LES DONNÉES DE RECHERCHE
Questions à Christine L. Borgman

Élise Lehoux

Christine L. Borgman est professeure en sciences de l'information et titulaire émérite d'une *Presidential Chair* dans la même discipline à l'UCLA. Spécialiste des données de la recherche, son ouvrage *Big Data, Little Data, No Data: Scholarship in the Networked World* (MIT Press, 2015) vient d'être traduit en français chez OpenEdition[1]. Elle répond aux questions d'Élise Lehoux, conservatrice des bibliothèques à l'université de Paris, à la suite d'une mission de deux mois passée dans un centre de recherche de l'université Harvard. Cet échange[2] vise à faciliter une contextualisation de la traduction française de son ouvrage, en soulignant les différences d'approches entre l'Europe et les États-Unis sur la question des données de la recherche.

Consulter le texte original :
https://www.enssib.fr/bibliotheque-numerique/documents/70118-research-data-questions-to-christine-l-borgman.pdf

[NDLR] Mise à jour du 12 janvier 2022

Solenn Bihan propose une traduction graphique de cet entretien dans une contribution au BBF, « Cherchez l'éléphant pour trouver les données : traduction graphique à partir de l'interview de Christine Borgman » : https://bbf.enssib.fr/matieres-a-penser/cherchez-l-elephant-pour-trouver-les-donnees-traduction-graphique-a-partir-de-l-interview-de-christine-borgman_70288

*

Élise Lehoux : Comment en êtes-vous arrivée à travailler sur les données de la recherche ? Quelles étaient les orientations épistémologiques lorsque vous avez commencé à travailler sur ce sujet ? Avez-vous observé des différences d'analyse entre les façons de l'envisager, en Europe et en Amérique du Nord ?

Christine L. Borgman : Le parcours qui m'a conduit à l'étude des données de recherche a été tortueux. Je me suis d'abord formée aux mathématiques et à la bibliothéconomie, commençant ma carrière dans l'informatisation des bibliothèques. L'étude de l'appréhension des systèmes de recherche d'informations par leurs utilisateurs m'a menée à entreprendre un doctorat en communication à l'université de

1 https://books.openedition.org/oep/14692?lang=fr. Voir aussi Élise LEHOUX, « Christine L. Borgman, "Qu'est-ce que le travail scientifique des données ? Big data, little data, no data" », *Bulletin des bibliothèques de France*, 22 mars 2021. En ligne : https://bbf.enssib.fr/critiques/qu-est-ce-que-le-travail-scientifique-des-donnees_69929

2 Traduction de l'anglais par Élise Lehoux.

Stanford, avec une double spécialisation en informatique et en sciences cognitives. La communication savante reste le fil rouge de ces décennies de travail, qui ont porté sur la recherche d'informations, la bibliométrie et les interfaces utilisateur dans les systèmes de recherche. Les chercheurs manipulent des données sous une multitude de formes, des inscriptions sur les tablettes cunéiformes aux photons détectés par les capteurs photographiques des télescopes spatiaux. Lorsque les données numériques sont devenues la monnaie d'échange de la recherche moderne, l'étude de la manière dont les gens acquièrent, traitent et interprètent les observations, afin de produire des données scientifiquement utiles, était une transition évidente – du moins, c'est ce qu'il me semble avec le recul.

On observe une grande variété d'orientations épistémologiques dans les manières dont les chercheurs travaillent avec les données. Les sciences humaines et les sciences diffèrent dans leurs questions de recherche et leurs méthodes, mais ce sont les variations à l'intérieur de chaque domaine qui sont les plus intéressantes. Les astronomes s'accordent sur l'existence d'un seul et même ciel comme principe opératoire, mais ils l'étudient avec un large éventail de technologies, de méthodes et de questions. Lors d'entretiens, les universitaires affirment souvent qu'ils suivent les pratiques usuelles de leur domaine, pourtant les approches varient d'une personne à l'autre. Dans une étude portant sur une collaboration multidisciplinaire pour la connaissance des fonds marins, nous avons constaté que les épistémologies ont évolué au fil du temps. Travaillant sur les mêmes paillasses de laboratoire, des chercheurs aux expertises complémentaires ont parfois obtenu le même résultat à travers des méthodes, des outils et des perspectives théoriques assez différents, influençant les perspectives de leurs collègues en cours de route (Darch & Borgman, 2016).

Élise Lehoux : Nous nous réjouissons de la traduction de votre ouvrage *Big data, little data, no data* en français chez OpenEdition. Depuis la publication de la version originale, avez-vous observé des changements dans la prise en compte organisationnelle ou politique de la gestion des données de la recherche au sein des pays ou des communautés de recherche que vous avez pu étudier ? Avez-vous vu émerger de nouvelles « provocations » – pour reprendre vos mots – ou points d'attention ou bien certains d'entre eux se sont-ils déplacés ?

Christine L. Borgman : Je suis également ravie que mon livre ait été traduit en français et soit en libre accès ; je suis reconnaissante au ministère d'avoir encouragé et financé ce projet[3]. Ayant collaboré avec des partenaires de recherche européens pendant une grande partie de ma carrière, je trouve que les différences les plus marquantes se situent au niveau des approches institutionnelles des universités, du financement et des politiques de recherche. En Europe, la gestion de l'université est plutôt centralisée à l'intérieur de chaque pays, associée à une coopération au niveau

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

européen. Aux États-Unis, il existe une forme de coordination au sein de certains États, comme le système de l'université de Californie, et au niveau des agences de financement nationales, mais le tout est superposé à un mélange complexe d'acteurs publics et privés. Chacune de ces approches présente bien sûr des avantages et des inconvénients. Il est plus aisé de mettre en œuvre des démarches de science ouverte dans les modèles centralisés.

Les six provocations présentées dans le chapitre I ont raisonnablement bien résisté à l'épreuve du temps. C'est sur la première provocation, qui concerne la reproductibilité, le partage et la réutilisation des données, que nous avons le plus progressé. Ce domaine continue d'être mon principal sujet de recherche. Les principes FAIR (*Findable, Accessible, Interoperable, Reusable*) pour les données de recherche, publiés l'année suivant la parution de mon livre (Borgman, 2015 ; Wilkinson *et al.*, 2016), ont accéléré ces tendances.

Les questions soulevées par mes deux dernières provocations, sur les infrastructures de la connaissance à court et à long terme, ont reçu le moins d'attention. À bien des égards, il s'agit des questions les plus épineuses que les acteurs de la recherche doivent aborder, car les infrastructures sont fragiles, et souvent vulnérables (Borgman *et al.*, 2016). Les questions économiques et politiques restent urgentes. Nous avons réexaminé ces problèmes d'infrastructures de la connaissance lors d'un atelier organisé au début de l'année dernière (Borgman *et al.*, 2020).

Élise Lehoux : En creux de votre ouvrage se situe l'activité des services d'accompagnement à la recherche – bibliothécaires, techniciens, ingénieurs – dans la sensibilisation et formation à la gestion des données, la curation des données, l'accompagnement des équipes de recherche. Quel rôle jouent ou devraient jouer les bibliothèques sur ces sujets dans les années à venir ? Comment et pourquoi œuvrer à rendre visible ce travail invisible ou invisibilisé ?

Christine L. Borgman : Les bibliothécaires, les archivistes et le personnel de soutien jouent en effet de nombreux rôles importants dans la gestion des données de recherche (GDR). L'invisibilité d'une grande partie de ce travail conduit à une sous-évaluation de leurs contributions.

Les professionnels de l'information ont plusieurs façons de rendre le travail de GDR plus visible. Ils peuvent être directement associés aux projets de recherche pour aider les équipes à gérer plus efficacement leurs données. Toutes les données ont besoin d'être gérées, qu'elles soient ou non partagées. Un autre moyen est de développer des modèles d'enseignement sur la GDR intégrés dans les formations disciplinaires des masters et des doctorats. Le fait de l'associer le plus tôt possible dans les carrières des chercheurs favorise un engagement à long terme.

Élise Lehoux : Le *data management plan* semble surtout envisagé comme un document administratif attendu par les financeurs. Cependant, en obligeant à considérer la vie des données tout au long d'un projet, ce document peut faire

émerger des questions relatives à sa méthodologie et à sa gestion. Comment voyez-vous son rôle et son éventuelle évolution ?

Christine L. Borgman : En effet, les plans de gestion des données sont trop souvent considérés comme un outil bureaucratique plutôt que comme une démarche constructive qui encouragerait les personnes à réfléchir au patrimoine que représentent leurs données.

Un entretien sur la gestion des données d'une heure entre un bibliothécaire et un chercheur est insuffisant pour créer une expertise partagée. Les données de recherche ne sont pas des documents génériques ; elles sont profondément imbriquées dans les connaissances disciplinaires. Les communautés pourraient tirer bénéfice en investissant dans le développement de profils de bibliothécaires spécialisés, de personnes diplômées et expérimentées dans un domaine qui deviennent des experts en information de cette même discipline. Une formation en physique est tout aussi nécessaire pour gérer des données astrophysiques qu'une formation en philologie l'est pour gérer des documents philologiques. La spécialisation des bibliothécaires est une vieille idée qui mérite d'être revisitée pour créer de nouvelles générations de professionnels des données et de l'information.

Élise Lehoux : Différentes initiatives ont vu le jour aux États-Unis dans les années 2010 pour développer la littératie des données. Pensez-vous que cela soit toujours un enjeu aujourd'hui ?

Christine L. Borgman : La littéracie des données d'hier est la science des données d'aujourd'hui, et elle est tout à fait au goût du jour. Les grandes universités américaines, dont l'université de Californie-Berkeley, le Massachusetts Institute of Technology (MIT) et l'université de Virginie ont créé des instituts de science de la donnée qui proposent des diplômes pour l'ensemble des cycles licence-master-doctorat (LMD). D'autres universités, comme l'UCLA, coordonnent les efforts déployés sur le campus pour que les différents départements et instituts proposent des parcours d'études en sciences de la donnée.

La science de la donnée d'aujourd'hui s'invite dans un large éventail de domaines relevant des sciences humaines et sociales et des sciences et techniques. Son étude se fait à l'aide de méthodes allant de l'anthropologie à l'épistémologie, en passant par les études critiques et statistiques. Le champ est désormais si vaste qu'il est plus facile de dire ce que la science des données n'est pas que ce qu'elle est, comme l'explique Xiao-Li Meng (2019) dans l'éditorial qui accompagne le lancement de la *Harvard Data Science Review*.

Élise Lehoux : Ouverture *versus* protection des données (RGPD en Europe), est-ce que l'on n'a pas ici un paradoxe autour de cette question d'ouverture des données de la recherche ?

Christine L. Borgman : Pour envisager conjointement l'ouverture et la protection des données, il faut porter une attention particulière au contexte et au *timing*.

Les pratiques de traitement des données relatives aux expériences sur la personne varient considérablement. Certaines données ne peuvent jamais être divulguées, tandis que d'autres peuvent être consultées ou réutilisées dans le cadre de protocoles appropriés, comme dans le cas des essais cliniques. Il est rare que les données de recherche soient « ouvertes, ouvertes, ouvertes », comme l'a rapporté l'un des participants dans notre étude sur une importante archive de données européenne (Borgman *et al.*, 2019). Au contraire, les données peuvent devenir ouvertes après un traitement suffisant, après des périodes d'embargo et associées à des articles scientifiques au moment de leur publication.

Le droit américain établit des distinctions importantes entre *informational privacy*, c'est-à-dire la protection des données concernant un individu, et l'*autonomy privacy*, c'est-à-dire la capacité d'un individu à conduire ses activités sans être observé. Ces distinctions sont utiles pour déterminer quelles données peuvent être divulguées, à qui et quand. Les tensions entre l'*informational privacy* et l'*autonomy privacy* sous-tendent des paradoxes apparents dans les environnements universitaires et de recherche (Borgman, 2018). Alors que nos agences de recherche biomédicale révisent leurs politiques de diffusion des données (National Academies of Sciences, 2021) et que la législation et les pratiques en matière de protection de la vie privée évoluent, des tensions apparaissent également entre les nombreuses épistémologies de la *privacy* à l'âge du numérique (Allen, 2021).

Élise Lehoux : Le mouvement de l'*Open Science* implique également une forme de normalisation des pratiques scientifiques par le biais de processus de standardisations. Quelles conséquences cela peut-il avoir sur les matériaux de la recherche et sur la façon de faire de la science ?

Christine L. Borgman : Les pressions visant à normaliser la pratique scientifique aux fins de la science ouverte sont controversées, comme vous le suggérez. Les normes d'échange de données favorisent la réutilisation, tandis que les normes appliquées trop strictement aux méthodes de recherche peuvent nuire à l'innovation. Le diable se cache dans les détails.

Élise Lehoux : J'ai vu aux États-Unis de nombreux mouvements pour favoriser la place des femmes dans les métiers de la *data*. Pouvez-vous nous parler de la place des femmes dans la science des données ?

Christine L. Borgman : Si l'on met de côté la difficulté de délimiter les contours de la science des données, que l'on évoquait plus haut, on peut remarquer que les conférences Women in Data Science (WiDS) se sont développées à l'échelle internationale, accueillant des dizaines d'événements rien qu'en 2021 (Women in Data Science Worldwide Initiative, 2021). Ces conférences attirent une diversité d'acteurs venant des universités, de l'industrie, du gouvernement et d'autres secteurs. Tout le monde peut y participer, mais tous les intervenants sont des femmes. La conférence

inaugurale que j'ai prononcée cette année lors du colloque WiDS organisé par l'université de Virginie a été une occasion bienvenue pour attirer l'attention de la communauté des sciences de la donnée sur des questions de sciences sociales (Borgman, 2021). Les vidéos et les supports de présentation de nombreux événements de 2021 – et des années précédentes – sont disponibles en ligne.

Élise Lehoux : On voit en France des réflexions émerger sur les données dites « négatives » ou non concluantes. Quelles questions voyez-vous également émerger dans le contexte nord-américain ?

Christine L. Borgman : Les questions sur comment, quand et pourquoi fournir un accès à des données nulles ou négatives sont omniprésentes dans le débat sur la communication savante. J'ai brièvement évoqué cette problématique dans mon livre, en présentant le concept de « no data », c'est-à-dire des cas où les données n'ont pas été enregistrées, publiées, ou n'ont pas fait l'objet d'une curation. Au cours de la dernière année, ces questions ont été soulevées à plusieurs occasions au sein des milieux scientifique, biomédical et des sciences humaines et sociales.

Pour simplifier à l'extrême un débat complexe, je propose ces quelques points :

- La publication de données non concluantes peut être bénéfique pour la communauté scientifique lorsqu'elle évite la duplication d'efforts aboutissant à des impasses.
- Les expériences qui aboutissent à des données non concluantes sont probablement beaucoup plus nombreuses que celles qui aboutissent à des résultats positifs.
- La publication de résultats non concluants nécessite des ressources comparables à celles de la publication de résultats positifs. Par conséquent, les auteurs et les éditeurs sont peu incités à investir leurs maigres ressources dans la publication de résultats non concluants.
- Le problème épistémologique majeur est de définir les résultats « non concluants ». Une avancée scientifique peut être le résultat de dizaines, de centaines ou de milliers de collectes de données effectuées au cours de nombreuses années (Strevens, 2020). Jusqu'au moment où l'accumulation des données permet de dégager un schéma d'ensemble significatif, chaque expérience individuelle a produit des résultats non concluants.
- La valeur aberrante, l'échec ou le résultat contradictoire peuvent eux-mêmes donner lieu à de l'innovation dans un second temps (Firestein, 2012).

Élise Lehoux : L'une des hypothèses principales de votre livre est que la « valeur des données réside dans leur usage ». Quelles formes de valorisation des données se développent actuellement en Amérique du Nord ?

Christine L. Borgman : La réponse la plus succincte à la question de savoir comment mesurer la valeur des données est que les données, en soi, ont peu de valeur.

Les tentatives d'évaluer les données en fonction de leur volume, leur variété, leur vitesse de production, ou de tout autre paramètre échouent car la valeur des données ne réside pas dans leurs octets mais dans leur contexte. Nous jugeons les données en fonction de ce que nous savons *à leur sujet*. Avons-nous confiance dans les personnes associées à la création de ces données ? À leur conservation ? À leur réutilisation ? Avons-nous confiance dans leur origine, parfois complexe ? Pouvons-nous inspecter ces données ? Ceux qui ont créé les données seront toujours ceux qui les connaîtront le mieux, et c'est là que reposent la confiance et la valeur (Pasquetto *et al.*, 2019). La capacité à réutiliser les données repose sur cette chaîne de valeur. •

RÉFÉRENCES

- Allen, A. L. (2021). HIPAA at 25-A Work in Progress. *New England Journal of Medicine*, 384 (23), 2169-2171. https://doi.org/10.1056/NEJMp2100900
- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.
- Borgman, C. L. (2018). Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier. *Berkeley Technology Law Journal*, 33 (2), 365-412. https://doi.org/10.15779/Z38B56D489
- Borgman, C. L. (2021). *Big Data, Little Data, or No Data? A Social Science Perspective on Data Science*. Women in Data Science. https://datascience.virginia.edu/pages/2021-women-data-science-charlottesville
- Borgman, C. L., Darch, P. T., Pasquetto, I. V., & Wofford, M. F. (2020). *Our knowledge of knowledge infrastructures: Lessons learned and future directions* (Alfred P. Sloan Foundation). University of California, Los Angeles. http://escholarship.org/uc/item/9rm6b7d4
- Borgman, C. L., Darch, P. T., Sands, A. E., & Golshan, M. S. (2016). The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. *Proceedings of the Association for Information Science and Technology*, 53, 1-10. http://dx.doi.org/10.1002/pra2.2016.14505301057
- Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, 70 (8), 888–904. https://doi.org/10.1002/asi.24172
- Darch, P. T., & Borgman, C. L. (2016). Ship space to database: Emerging infrastructures for studies of the deep subseafloor biosphere. *PeerJ Computer Science*, 2, e97. https://doi.org/10.7717/peerj-cs.97
- Firestein, S. (2012). *Ignorance: How It Drives Science*. Oxford University Press.
- Meng, X.-L. (2019). Data Science: An Artificial Ecosystem. *Harvard Data Science Review*, 1 (1). https://doi.org/10.1162/99608f92.ba20f892
- National Academies of Sciences. (2021, April 28). *Changing the Culture of Data Management and Sharing A Workshop*. https://www.nationalacademies.org/event/04-29-2021/changing-the-culture-of-data-management-and-sharing-a-workshop

- Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, 1 (2). https://doi.org/10.1162/99608f92.fc14bf2d
- Strevens, M. (2020). *The Knowledge Machine: How Irrationality Created Modern Science* (Illustrated edition). Liveright.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. http://dx.doi.org/10.1038/sdata.2016.18
- *Women in Data Science Worldwide Initiative*. (2021). Women in Data Science (WiDS) Conference. https://www.widsconference.org/