BBF • Contribution | 12 October 2021

RESEARCH DATA
Questions to Christine L. Borgman

Élise Lehoux

Christine L. Borgman is Distinguished Research Professor and Presidential Chair Emerita in Information Studies at the University of California, Los Angeles (UCLA). She is a specialist of research data and a French translation of her book *Big Data, Little Data, No Data: Scholarship in the Networked World* (MIT Press, 2015) has just been published by OpenEdition[1]. She is interviewed by Élise Lehoux, research data librarian at the University of Paris, at the occasion of a two-month residency at a Harvard University research center. This exchange aims to facilitate a contextualization of the French translation of her book, highlighting the differences in approach to research data between Europe and the United States.

For the French version:
https://bbf.enssib.fr/matieres-a-penser/les-donnees-de-recherche_70116

*

Élise Lehoux: How did you come to work on research data? What were the epistemological orientations of the researchers in this field when you started working on it? Have you observed any differences in the way research data is analyzed in Europe and North America?

Christine L. Borgman: My route to studying research data was circuitous, starting with degrees in mathematics and librarianship for a first career in library automation. Addressing the human side of information retrieval systems led me to the PhD in Communication at Stanford University, with specializations in computing and cognitive sciences. Scholarly communication is a common thread through my several decades of research, exploring information-seeking, bibliometrics, and user interfaces for search systems. Scholars handle data in myriad forms, from inscriptions on cuneiform tablets to photons detected by imagers on space telescopes. As digital data became the currency of modern scholarship, studying how people acquire, process, and interpret observations to produce scientifically useful data was an obvious transition – or so it appears in retrospect.

1 https://books.openedition.org/oep/14692?lang=fr. See also 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. Online : https://bbf.enssib.fr/critiques/qu-est-ce-que-le-travail-scientifique-des-donnees_69929

Researchers bring a fascinating variety of epistemological orientations to their data practices. The humanities and sciences differ in research questions and methods; it is the variance within fields that are most interesting. Astronomers agree on the existence of "a single sky" as an organizing principle, but they study that sky with a profound array of technologies, methods, and questions. In interviews, scholars often claim that they follow common practices of their field, and yet we find distinct approaches from one person to the next. In a study of a multidisciplinary collaboration on ocean floor science, we found that epistemologies evolved over time. Working side-by-side at lab benches, researchers with complementary expertise sometimes arrived at the same result via contrasting methods, tools, and theoretical perspectives, influencing their partners' perspectives along the way (Darch & Borgman, 2016).

Élise Lehoux: We are delighted that your book *Big data, little data, no data* has been translated into French by OpenEdition. Since the publication of the original version, have you observed any changes in the way the countries or research communities you have studied treat research data management, on a political or organizational plan? Have you seen any new "provocations" – to use your words – or points of attention emerge? Have some of them shifted?

Christine L. Borgman: I am equally delighted that my book was translated into French, as an open edition, and grateful to the Ministry[2] for suggesting and funding the project. Having been engaged in European research collaborations for most of my career, the most striking differences are in institutional approaches to universities, research funding, and policy. Europe is more centralized within countries, with pan-European cooperation. The U.S. has some coordination within states, such as the University of California system, and national funding agencies, all of which are layered on a complex mix of public and private stakeholders. Each of these approaches has advantages and disadvantages, of course. Open science is easier to implement in centralized models.

The six provocations in Chapter 1 have stood the test of time reasonably well. We have made the most progress on the first provocation, which addresses reproducibility, sharing, and reuse of data. This area continues to be my primary research focus. The FAIR principles (Findable, Accessible, Interoperable, Reusable) for research data, published the year after my book was published (Borgman, 2015 ; Wilkinson *et al.*, 2016), have accelerated these trends.

The latter two provocations, on knowledge infrastructures in the near and far terms, have received the least attention in the interim. In many ways, these are the most critical issues for stakeholders to address, as infrastructures are fragile, and often brittle (Borgman *et al.*, 2016). The economic and policy issues remain urgent.

2 The translation was funded by the Ministry of Higher Education, Research and Innovation as part of the National Plan for Open Science.

We revisited these knowledge infrastructure concerns in a workshop conducted early last year (Borgman *et al.*, 2020).

Élise Lehoux: Research support services – librarians, technicians, project officers – play an important role in the activities you analyze in your book, by raising awareness, providing training, managing and curating data, and assisting the research teams. What role do or should libraries play in these areas in the years to come? How and why should we endeavor to make this invisible (or invisibilized) work visible?

Christine L. Borgman: Librarians, archivists, and support staff indeed play many important roles in research data management (RDM). The invisibility of much of this work leads to under-valuing their contributions.

Information professionals can make RDM work more visible in several ways. We can partner with research groups to aid them in managing their own data more effectively. All data need to be managed, whether or not shared with others. Another way is to develop instructional models on RDM that can be incorporated into post-graduate education within individual fields. Partnering with researchers earlier in the careers promotes long-term engagement.

Élise Lehoux: The *data management plan* seems to be regarded, above all, as an administrative document needed to comply with the demands of the funders. However, because it requires envisioning the life of data throughout the duration of research projects, it can elicit some interesting questions concerning their methodology or their management. How do you consider the role of data management plans and their possible evolution?

Christine L. Borgman: Indeed, data management plans too often are a bureaucratic tool rather than a constructive mechanism to encourage people to think about their data assets.

A one-hour RDM interview between a librarian and a researcher is insufficient to create shared expertise. Research data are not generic documents; they are entities deeply seated in disciplinary knowledge. Communities can benefit by investing in subject librarianship, where individuals with degrees and experience in a domain become information experts in that domain. Physics training is as necessary to manage astrophysics data as is philology training to manage philological materials. Subject librarianship is an old idea worth revisiting to create new generations of data and information science professionals.

Élise Lehoux: There were many initiatives in the US in the 2010s to develop data literacy. Do you think this is still an issue today?

Christine L. Borgman: Yesterday's data literacy is today's data science, and it is very much au courant. Major universities in the U.S., including University of California-Berkeley, Massachusetts Institute of Technology (MIT), and the University of Virginia, have established entire schools of data science that offer undergraduate and post-graduate degrees. Other universities, such as UCLA, are coordinating efforts across the campus for individual departments and schools to offer data science curricula.

Today's data science cuts a broad swath across the sciences, social sciences, humanities, and technology. Approaches range from anthropological to epistemological to critical to statistical. The field is now so broad that it is easier to say what data science is not than what it is, as explained by Xiao-Li Meng (2019) in his opening editorial to launch the *Harvard Data Science Review*.

Élise Lehoux: Data openness *versus* data protection (RGPD in Europe): isn't there something paradoxical about these two needs for open research data?

Christine L. Borgman: In comparing open data and data protection, context and timing matter. Practices for handling human subjects data vary widely. Some data never can be released, while others can be viewed or reused under proper protocols, such as clinical trials. Rarely are research data "open, open, open" as one of our participants reported in a study of a major European data archive (Borgman *et al.*, 2019). Rather, data may become open after sufficient processing, after embargo periods, and with associated journal articles at the time of publication.

U.S. law makes important distinctions between "informational privacy", roughly information about oneself, and "autonomy privacy", roughly the ability not to be observed. These distinctions are useful in determining what data should be released, to whom, and when. Tensions between informational and autonomy privacy underlie these apparent paradoxes in university and research environments (Borgman, 2018). As our biomedical research agencies revise their data release policies (National Academies of Sciences, 2021), and as privacy law and practice evolves, tensions also are arising between the many epistemologies of privacy in our digital age (Allen, 2021).

Élise Lehoux: The Open Science movement also involves a form of normalization of scientific practices through standardization processes. What consequences might this have on research materials and on the way science is done?

Christine L. Borgman: Pressures to standardize scientific practice for the purposes of open science are controversial, as you suggest. Standards for data exchange promote reusability, whereas standards applied too strictly to research methods may undermine innovation. The devil is in the details.

Élise Lehoux: I have seen many movements in the U.S. that promote the place of women in data professions. Can you tell us about the place of women in data science?

Christine L. Borgman: Setting aside the challenge of scoping "data science", as discussed above, the Women in Data Science (WiDS) conferences have expanded internationally, hosting dozens of events in 2021 alone (Women in Data Science Worldwide Initiative, 2021). These conferences attract diverse participation from universities, industry, government, and other sectors. Anyone can attend, but all the speakers are women. My keynote this year, to the WiDS conference hosted by the University of Virginia, provided a welcome opportunity to engage the data science community in social science perspectives (Borgman, 2021). Videos and slides are available for many of the 2021 and earlier events.

Élise Lehoux: In France, reflections are developing on so-called negative or inconclusive data. Is this issue currently debated in the North American context? Or are there any other interesting emerging trends?

Christine L. Borgman: Questions about the how, when, and why of providing access to null or negative data pervade discussions of scholarly communication. I touched on these briefly in my book when presenting the concept of "no data", in which data were not captured, not released, or not curated. Over the course of the last year or so, these questions have arisen in venues of science, biomedical, humanities, and social sciences research.

To oversimplify a complex debate, I offer a few points:

- Releasing null data can benefit the community by avoiding duplication of efforts that result in dead ends.
- Experiments resulting in null data probably are far more common than are those resulting in positive findings.
- Publishing null findings requires comparable resources to publishing significant positive findings. As a consequence, authors and publishers have few incentives to invest scarce resources in publishing null findings.
- The overwhelming epistemological matter is defining "null" results. A single scientific breakthrough may be the result of tens, hundreds, or thousands of data collection efforts conducted over the course of many years (Strevens, 2020). Until the cumulative pattern becomes apparent, each of these experiments had null results.
- The outlier, the failure, or the contradictory result may itself become the innovation later (Firestein, 2012).

Élise Lehoux: One of the main hypotheses of your book is that the "value of data lies in its use". What forms of data valorization are currently developing in North America?

Christine L. Borgman: The most succinct answer to the question of how to measure the value of data is that data, per se, have little value. Attempts to weigh data by volume, variety, velocity or other parameter fail because the value of data lies not in their bits but in their context. We judge data by what we know *about* them. Do we trust the people associated with creating those data? Curating them? Reusing them? Do we trust the provenance chain? Can we inspect the data? Those who created the data always will know them best, and therein lies the trust and value (Pasquetto *et al.*, 2019). The ability to reuse data rests on that value chain. •

REFERENCES

- 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/