Christine L. Borgman

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

OpenEdition Press

9. Reconnaissance, attribution et découvrabilité des données

DOI : 10.4000/books.oep.14782
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2020
Date de mise en ligne : 18 décembre 2020
Collection : Encyclopédie numérique
ISBN électronique : 9791036565410

http://books.openedition.org

Référence électronique
BORGMAN, Christine L. *9. Reconnaissance, attribution et découvrabilité des données* In : *Qu'est-ce que le travail scientifique des données ? Big data, little data, no data* [en ligne]. Marseille : OpenEdition Press, 2020 (généré le 21 janvier 2021). Disponible sur Internet : <http://books.openedition.org/oep/14782>. ISBN : 9791036565410. DOI : https://doi.org/10.4000/books.oep.14782.

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-ICSTI[1], « 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 taraude 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 XIXe 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 XXe 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 citateur 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 adéquatement 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 1er 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 édifiée. Les nombreuses typologies ne peuvent être fondues 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 minimes 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 citateur ou la citatrice, 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 citateur 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 citateurs et citatrices 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, glosés, 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 employées 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, compilateurs 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 usagères 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 référençant 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.

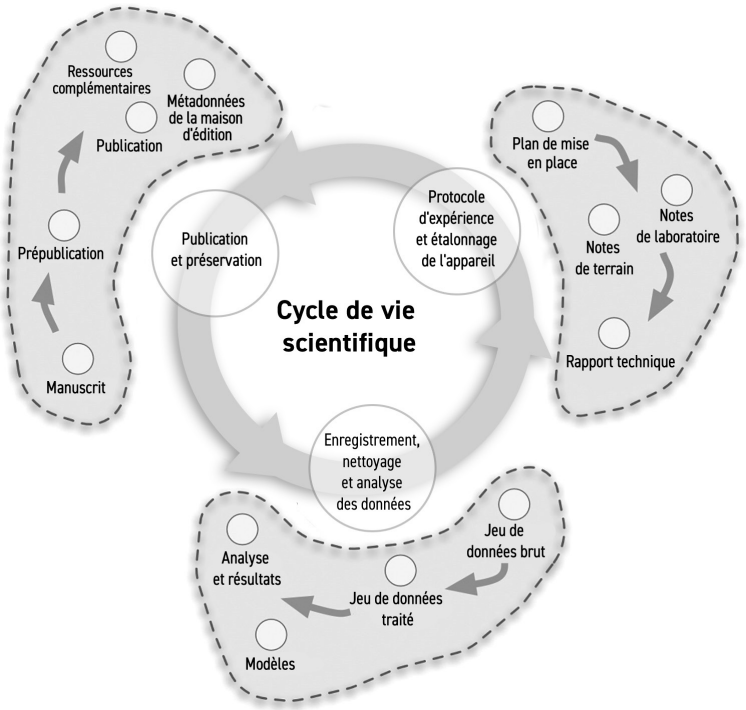

Figure 9.1. Exemple de cycle de vie scientifique du Center for Embedded Networked Sensing

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 inéprouvées.

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 reproposés 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.