À propos des données de recherche en SHS

Joachim Schöpfel

A propos des données de recherche en SHS

Joachim Schöpfel, Université de Lille

Un objet-frontière

S'il y a un consensus sur les données de recherche, c'est qu'il est quasi-impossible de s'accorder sur une seule définition, en particulier en SHS[1], où il s'agit en partie (souvent) d'artefacts ou de productions d'objets d'intérêt, « des 'construits' selon Bruno Latour, et même des 'capta' comme le souligne Joanna Drucker[2]. Pour les uns, tout est donnée ou pourrait l'être. Pour d'autres, il s'agit de la description élémentaire d'une réalité, à la base de toute information et connaissance[3]. Et d'autres encore, à l'instar d'un M. Jourdain qui fait de la prose sans le savoir, produisent et gèrent des données de recherche sans jamais prononcer ce terme ; ils évoquent plutôt des sources, témoignages, corpus, références, images etc.[4]

Le dépouillement des enquêtes ou l'analyse des répertoires d'entrepôts de données révèlent également une large diversité, en fonction des thématiques, disciplines, outils, méthodes et dispositifs ; parmi les différents types de données, citons par exemple les corpus de textes, photos, enregistrements audio-visuels, modèles 3D, tableaux statistiques et bases de données, logiciels, résultats d'enquêtes, partitions, graphiques, annotations, cartes etc. : tout cela fait partie des données issues de la recherche en SHS et stockées sur des serveurs, disques durs, clés USB ou ailleurs[5].

En principe, tout peut être ou devenir données même si aujourd'hui il s'agit avant tout de représentations d'objets numériques qui peuvent être traités par des machines et qui sont interopérables de manière à pouvoir être traités de façon durable dans des systèmes et des collections qui conservent encore leur trace d'origine (provenance) et des couches complexes de signification[6]. Certaines données sont relativement « simples » (statistiques etc.) mais la particularité des SHS est la construction et l'utilisation de données complexes (« agrégations spécialisées ») comme des éditions scientifiques, des corpus de textes, des textes structurés, des collections scientifiques thématiques, des données avec analyses et/ou annotations, ou encore des outils de découverte (bibliographies…). Une autre particularité est l'importance de la dimension interprétative pour l'encodage des données (dans une édition critique numérique, ce qui est retenu comme faisant partie du texte qu'il s'agit d'établir à partir de divers manuscrits par exemple), pour les métadonnées descriptives (qui dérivent des principes théoriques sur ce qu'est une édition ou des principes

---

[1] Borgman, C. (2016). *Big data, little data, no data: scholarship in the networked world*. Cambridge, Mass. : The MIT Press.

[2] Bachimont, B. (2017). L'archive et la massification des données : Une nouvelle raison numérique. *Gazette des Archives*, (245), 27–43. https://cours.ebsi.umontreal.ca/sci6116/Ressources_files/245_Gazette_Bachimont_2.pdf

[3] Abiteboul, S. (2012). *Sciences des données : de la logique du premier ordre à la Toile. Leçon inaugurale prononcée le jeudi 8 mars 2012*. Paris : Collège de France. https://doi.org/10.4000/books.cdf.506

[4] Malingre, M.-L., Mignon, M., Pierre, C., & Serres, A. (2019). Construction(s) et contradictions des données de recherche en SHS. *Recherche d'information, Document et Web Sémantique*, 2(1), 1–21. https://doi.org/10.21494/iste.op.2019.0336

[5] Prost, H., & Schöpfel, J. (2015). *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3.* https://hal.archives-ouvertes.fr/hal-01198379 Serres, A., Malingre, M.-L., Mignon, M., Pierre, C., & Collet, D. (2017). *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2*. https://hal.archives-ouvertes.fr/hal-01635186

[6] Flanders, J., & Munoz, T. (2019). An Introduction to Humanities Data Curation. In *DH Curation Guide: a community resource guide to data curation in the digital humanities*. https://guide.dhcuration.org/contents/intro/

éditoriaux), pour les annotations, etc. Une autre particularité est l'importance de la contextualisation des données en question, c'est-à-dire la connaissance précise de leur acquisition, production ou construction, la contribution et responsabilité intellectuelle des personnes à l'origine des données, l'historique des données (éditions, versions…) etc.

Néanmoins, la diversité de types de données et leurs particularités ne doivent pas cacher les enjeux communs, politiques, scientifiques, techniques et organisationnels. Peut-être faut-il considérer les données de recherche comme une sorte d'objet-frontière, au sens sociologique, suffisamment flexible et souple pour s'accommoder des besoins et usages des divers groupes qui l'utilisent, et assez « robuste » pour maintenir une certaine identité et cohésion entre les différentes disciplines ou communautés[7].

L'enquête sur les données de recherche en SHS à l'Université de Rennes 2 conclut que c'est le regard du chercheur qui donne sens et valeur à la donnée ; elle est construite par l'observateur[3]. Cependant, pour pouvoir communiquer et surtout, pour pouvoir développer des services de données utiles aux chercheurs, il faut un certain cadre partagé, une compréhension mutualisée qui peut servir de référence et rendent possibles « la constitution d'équivalences entre des mondes hétérogènes »[8].

Deux définitions de référence

Parmi un grand nombre de définitions, celle proposée par l'OECD en 2007 est devenue une sorte de référence : « *Les données de recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche* »[9]. La nouvelle Directive européenne sur les données ouvertes (Open Data Directive)[10] reprend les grandes lignes de cette définition, considérant dans l'article 2 comme données de recherche « *des documents se présentant sous forme numérique, autres que des publications scientifiques, qui sont recueillis ou produits au cours d'activités de recherche scientifique et utilisés comme éléments probants dans le processus de recherche, ou dont la communauté scientifique admet communément qu'ils sont nécessaires pour valider des conclusions et résultats de la recherche* ». Les deux définitions ont en commun qu'elles établissent un lien étroit entre les données, leur fonction pour le processus de recherche et les communautés scientifiques. Mais il y a d'autres facettes[11].

Cinq facettes majeures

Pour mieux comprendre les enjeux des données de recherche, il est utile de distinguer plusieurs aspects essentiels :

- Le lien avec le processus de recherche : certaines données sont collectées (« primaires »), d'autres sont produites (« secondaires ») ; mais elles servent toutes d'une manière ou d'une autre à valider les résultats. Ce lien n'est pas nouveau – l'analyse d'un corpus de documents

[7] Latzko-Toth, G., & Millerand, F. (2015). Objet-frontière. In F. Bouchard, P. Doray, & J. Prud'homme (Eds.), *Sciences, technologies et sociétés de A à Z* (pp. 163–165). Montréal : Les Presses de l'Université de Montréal. https://doi.org/10.4000/books.pum.4333

[8] Vinck, D. (2009). De l'objet intermédiaire à l'objet-frontière. Vers la prise en compte du travail d'équipement. *Revue d'anthropologie des connaissances*, 3(1), 51–72. https://www.cairn.info/revue-anthropologie-des-connaissances-2009-1-page-51.htm

[9] OECD (2007) http://www.oecd.org/fr/science/inno/38500823.pdf

[10] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

[11] Cf. aussi Schöpfel, J., Kergosien, E., & Prost, H. (2017). « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. *INFORSID 2017*, 31 Mai 2017, Toulouse. https://hal.archives-ouvertes.fr/hal-01530937

par exemple existe depuis longtemps, bien avant l'arrivée du numérique. Et même si l'essentiel du débat tourne aujourd'hui autour des données numériques, il ne faut pas perdre de vue la fonction spécifique de « preuve questionnable » au sein des différentes disciplines.

- Le lien avec la publication : dans la mesure où les données contribuent à produire de la connaissance, il y a un lien avec la publication des résultats, que ce soit sous forme d'annexes dans les thèses, de fichiers dans les articles enrichis ou de jeux de données déposés dans un entrepôt. Dans ce contexte on pourrait évoquer le développement des revues et articles de données (ou « data papers »), même s'ils sont encore peu nombreux en SHS[12].
- Le lien avec l'environnement scientifique : le concept de données est étroitement lié au contexte réel, à une communauté (« *data community* »[13]) autour d'un équipement, une infrastructure, une méthodologie, une thématique, un établissement etc. L'une des particularités des SHS est l'absence de grands équipements producteurs massifs de données comme des observatoires ou des accélérateurs de particules[14].
- La qualité : il s'agit d'une matière suffisamment qualifiée pour être utilisée ; en d'autres termes, tout peut être donnée mais toute donnée n'est pas utile pour la recherche. Il y a donc un choix à faire, une sélection, au sein d'un cadre dont la communauté concernée est le garant. Ceci concerne également la documentation du processus, c'est-à-dire la richesse des métadonnées, indispensable à la qualité des données elles-mêmes.
- La réutilisation : le dernier aspect est la réutilisation potentielle des données, l'intérêt pour d'autres chercheurs et d'autres recherches, issues de la même disciplines ou d'autres domaines scientifiques et sociaux. Certaines données ne peuvent pas être diffusées librement mais toutes devraient être partageables, dans des conditions à définir par les chercheurs (normes communautaires) et dans le cadre légal et réglementaire. Rappelons ici le principe d'ouverture et de partage des données « aussi ouvert que possible et aussi fermé que nécessaire » qui laisse l'appréciation du partage aux chercheurs et à leurs communautés et institutions.

Des enjeux multiples

Les enjeux des données de recherche sont multiples. Pour les chercheurs il s'agit de pouvoir (et savoir) produire, gérer, stocker et le cas échéant partager leurs données d'une manière performante, à la hauteur des exigences de la science, et dans le cadre des contraintes imposées par la loi, les institutions et les agences. Savoir gérer les données fait désormais partie des bonnes pratiques scientifiques, en particulier dans les projets d'envergure (H2020, ANR etc.).

Pour les professionnels de l'information il s'agit de développer des services de données performants et adaptés aux besoins des chercheurs – des sites d'information et de veille, des répertoires, une offre de formation, des entrepôts etc.[15] Il s'agit également de contribuer à la description et à la préservation de ces données et, d'une manière plus générale, à la normalisation des formats et outils. Contribuer à la réussite de ces projets pourrait être une opportunité pour ces professionnels,

---

[12] Schöpfel, J., Farace, D., Prost, H., & Zane, A. (2019). Data Papers as a New Form of Knowledge Organization in the Field of Research Data. *Colloque ISKO France 2019*, 9-11 Octobre 2019, Montpellier. https://hal.archives-ouvertes.fr/halshs-02284548

[13] Une analyse d'ITHAKA S+R de 2019 décrit la *data community* comme un réseau fluide et informel de chercheurs qui partagent et utilisent un certain type de données. https://sr.ithaka.org/publications/data-communities/

[14] C'est la raison pourquoi on parle aussi de la « longue traîne » des données en SHS : beaucoup de jeux de données hétérogènes, d'origine, de format et de nature très différents.

[15] Cf. Cat-OPIDoR, le wiki des services dédiés aux données de la recherche https://cat.opidor.fr/

qui, par exemple, pourraient aider à la rédaction des plans de gestion ou former les jeunes chercheurs aux bonnes pratiques[16].

Pour les responsables scientifiques et politiques, il s'agit de créer un environnement propice au développement de communautés scientifiques autour de certains types de données (textuels, objets, modélisations…) ou de certaines thématiques, afin de faciliter l'émergence de normes, outils et pratiques communautaires sur le terrain, dans une démarche « bottom-up ». Pour eux, dans un monde scientifique fortement interconnecté, il s'agit également d'assurer une certaine coordination, notamment par rapport à l'intégration dans les réseaux et infrastructures internationaux.

Il y a d'autres enjeux, économiques (réduire le risque de perdre des données), éthiques (contribuer à la transparence et l'intégrité de la recherche), informatiques (assurer l'interopérabilité des dispositifs), pour en citer quelques-uns. Les métadonnées sont l'un des facteurs clés pour le futur développement de la gestion et du partage des données de recherche[17]. Un autre facteur clé est la création de consortia avec des unités de recherche (laboratoires…), des services de données (SCD, INIST…), des producteurs de données (musées, archives…) et d'infrastructures (Huma-Num, Progedo…) capables à relever les défis de la standardisation, de la conservation et du développement de nouveaux services à forte valeur ajoutée (linked data, data mining etc.).

Mais l'attention à cet objet frontière permet de tenir à la fois ce qu'il y a de commun et ce qu'il y a de spécifique. En SHS la dimension critique et interprétative des données, les modalités particulières d'administration de la preuve en appui à un raisonnement ou démonstration donne lieu à des pratiques de recherche qui sont des formes de « gestion critiques des données » élaborées au fil du temps au sein disciplines et de leurs us et coutumes. La robustesse et la flexibilité nécessaires en mettant plus qu'ailleurs le chercheur au centre du dispositif.

[16] Schöpfel, J. (2018). *Vers une culture de la donnée en SHS*. Villeneuve d'Ascq : Université de Lille. https://hal.archives-ouvertes.fr/hal-01846849

[17] Cf. les principes FAIR http://www.donneesdelarecherche.fr/spip.php?mot124