Christine L. Borgman

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

OpenEdition Press

2. Qu'est-ce qu'une donnée ?

DOI : 10.4000/books.oep.14732
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2020
Date de mise en ligne : 18 décembre 2020
Collection : Encyclopédie numérique
EAN électronique : 9791036565410

2. Qu'est-ce qu'une donnée ?

Introduction

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

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

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

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

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

Définitions et terminologie

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

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

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

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

Définitions par l'exemple

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

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

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

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

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

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

Définitions opérationnelles

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

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

1. J. L Borges, *Œuvres complètes*, t. I, P. Bénichou, S. Bénichou-Roubaud, J.-P. Bernès et R. Caillois (traduction), Paris, Gallimard (La Pléiade), 1957, p. 749.
2. Pour la version française, voir Consultative Committee for Space Data Systems, 2017, p. 1-10.

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

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

Le modèle OAIS distingue donnée et information comme suit :

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

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

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

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

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

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

Définitions catégorielles

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

Degrés de traitement

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

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

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

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

Figure 2.1. Niveaux de traitement EOS DIS de la NASA

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

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

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

Origine et valeur de préservation

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

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

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

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

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

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

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

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

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

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

Ensembles de données

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

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

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

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

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

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

Distinctions conceptuelles

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

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

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

Sciences exactes et sciences sociales

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

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

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

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

Sciences humaines

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

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

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

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

Conclusion

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

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

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

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