Christine L. Borgman

Qu'est-ce que le travail scientifique des données ?
Big data, little data, no data

OpenEdition Press

10. Que garder et pourquoi ?

DOI : 10.4000/books.oep.14787
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2020
Date de mise en ligne : 18 décembre 2020
Collection : Encyclopédie numérique
EAN électronique : 9791036565410

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 éléphantesque.

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 ingénieures et ingénieurs 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 commodes, 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'efficience, 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 entretiendront 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*.