Christine L. Borgman

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

OpenEdition Press

Preface to the French translation of *Big Data, Little Data, No Data*

DOI : 10.4000/books.oep.14707
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

Preface to the French translation of *Big Data, Little Data, No Data*

The origin of this translation into French was the invitation from Marin Dacos and Françoise Genova to keynote the 2018 Paris conference launching the National Open Science Plan for France, subtitled "from Strategy to Action." This stimulating three-day event, organized in partnership with the Research Data Alliance, drew participants from France, Europe, and the U.S. to explore emerging issues around open access to publications, data, software, and other scientific research objects. Top-down approaches to open science, in the form of government science policies, met bottom-up approaches in the form of community practices. The ambitious National Plan combines country-specific actions with European initiatives: "France is committed to making scientific research results open to all – researchers, companies, citizens."

In the two years since that launch event, open science continues to evolve in concept, policy, and practice. Notions of "openness," whether referring to publications, data, software, or other entities, are as amorphous as ever. What is open to whom, when, why, and under what circumstances varies widely. Open access (OA) to publications is the usual starting point for open science, as is the case in the plan for France. OA publishing comes in many flavors, however. Some plans shift the costs to government funding agencies, some to universities, some to authors, and some to publishers. Other flavors promote preprint servers and institutional repositories as complements to paid subscriptions. Government initiatives such as *Plan S*, due to take effect in 2021, are controversial, resulting in complex compromises among stakeholders (Kwon, 2018; Noorden, 2020; *Plan S*, 2019).

The second pillar of most open science plans, including that of France, is open data. Access to research data, a central theme of this book, is yet more complex than OA publishing. Scholarly publishing has a long history, dating back millennia for books and centuries for journal articles. Data also have long histories, but more as process than as scholarly products to be exchanged. Research data can be embodied in artifacts, but they also can be abstractions or simulations. Almost any entity can be used as evidence of some phenomena. One person's signal is another's noise. Humans are in the loop throughout the entire lifecycle of data, from creation to curation to decay or disposal (Borgman, 2019).

In the time since this book's original publication in 2015, stakeholders have come to acknowledge the messiness of openness (Aspesi and Brand, 2020). While most disciplines now accept open access publishing, the means and adoption rates vary by domain, funding source, country, and other factors. More stakeholders recognize the messiness of scientific data, with advances in research on philosophical, epistemological, social, technical, and cultural aspects of data practices (Borgman, 2019; Drucker, 2014; Lane *et al.*, 2020; Leonelli, 2019a, 2019b; Pasquetto *et al.*, 2019; Rosenberg, 2018). The FAIR principles, first promulgated in 2016 (Wilkinson *et al.*, 2016), established a framework for disseminating data openly. These principles were quickly adopted into European science policy (European Union Publications Office, 2018). Practical, on-the-ground efforts to implement FAIR reveal the aspirational nature of the enterprise. For data to be *Findable*, communities must agree on methods for description, search, and retrieval. For data to be *Accessible* and *Interoperable*, stakeholders must agree on technical and legal frameworks. For data to be *Reusable*, originators must provide access to adequate documentation, and often to associated software, instrumentation, and other technologies. Reusing data is a goal more achievable than reproducible or replicable science, all of which remain contested concepts. The FAIR principles are subject to temporal factors. The longer the time from origin, the more difficult data are to find, access, interoperate, or reuse. Data creators retain significant advantages in the ability to reuse research data (Pasquetto *et al.*, 2017, 2019).

Several new journals devoted to interdisciplinary investigations of data have launched in the last five years, such as the *Harvard Data Science Review*, *Scientific Data*, and the *Journal of Data and Information Science*, plus countless special issues of discipline-specific journals. Interest in software preservation, software citation, and data citation continues to grow (Bouquin *et al.*, 2020; Davenport *et al.*, 2020; Smith *et al.*, 2016; *Software Heritage Foundation*, 2019; Wofford *et al.*, 2020). These venues, plus conferences, government reports, and funding initiatives serve to broaden the conversation about research data. Theoretical, technical, and practice topics abound, such as costs and benefits of data preservation, ethics and values of providing access to human subjects data, incentives and disincentives to share or reuse data, tradeoffs between launching new missions and preserving the data of current missions, how practices vary within and between domains, and whether data are best managed by universities, disciplinary repositories, government agencies, or commercial ventures, to name a few.

In parallel with the growth of research into data science is the expanding array of career tracks in data science, data management, and other areas of data practice. Universities in France, Europe, North America, Australasia, Asia, and elsewhere are investing in data science programs at the undergraduate and post-graduate levels. Some universities expect students in all fields to take at least one data science course as a

core requirement. These courses and degrees vary in theoretical, computational, and practical orientation. Many provide general knowledge in computing, statistics, digital scholarship, or related areas. Some are discipline-specific, such as bioinformatics. Yet other programs are professional, such as those in business, management, librarianship, and information sciences. Academic libraries and research institutions are hiring data management specialists. These data professionals play important roles in curating data, developing data archives, aiding researchers in managing their own data, and strategic planning for open science.

Perhaps the most significant advance in this time frame is broader recognition of the knowledge infrastructures (KI) in which scholarship occurs. As explained in Chapter 1 herein, KI encompass human, social, technical, policy, and institutional components of intellectual work and the many interactions between them. Infrastructures develop, evolve, and adapt in complex ways over long periods of time. To consider any component in isolation, whether data, publications, or individual systems, is to see but one feature of the elephant. Silos can emerge and become isolated. Components designed to interoperate seamlessly can become brittle, leading to catastrophic breakdowns. Robust networks can be self-healing, at least for awhile. Infrastructures are inherently fragile, requiring continual maintenance and regular repair. Durability is an accomplishment. Research data are especially fragile because they rarely stand alone. Lacking context, structure, or documentation, a dataset may be little more than a string of numbers. For data to be findable, accessible, interoperable, and reusable, myriad stakeholders must invest in KI components such as metadata, formats, archives, software preservation, data curation, skilled labor. In turn, these stakeholders must work together in ways that keep these KI components working together. As the array of players expands to include public, private, government, and niche partners, KI complexity increases. Infrastructures are notoriously invisible until they break down (Borgman *et al.*, 2020, 2016; Edwards, 2010; Edwards *et al.*, 2013; Mayernik *et al.*, 2017; Scroggins *et al.*, 2020; Scroggins & Pasquetto, 2020; Star & Ruhleder, 1996).

This new edition of *Big Data, Little Data, No Data: Scholarship in the Networked World* appears at a critical juncture in scholarship, open science, and knowledge infrastructures. Charlotte Matoussowsky is an exceptionally conscientious translator, peppering me with thoughtful questions throughout the process. May this volume provoke new conversations to advance our understanding, use, and reuse of research data for the generations to come.

Christine L. Borgman
UCLA, Los Angeles
November 13, 2020