflexions, peut-être plus qu'ils ne le croient, je compte William Barford, Grant Blank, Dame Lynne Brindley, Roger Cashmore, Sir Iain Chalmers, Carol Clark, Douglas Dupree, Timothy Endicott, David Erdos, Bertrand Fauchaux, James Forder, Brian Foster, John-Paul Ghobrial, Sir Anthony Graham, Leslie Green, Daniel Grimley, Keith Hannabus, Christopher Hinchcliffe, Wolfram Horstmann, Sunghee Kim, Donna Kurtz, Will Lanier, Chris Lintott, Paul Luff, Bryan Magee, Helen Margetts, Philip Marshall, Ashley Nord, Dominic O'Brien, Dermot O'Hare, Richard Ovenden, Denis Noble, Seamus Perry, Andrew Pontzen, Rachel Quarrell, David Robey, Anna Sander, Brooke Simmons, Rob Simpson, Jin-Chong Tan, Linnet Taylor, Rosalind Thomas, Nick Trefethen, David Vines, Lisa Walker, David Wallace, Jamie Warner, Frederick Wilmot-Smith et Timothy Wilson.

Je dois remercier plus particulièrement mes collègues qui ont contribué de manière substantielle aux études de cas des chapitres 5, 6 et 7. L'exemple de l'astronomie au chapitre 5 est largement fondé sur le travail d'Alyssa Goodman, du Harvard-Smithsonian Center for Astrophysics, et de ses collaborateurs, parmi lesquels Alberto Accomazzi, Merce Crosas, Chris Erdmann, Michael Kurtz, Gus Muench et Alberto Pepe. Cette étude de cas s'appuie également sur les recherches de l'équipe Knowledge Infrastructures de l'université de Californie à Los Angeles (UCLA). Les brouillons de ce chapitre ont également bénéficié de multiples relectures par la professeure Goodman et de révisions par d'autres astronomes ou historiens de l'astronomie, dont Alberto Accomazzi, Chris Lintott, Michael Kurtz, Patrick McCray et Brooke Simmons. Les astronomes George Djorgovski, Phil Marshall, Andrew Pontzen et Alex Szalay ont également aidé à éclaircir certains points scientifiques. Le cas des réseaux de capteurs au chapitre 5 s'appuie sur des travaux antérieurs sur le CENS. Le manuscrit a été révisé par des collaborateurs et collaboratrices et des chercheurs et chercheuses du CENS, dont David Caron, Eric Graham, Thomas Harmon, Matthew Mayernik et Jillian Wallis. La première étude de cas de sciences sociales du chapitre 6, sur la recherche sur Internet, se fonde sur des entretiens avec les chercheurs et chercheuses de l'Oxford Institute Research Grant Blank, Corinna di Gennaro, William Dutton, Eric Meye et Ralph Schroeder, qui ont bien voulu relire les brouillons de ce chapitre. La deuxième étude de cas, sur la sociologie des techniques, est fondée sur des travaux antérieurs menés avec des collaborateurs, tels qu'ils sont cités, et ont été revus par Matthew Mayernik et Jillian Wallis. L'étude de cas de sciences humaines au chapitre 7 a été réalisée pour le présent ouvrage. Celle de CLAROS s'appuie sur des entretiens et des éléments fournis par Donna Kurtz, de l'université d'Oxford, ainsi que sur des sources supplémentaires proposées par David Robet et David Shotton. L'analyse du Griffon de Pise repose sur des entretiens et des éléments fournis par Peter Northover, également d'Oxford, et des sources supplémentaires proposées par Anna Contadini de la SOAS à Londres. La dernière étude de cas, sur les études bouddhiques, doit tout à l'enseignement patient de Stefano Zacchetti, titulaire de la chaire Yehan Numata d'études bouddhiques à Oxford, qui m'a fait découvrir son sanctuaire. Des chercheurs et chercheuses en sciences humaines ont eu la bonté de réviser le chapitre 7, parmi lesquels Anna Contadini, Johanna Drucker, Donna Kurtz, Peter Northover, Todd Presner, Joyce Ray et David Robey.

Beaucoup d'autres m'ont fait bénéficier de leur expertise sur des sujets spécialisés. En matière biomédicale, il y a Jonathan Bard, Martin Burton, Iain Chalmers, Panagis Filippakopoulos et Arthur Thomas. Le docteur Filippakopoulos a relu les manuscrits de plusieurs chapitres. Sur les technologies internet et les mécanismes de citation, on compte Geoffrey Bilder, Blaise Cronin, David de Roure, Peter Fox, Carole Goble,

Peter Ingwersen, John Klensin, Carl Lagoze, Salvatore Mele, Ed Pentz, Herbert Van de Sompel et Yorick Wilks. Le chapitre 9 s'est vu amélioré grâce aux remarques de Blaise Cronin, Kathleen Fitzpatrick et John Klensin. Paul Edwards et Marilyn Raphael ont été mes consultants sur la modélisation climatique. Les sections sur la propriété intellectuelle et l'accès ouvert doivent beaucoup à mes discussions avec David Erdos, Leslie Green, Peter Hirtle, Peter Murray-Rust, Pamela Samuelson, Victoria Stodden et John Wilbanks. Christopher Hinchcliffe m'a aidé à affiner ma compréhension des ressources communes à la suite de mes discussions sur l'économie avec Paul David, James Forder et David Vines. De longs entretiens avec mes collaborateurs Geoffrey Bowker, Paul Edwards, Thomas Finholt, Steven Jackson, Cory Knobel et David Ribes ont influencé mes idées sur les infrastructures de la connaissance. De même, ma réflexion sur les politiques en matière de données doit beaucoup à ma participation au Board on Research Data and Information, au CODATA et à l'Electronic Privacy Information Center, ainsi qu'aux indications de Francine Berman, Clifford Lynch, Paul Uhler et Marc Rotenberg. Sur la question des bibliothèques et des archives, j'ai consulté Lynne Brindley, Johanna Drucker, Anne Gilliland, Margaret Hedstrom, Ann O'Brien, Susan Parker, Gary Strong et Sarah Thomas. Jonathan Furner a clarifié pour moi des concepts philosophiques que j'avais découverts lors de nombreuses conversations à Oxford. Will Jones m'a initiée aux complexités éthiques de la recherche sur les réfugiés. Abdelmonem Afifi, Mark Hansen et Xiao-Li Meng ont contribué à ma compréhension des risques statistiques en analyse de données. Clifford Lynch, Lynne Markus, Matthew Mayernik, Ann O'Brien, Katie Shilton et Jillian Wallis ont lu et commenté de larges portions du manuscrit, tout comme plusieurs évaluateurs anonymes recrutés par Margy Avery des éditions MIT Press.

Ce serait une négligence de ne pas citer le travail invisible de celles et ceux qui sont rarement mentionnés en tant qu'auteurs. Je pense aux organismes de financement et aux responsables de programme qui ont rendu ces travaux possibles. Au sein de la National Science Foundation, Daniel Atkins, Stephen Griffin et Mimi McClure ont particulièrement veillé au développement des recherches sur les données, la recherche et les infrastructures. Tony Hey et son équipe au sein de Microsoft Research se sont montrés collaboratifs et de bon conseil et nous ont apporté leur soutien financier à des moments critiques. J'adresse mes remerciements à Lee Dirks, Susan Dumais, Catherine Marshall, Catherine Van Ingen, Alex Wade et Curtis Wong de Microsoft Research. Josh Greenberg de la Sloan Foundation m'a octroyé des fonds, une liberté et des conseils dans mon étude des infrastructures de la connaissance. Sont également invisibles les nombreuses personnes qui m'ont invitée à parler de mon futur livre et celles qui ont assisté à ces communications. Je suis très reconnaissante de ces occasions de riches débats. Rebekah Cummings, Elaine Levia et Camille Mathieu ont établi l'énorme bibliographie, qui sera rendue publique dans un groupe Zotero

(Borgman *Big Data, Little Data, No Data*) à la publication du présent ouvrage, dans l'esprit de l'accès ouvert.

La dernière personne que je dois remercier, et non des moindres, est mon mari George Mood, qui a retravaillé le manuscrit et tout ce que j'ai publié depuis 1977. Il a cependant l'habitude d'effacer son nom de la section des remerciements. Pour une fois, que le travail invisible soit révélé au grand jour.

Sommaire

Note de la traductrice	7
Preface to the French translation of <i>Big Data, Little Data, No Data</i>	9
Préface à l'édition française	13
Préface	17
I. Données et recherche	21
1. Provocations	23
2. Qu'est-ce qu'une donnée ?	39
3. Le travail scientifique des données	55
4. La diversité des données	81
II. Études de cas en travail scientifique des données	111
5. Le travail scientifique des données dans les sciences exactes	113
6. Le travail scientifique des données dans les sciences sociales	161
7. Le travail scientifique des données dans les sciences humaines	203
III. Politiques et pratiques en matière de données	251
8. Le partage, la diffusion et la réutilisation des données	253
9. Reconnaissance, attribution et découvrabilité des données	295
10. Que garder et pourquoi ?	329
Bibliographie	349
Remerciements	413

Réalisé sur les presses de l'imprimerie Ciaco
en impression à la demande
depuis le mois de décembre 2020
à Louvain-la-Neuve, Belgique
N° d'imprimeur : 89470