SCIENCE in the Archives

Pasts, Presents, Futures

::

EDITED BY

LORRAINE DASTON

SCIENCE IN THE ARCHIVES

Pasts, Presents, Futures

Edited by Lorraine Daston

The University of Chicago Press
Chicago and London

The University of Chicago Press, Chicago 60637
The University of Chicago Press, Ltd., London

Published 2017.
Printed in the United States of America

26 25 24 23 22 21 20 19 18 17 1 2 3 4 5

ISBN-13: 978-0-226-43222-9 (cloth)
ISBN-13: 978-0-226-43236-6 (paper)
ISBN-13: 978-0-226-43253-3 (e-book)
DOI: 10.7208/chicago/9780226432533.001.0001

Library of Congress Cataloging-in-Publication Data
Names: Daston, Lorraine, 1951– editor.
Title: Science in the archives : pasts, presents, futures / edited by Lorraine Daston.
Description: Chicago ; London : The University of Chicago Press, 2017. | Includes bibliographical references and index.
Identifiers: LCCN 2016028698 | ISBN 9780226432229 (cloth : alk. paper) | ISBN 9780226432366 (pbk. : alk. paper) | ISBN 9780226432533 (e-book)
Subjects: LCSH: Scientific archives. | Scientific archives—History. | Science—History.
Classification: LCC Q224 .S35 2017 | DDC 026/.5—dc23 LC record available at https://lccn.loc.gov/2016028698

♾ This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

CONTENTS

SEVEN

The "Data Deluge": Turning Private Data into Public Archives

Bruno J. Strasser

The "data deluge" is an interesting metaphor (fig. 7.1). Widely used since the 1990s, it attempts to capture the process resulting in the current "data flood," the immense amount of digital data about nature, people, and societies. But one of the most puzzling connotations of this metaphor is that data, like rain in the most famous deluge of all, seems to pour down naturally, pulled only by the laws of gravity, submerging the earth. All those who have paid close attention to archives and databases have however seen the data deluge in a very different light. Data did not fall naturally upon them. Data was something that had to be actively sought out. Creating a flow of data from source toward repositories was far more challenging than just waiting for rain to fall from the heavens. This chapter aims to understand how such data flows were created and sustained in the late twentieth-century experimental life sciences. It argues that the data deluge was not simply the product of technological revolutions in the modes of data production and a general increase in the amount of data being produced ("big data"). Rather, these developments resulted from two historically significant transformations: a redefinition of what counts as "data" and also of the obligations attached to possessing "data." These changes deeply affected how data came to be collected. But they did not upset the existing moral economy of the experimental sciences, based on individual authorship, credit, and rewards. Instead, individual and collective interests aligned in new ways. This chapter addresses a crucial aspect of the sciences of the archives: how the archives became filled with data available for public use.[1]

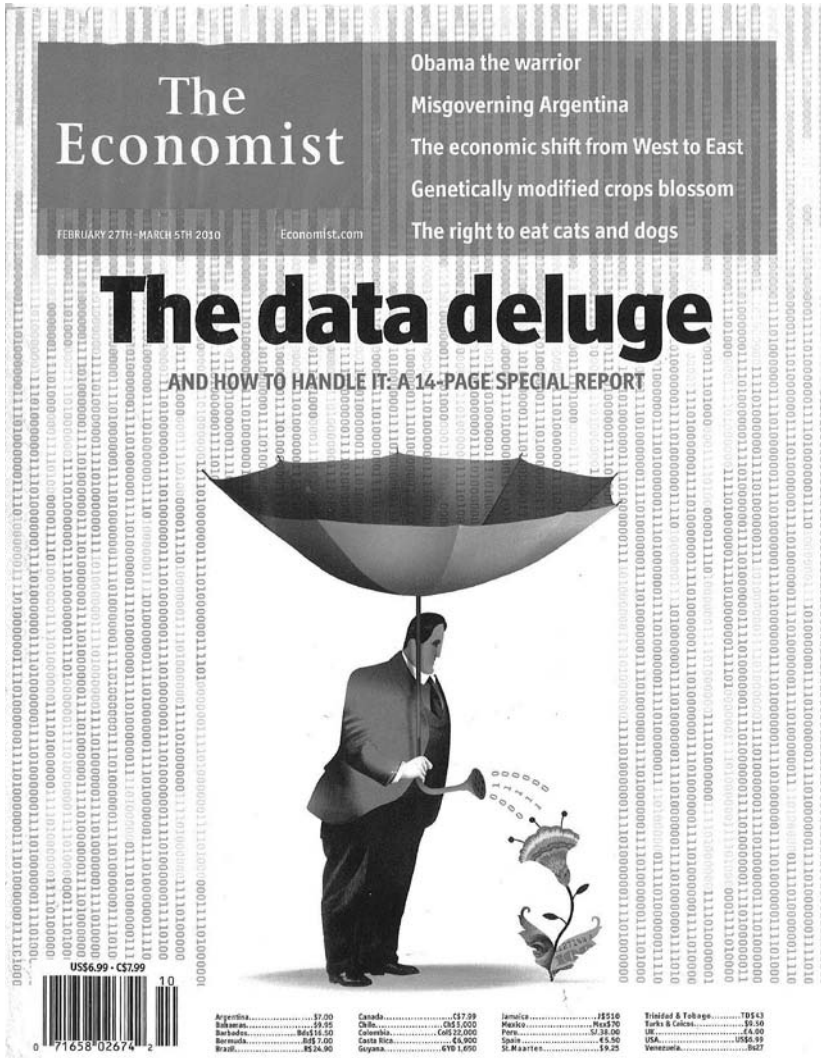

Fig. 7.1 The Economist, February 27, 2010, reported on the effects of the "data deluge" in a variety of fields, from the stock market to national security.

The literature about "big data" is growing almost as fast as "big data" itself. The popular *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (2013), by Internet analyst Viktor Mayer-Schönberger and media commentator Kenneth Cukier, or the more scholarly *Reinventing Discovery: The New Era of Networked Science* (2012) by the physicist and writer Michael Nielson and *Too Big to Know* (2012) by the Harvard Internet scholar

David Weinberger contain many insights into how the availability of big data can change how knowledge is produced.[2] But all of these authors take for granted that the amount of data available publicly is solely determined by the rate of data production. Nielson, for example, summarized the situation for genetic data: "Each time [researchers] obtained a new chunk of genetic data in their laboratories, they uploaded that data to a centralized online service such as GenBank."[3] Had these authors been less impressed by the growing rate in the production of data, they might have paid closer attention to how data was actually collected and made publicly available. They might also have realized that researchers did not simply "upload that data to a centralized online services" once they had obtained it. Indeed, the single most important concern among all those who have developed the databases and other infrastructures of big data science was how to compel researchers to share their data.[4] In 1976, for example, the managers of the first international database in protein crystallography (the Protein Data Bank) admitted to the research community that "in spite of our recent rapid rate of growth . . . we are aware that the Bank lacks data." They went on to "urge investigators to deposit" their data as it became available.[5] Parting ways with recent concern over the overabundance of data, this chapter focuses on the *scarcity* of data, examining how the data deluge, far from being a natural phenomenon or the result of a technological revolution, was achieved only slowly and in the teeth of resistance by those who envisioned how publicly available big data could transform the production of scientific knowledge.

Under specific historical circumstances, various collections of scientific things and "data," to use today's term, have been turned into "archives," making possible the development of "sciences of the archives." Since the Middle Ages, the word "archives" has designated both a place and what it contains, namely documents about the history of a people or institution.[6] Archives were established as a link between the past and the future with imagined uses and users (see Lorraine Daston's introduction to this volume). State genealogical archives, for example, have permitted the authentication of family relations and patrimonial inheritance. In the sciences, when things and "data" began to be collected, not only as prized possessions for the present, but as a resource for the future, they became scientific "archives." In natural history, the term "archive" was commonly used to designate a collection of specimens (see David Sepkoski, chapter 2 in this volume) and these "scientific archives," such as herbaria and zoological collections, have played an essential rôle in the production of knowledge. Collecting, comparing, and classifying have been key epistemic practices applied to the archives of natural history. By contrast, in the experimental

life sciences, scientific archives have been marginal, and the term "archive" rarely used, with the notable exception of experimental medicine, in which a number of journals were titled "Archiv," constituting a public collection of results intended for future use (see J. Andrew Mendelsohn, chapter 3 in this volume). However, there seems to have been no strict equivalent in the rest of the experimental life sciences, in which knowledge was produced through new experiments, not the systematic comparison of results of previous ones. Only in the late twentieth century did such comparisons of experimental results become a common practice in the experimental life sciences, transforming them into "sciences of the archives," as one might call them in view of how they function epistemically. Given that the experimental life sciences had rested on a very different epistemic tradition for most of their history, it is unsurprising that this recent transformation, culminating in the current data deluge, was far more laborious than current commentators have imagined.

COLLECTIVE COLLECTIONS

To understand the specific historical changes in data collection at the end of the twentieth century, it is useful to take a broader view and examine how things and data were collected in previous centuries and stored in scientific collections and archives—almost always a collective enterprise.[7] The sciences that relied on collections, from astronomy to zoology, have often involved a very wide range of participants, most of whom could be, since the nineteenth century, labeled as "amateurs," in that they did not make a living through their collecting practices. All of the great natural history collections of rocks, plants, and animals were gathered with the help of countless amateurs, often with exceptional levels of "lay expertise" in their field of specialty.[8] The same holds true for astronomy, in which amateurs were crucial, for example, to the collection of comet observations, or for meteorology, in which they played an essential role in the systematic recoding of rare and common phenomena across large spaces.[9]

The extensive involvement of amateurs in the sciences of the archives was no historical accident. Because these sciences required the collection of observations and things across vast geographic expanses or even the entire world, they needed observers who were physically present in diverse locales. The great scientific expeditions could collect large amounts of observations and things, but only at great expense and for a limited time period. They could hardly compete with the long-term presence of observers around the globe, provided these could be trained to supply standardized observa-

tions.[10] Thanks to intimate knowledge of their immediate surroundings, these "resident observers" could gather observations that the "traveling observer" would often overlook.[11]

The enrollment of large numbers of amateurs in a collective research project depended on the possibility of providing some kind of financial, symbolic, or personal reward. The commodification of natural objects stimulated a growing market where researchers could simply buy specimens from plant and animal dealers, recreational hunters, and private collectors.[12] More importantly, these sciences produced knowledge based on objects—ferns, crystals, clouds—that were visible to the (trained) human eye and had long been part of a vernacular culture. In nineteenth-century England, nonscientists could be passionate about ferns, discuss them in pubs, and hold them in their homes as prized cultural items, whereas for scientists these ferns were specimens to be named, classified, and theorized about.[13] Corals in eighteenth-century France tell a similar story: simultaneously beautiful ornaments of aristocratic salons and scientific objects for naturalists.[14]

With the rise of the experimental sciences, especially since the late nineteenth century, the relationship between professional and amateurs changed drastically. From its origins, the laboratory was a private space, located in the home of an experimentalist.[15] Strict control over who could access the laboratory was key to its epistemic function. The laboratory was accessible only to a select few gentlemen, not the broader public.[16] As the laboratory came to hold increasingly complex, expensive (and sometimes dangerous) equipment, it became almost exclusively located in research institutions, thus deepening the divide between professional scientists and the public. In the twentieth century, a few sciences, mainly those in the natural history tradition, still relied extensively on amateurs, but these sciences were becoming increasingly marginalized, in terms of both budgets and prestige, by the experimental sciences. The experimental sciences, on the other hand, became fully professionalized: there was no community of amateurs to rely on for the collection of experimental data. The production of experimental data was a matter for professionals. This fundamental difference between the experimental and the naturalist sciences had deep consequences for data collections.

WITHHOLDING DATA

In the molecular life sciences in the 1970s and 1980s, a number of new technologies resulted in improved methods for determining the structure of macromolecules. In 1977, for example, two new methods permitted de-

termination of DNA sequences. Combined with the wide interest in the biological meaning of DNA sequences, these new methods resulted in an exponential growth in the production of sequence data.[17] In crystallography, more powerful methods, relying on digital computers, were also developed in the 1970s, speeding up the production of crystallographic data and the number of solved molecular structures.[18] This burgeoning store of data was often shared informally among researchers. It made possible a deep epistemic transformation in how data was used. As I have argued elsewhere, one of the most significant changes in the experimental sciences during the twentieth century was the increasing reliance on practices of producing knowledge based on collecting, computing, comparing, classifying, and curating large and diverse amounts of data.[19] These practices, so common in natural history and other observational sciences, became key to the experimental sciences as well. By adopting a different set of epistemic practices, whose potential became ever clearer starting in the 1970s, the experimental sciences were also confronted with new material and social challenges. But unlike the practitioners of natural history and other collecting sciences, who had long experience in resolving such problems, the experimentalists were at a loss to find solutions within the specific moral economies of their communities.[20] The single most important challenge was how to collect in a single place and make public the massive amount of data required to feed these epistemic practices.

Whereas most sciences of the archives had relied on and cultivated large networks of professional and amateurs, the experimental sciences had severed their ties with amateurs more than a century ago and consigned them to the role of distant spectators.[21] Thanks to the growth of the scientific workforce in the second half of the twentieth century, there was a much larger community of professionals able to contribute to a large collecting effort—on the condition that they could be persuaded to participate in this collective effort. Given a professional identity that valued individual achievement over collective participation, this was no simple task, especially if it implied sharing data difficult to produce and potentially rich in the epistemic rewards of new publications.

Since the Scientific Revolution of the sixteenth and seventeenth centuries, experimental results, "data" one could say, had been treated as the private property of the investigator who produced and carefully guarded it in laboratory notebooks. Data was disclosed publicly in exchange for scientific credit through oral communication in academies or publications in printed journals. This moral economy was still very much at work in the twentieth century. In 1968, American biologist James Watson revealed in his tell-all autobiography, *The Double Helix*, that he "was more aware of

[British crystallographer Rosalind Franklin's] data than she realized" and that "Rosy, of course, did not directly give us her data."[22] The data in question, communicated to Watson and his British collaborator Francis Crick though a confidential activity report, proved essential for the determination of the double helix structure of DNA. Scientists who reviewed Watson's book were almost unanimous in condemning his behavior and his shameless bragging about it.[23] In their view, Watson robbed Franklin of her data and thereby of her due credit. Data belonged to individuals, not to the scientific community as a whole.

COLLECTING DATA

Those who attempted to set up large collections of molecular data in the second half of the twentieth century experienced this problem firsthand, as the history of the Protein Data Bank and of GenBank, two of the major databases in the life sciences, make abundantly clear. The Protein Data Bank was set up in Brookhaven National Laboratory in 1973 to store all the existing data about the three-dimensional structure of proteins as determined by crystallographic methods. At the time of its creation, only a dozen protein structures had been determined by a small community of protein crystallographers who gathered at Cold Spring Harbor in 1971 for a symposium on the topic. Most of the researchers who had pioneered the determination of protein structures were present, including Max Perutz (hemoglobin), David Phillips (lysozyme), Frederic Richards (ribonuclease A), and William Lipscomb (carboxypeptidase A). Walter Hamilton, the young president of the American Crystallographic Association, aired the idea of a data bank for protein structures, a proposal made by two even younger colleagues, Helen Berman and Edgar Meyers. His colleagues responded very favorably to the idea, and the initial set of data was collected on the strength of friendships among Hamilton and some of the crystallographers present at the meeting.[24]

By May 1973, the data bank was "about ready to begin distribution" and a formal announcement was published in *Acta Crystallographica* and in the *Journal of Molecular Biology*.[25] The Protein Data Bank contained the coordinates of just nine proteins and anyone could obtain the entire data bank on a magnetic tape for a modest sum, covering shipping and the cost of a blank tape. The announcement repeated the call made two years earlier: the "usefulness of the system" would depend on "the response of the protein crystallographers supplying the data."[26] In the following months, researchers who determined structures began to acknowledge in their pub-

lications that they had deposited the coordinates in the Protein Data Bank. The number of available structures grew, but slowly. In January 1974, there were just twelve structures available, a year later fifteen, and the following year twenty-three.[27]

Most authors, who had been friends of Hamilton, agreed to release the data. However, not all of them must have been entirely comfortable with their decision. One crystallographer, for example, authorized the release of his data, but asked that he be informed about who would access it.[28] Others, such as Max F. Perutz, wished to hold back the data until it was further refined.[29] The Protein Data Bank managers tried to persuade him to release the data because "coordinates at any stage of refinement will be extremely interesting and very useful to many people,"[30] and Perutz eventually agreed, but only after a year (and after having submitted another paper based on the further refinement of the same data).[31]

In order to encourage researchers to deposit their data, while recognizing their interest in keeping them private so they could exploit them further, the managers of the Protein Data Bank, together with some journal editors, devised an original system in the early 1980s. Journal editors asked prospective authors that the data supporting the conclusions of their scientific paper be submitted to the Protein Data Bank. To overcome proprietary resistance, the data bank managers offered researchers the option to deposit their data but to restrict its access to the public for up to four years after the publication of a paper based on the data. In 1989, more than 75 percent of those depositing data chose to keep it private for the maximum period of four years.[32] Clearly, most of the community of crystallographers was not ready to make data communal property at the time of publication unless forced to do so.

To make matters worse, the very definition of what counted as data, especially "raw data," was not universally agreed upon.[33] Should only the atomic coordinates of the protein model be considered "raw data"? Or the "structure factors," a set of calculated data from measured diffraction intensities, from which the atomic coordinates were derived? Or the intensities of the diffraction spots measured on the diffraction images used to calculate the structure factors? Or the diffraction images themselves, produced directly by the x-rays going through a protein crystal? At first, the Protein Data Bank focused on the atomic coordinates that described with precision the position of each atom in a protein model. But these data were far from being "raw." They were the result of many steps of measurements, calculations, and interpretations. Without the structure factors, the proposed structure could not be challenged. Thus some crystallographers argued that coordinates were not data at all, but results derived from data.

As the American crystallographer Richard E. Dickerson put it in a letter to the president of the American Crystallographic Association, "Results without data are unproven, and interpretations without results are hearsay."[34] The Protein Data Bank managers therefore began to ask crystallographers to include structure factors along with atomic coordinates. But many researchers resisted, arguing that the structure factors should be considered research notes, not data, and thus felt no obligation to share them.

The situation was very similar among molecular geneticists who were facing their own data deluge in the same period. In 1983, the NIH funded the creation of a central database named GenBank for all DNA sequences.[35] Located at Los Alamos, its main architect, the physicist Walter Goad, had promised funding agencies that all the data published in the scientific literature would be included in GenBank "within a year" and all new data would be integrated as soon as they were published in journals.[36] But three years later, only 19 percent of the sequences published the previous year were publicly available in GenBank.[37] The gap between the data available in printed journals and in electronic databases was constantly growing. And there was an unknown amount of data that was neither published nor deposited in databases. Unlike the field of crystallography, much of the data associated with a published article was included in the scientific journals and thus publicly available. But printed DNA sequences (long strings of As, Ts, Gs, and Cs) were of little use for whoever wanted to analyze them with a computer. Worse, almost all the original published sequences were inaccurate (typographic errors were impossible to spot by a copy editor). Since researchers were eager to have correct sequences in electronic format available from a database like GenBank, managers of GenBank were faced with the daunting task of typing sequences manually from a journal into a computer, which was not only time consuming but also added its own set of errors.

As they were increasingly falling behind the growing amount of data available in the scientific literature, the managers of GenBank expected to enroll the community in the collecting effort. Like the managers of the Protein Data Bank, those of GenBank hoped that researchers would directly send their data in an electronic format to the database at the same time as they submitted a manuscript for publication in a journal—or at least that they would use the papers forms, sent out by journal editors, on which authors could carefully write down their DNA sequences and return it to GenBank. In numerous calls published in scientific journals, GenBank managers cajoled researchers to comply. They encouraged, threatened, persuaded, cheered, and appealed to moral obligations in order to change the behavior of experimentalists. Journal editors, who were swamped by

sequence data and preferred to have it deposited in GenBank rather than printed in their pages, echoed these calls. In *Proceedings of the National Academy of Sciences*, for example, an editorial reminded the readers that "scientists who generate sequences . . . are also the users of sequences . . . self-interest should . . . dictate compliance."[38] All to little avail. As the editor of *Nucleic Acids Research* remarked pointedly, "Scientists would like access to everyone else's data through they do not necessarily wish to reciprocate."[39]

There were many reasons for the researchers' resistance to sharing data. Some simply felt that it was not worth the time needed to format and deposit the data in GenBank. Others were attempting to protect their data from potential competitors. Still others were concerned that it contained errors that could be spotted by others and tarnish their reputation. In the case of crystallographic data, researchers often wanted to "refine" the data further (a mathematical procedure), making it more precise, before they released it to the public. In practice, the basic principle that a publication requires all the data upon which the conclusions rest be made public (in print or upon request) was honored only in the breach. As Dickerson put it in no uncertain terms, "By the standards normally applied in other branches of science [the structures published without available data] are not really published at all, in the literal sense of making the information public."[40]

OPEN SCIENCE

The rise of open science in the last decade of the twentieth century was not the result of a spontaneous surge of altruism among researchers or a transformation in the moral economy of experimental science. Data collectors, such as those working for GenBank and the Protein Data Bank, began to pursue a new strategy more in line with the existing moral economy of the experimental life sciences. For many years, they had hoped that the communal ethos in science would allow them to collect data much as earlier natural history collectors had done. There, amateurs openly shared their "data," in the form of specimens or observations, with, for example, collectors affiliated with natural history museums or local naturalist societies. These amateurs received little, if any, credit for their contribution: at best, an acknowledgment in print or, exceptionally, a new species named after them. But in the experimental sciences, where professionals made a career (and a living) by turning data into credit, the naturalist system of data collection was bound to fail.

In this respect, the end of the 1980s marked a turning point. The National Institutes of Health (NIH) were particularly sensitive to the availa-

bility of data concerning genes or proteins related to diseases of great public concern, such as cancer or AIDS. In 1988, a group of researchers published a paper describing the structure of a protein from a cancer-causing gene, Ras. Two years later, the atomic coordinates describing the structure were still unavailable in the Protein Data Bank. When questioned about this omission, the article's lead author argued that he still needed to resolve some problems with the structure before depositing the data. An NIH official was annoyed by this prevailing attitude: the "data are good enough so that conclusions that are drawn from them can be published but not good enough to see the light of the day."[41]

To combat this problem, an NIH agency passed a resolution recommending that all grantees make their crystallographic data available within one year of publication, and that funding be withheld from those who did not comply.[42] Although this sanction sounded severe, it did not include any systematic enforcement measures. But it contributed, together with the growing pressure from professional societies, to the new system of data collection dependent on journal editors. By 1990, a number of them began to adopt policies mandating the sharing of crystallographic data with the Protein Data Bank. This proved far more effective.

This change in policy among journal editors resulted from the desperate efforts of database managers to solve their data collection problem and those of journals to preserve their epistemic authority while avoiding the costs of publishing growing amounts of data. Journal editors not only adopted but enforced mandatory submission policies, simply by deciding to publish only papers by authors who complied. These policy changes were almost always initiated by editorial board members with close ties to the data banks and who understood that data sharing was in the best interest of the research community. The molecular biologist Richard J. Roberts, for example, on the advisory board of GenBank and executive editor of *Nucleic Acids Research*, introduced the policy for that journal in 1988.[43] The principle was very simple. Journal editors would publish a paper only if the authors could provide an "accession number" demonstrating that the supporting data had been submitted to a public database, like GenBank or the EMBL data library, its European equivalent. Some journals, such as *Nature*, persisted in opposing any mandatory submission policy, and its editor-in-chief, John Maddox, encouraged other journals to resist "being turned into instruments of law-enforcement."[44] But *Nature* was becoming increasingly isolated. Authors also seemed to have some misgivings about having their data released too quickly, as almost 50 percent of those who submitted data to GenBank asked for confidentiality until their papers appeared in print.[45] Overall, the efforts of the EMBL and GenBank persuaded enough journals

to adopt submission policies, essentially solving the problem of data collection for sequence databases. These policies had an immediate and dramatic effect: in 1990, 75 percent of all data submitted to GenBank came directly from authors. By attaching the rewards (priority, credit, and authorship) that go with publishing in a journal to data deposition, the experimental sciences solved in their own way an old challenge of the sciences of the archives.

DATA PUBLICATION

The open science revolution was thus no revolution at all, in the sense of a profound transformation in the political and cultural values governing individual and collective behavior. The rise of open science resulted from a new alignment between individual and collective interests within the existing moral economy of science. Researchers began to share data because it became a requirement in order to publish papers and thus reap the associated credit. More recent attempts to encourage data sharing illustrate further the conservative nature of the open access transformation. The mandatory data sharing enforced by journal editors only concerned data that was used as evidence for claims made in a scientific paper. The vast majority of data produced was unaffected by this policy and remained in private laboratory notebooks and computers. Thus scientists and science administrators imagined two different models to tie data sharing to the existing reward system in science: data authorship and data citation.

In the last decades of the twentieth century, as the amount of data produced by scientists increased dramatically, printed journals began to exclude the possibility of publishing data alone. Only when data was used as supporting evidence for a broader claim could it be included in a scientific paper or deposited in a public database. The production of data alone was no longer considered an intellectual achievement that could be rewarded by granting scientific authorship. A few decades earlier, the situation had been very different. In the 1950s, the biochemist Frederick Sanger published a series of papers describing, for the first time, the sequence of a protein, insulin—a publication rewarded by the Nobel Prize in Physiology or Medicine in 1958. But by the end of the century, protein sequences, and even more so DNA sequences, were determined in numerous laboratories, often through automated methods, without necessarily being published. In order to encourage researchers to submit these data and make them public, database managers attempted to rely on the same incentive as journal editors: the granting of authorship. The Protein Data Bank, for example,

made it possible to cite "an entry without a published reference," including the name(s) of the author(s), a descriptive title, and a Protein Data Bank unique identifier (or a Digital Object Identifier, DOI).[46] A data entry could then be listed in an author's publication list, along with articles published in scientific journals. Journal editors have responded to databases' challenges to their exclusive rights to grant data authorship by launching new journals solely for the publication of data. Nature Publishing Group, for example, started *Scientific Data*, in 2014, for that purpose.[47] Databases and data journals both grant authorship, thus allowing researchers to claim the professional rewards attached to publications. Although this model for data sharing rests on the traditional reward system in the sciences based on publication records, its impact is limited by the fact that the value attributed to a publication depends on the reputation of the journal where it is published, a reputation that reflects how selective the peer-review process of the journal is perceived to be and various metrics of the journal's influence.[48] Although data deposition in a database or journal could be counted as a publication, its value in the scientific reward system thus remains low. So does the incentive to deposit data, limiting the impact of data authorship on data sharing.

An alternative model, based on data citations, was developed to overcome the limitations of data authorship. Along with the publication record, the scientific reward system has been based on citation records. Since 1964, the Science Citation Index, created by the American linguist Eugene Garfield, has tracked the number of times a given article is cited in the scientific literature.[49] This number is being used, in various combinations, as a way to measure quantitatively a scientist's impact on a scientific field. Since 2005, the number of citations is used to calculate the *h*-index, a measurement of a scientist's productivity and impact that has become a standard part of a scientist's resumé and that is often required by science funding agencies and academic search committees. As quantitative measurements of citations became increasingly influential in shaping scientific careers, proponents of data sharing—including database managers, journal editors, and funding agencies—encouraged authors to cite individual data entries, including the name of the researchers who had deposited the data, as they would for published papers. In the 1960s, when the first databases in the life sciences were established, researchers resented the fact that the database as a whole was cited, instead of the paper where they had first published the data, thus depriving them of the credit associated with a scientific citation. This citation practice also discouraged data sharing with a public database since it "anonymized" its origins.[50] Efforts by the Committee on Data for Science and Technology (CODATA), the US National Academy of Sciences,

and various other groups of scientists led to the publication of a Joint Declaration of Data Citation Principles in 2013.[51] It emphasized the importance of minimal standards for data citations. Just a few months earlier, the media multinational Thompson Reuters, which maintains the bibliographic record and citation index Web of Science, launched its Data Citation Index.[52] This new tool made it possible to measure how often a particular data set is cited in the scientific literature, thus encouraging data sharing, even when data is not associated with the publication of research findings.

CONSERVATIVE REVOLUTION

It is too early to say whether these initiatives will have a significant impact on data sharing practices. But what seems historically significant is the degree to which they have retreated from the idealistic attempts of the 1960s and 1970s to transform the moral economy of experimental science. Instead, they all rely on the existing reward system based on the granting of authorship by community-based journals (or databases) and the citations of published work by members of the scientific community. Thus scientific journals, through their almost exclusive power to grant authorship, still hold the key to this reward system. Databases managers, after relying on journals to enforce mandatory data submission policies, began to challenge the exclusive rights of journals by arrogating to themselves the power to grant a form of authorship for data. The scholarly literature about the rise of open science has focused on the policies elaborated by governments and science funding agencies, overlooking the role of journal editors and database managers. In *Reinventing Discovery*, for example, Nielson claimed that "the granting agencies are the de facto governance mechanism in the republic of science, and have great power to compel change, more power even than superstar scientists such as Nobel prizewinners."[53] Others have described the "open science revolution" as essentially spontaneous, a revolution "from below," where individual researchers became committed to open science and shared data voluntarily in the best interest of the scientific community. This chapter argues that there was no revolution at all, or only a conservative one. The open science revolution might have changed how much data was made available publicly, a great collective benefit, but not the reason individual researchers shared data. By and large, researchers have shared data because it became in their own interest to do so, as defined by the existing reward system in the experimental sciences. Far from upsetting the current moral economy of science, the rise of open science illustrates how much it remained entrenched in current scientific practice. For this

reason, researchers have even suggested that the term "data sharing," with its communitarian overtones, be abandoned and replaced by "data publication," a term perfectly in line with the individualistic ethos prevalent in the experimental sciences.[54]

In short, the data deluge was a product of two different transformations. First, it represents an expansion of what falls under the category of "data." Many research notes, preliminary measurements, and private observations did not count as data until the end of the twentieth century. The current data deluge is not simply the product of the increased amount of data being produced. It is also the result of the enlargement of what counts as data and thus merits preservation. The exact definition of "data," however, remains a moving target for all those involved in data policies, such as the National Science Foundation. Although since 2011 the NSF has required a "data management plan" (DMP) from its grantees, nowhere did it provide a definition of what counts as data, expecting that norms would be developed by the different research communities.[55] The NSF was treading lightly because labeling something as "data" created obligations to preserve it and make it publicly accessible. Second, the data deluge depended on the coupling of these obligations to the existing moral economy of the experimental life sciences. Data sharing became an obligation tied to gaining authorship and citations, the key components of the experimental sciences' individualistic reward system.[56] By making individual and collective interests coincide, the proponents of "open science" engineered a "data deluge," allowing the experimental science of the archives to flourish at the beginning of the twenty-first century.

NOTES

1. On the sciences of the archive, see Lorraine Daston, "The Sciences of the Archive," *Osiris* 27, no. 1 (January 1, 2012): 156–87; on collecting sciences, Bruno J. Strasser, "Collecting Nature: Practices, Styles, and Narratives," *Osiris* 27, no. 1 (January 2012): 303–40; on data flows, Stephen Hilgartner and Sherry I. Brand-Rauf, "Data Access, Ownership, and Control: Toward Empirical Studies of Access Practices," *Knowledge: Creation, Diffusion, Utilization* 15, no. 4 (1994): 355–72.

2. Michael A. Nielsen, *Reinventing Discovery: The New Era of Networked Science* (Princeton: Princeton University Press, 2012); Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston: Houghton Mifflin Harcourt, 2013); David Weinberger, *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room* (New York: Basic Books, 2011).

3. Nielsen, *Reinventing Discovery*, 7.

4. See for example Rita R. Colwell, David G. Swartz, and Michael Terrell MacDonell, *Biomolecular Data: A Resource in Transition* (Oxford: Oxford University Press, 1989).

5. Protein Data Bank, *Newsletter* 3 (1976): 2.

6. Frédéric Eugène Godefroy, *Dictionnaire de l'ancienne langue française et de tous ses dialectes du 9e au 15e siècle* (Paris: F. Vieweg, 1881).

7. Nicholas Jardine, James A. Secord, and Emma C. Spary, eds., *Cultures of Natural History* (Cambridge: Cambridge University Press, 1996).

8. On amateurs in natural history, see David Elliston Allen, "Amateurs and Professionals," in *The Cambridge History of Science: The Modern Biological and Earth Sciences*, ed. Peter J. Bowler and John Pickstone (Cambridge: Cambridge University Press, 2009), 15–33; David Elliston Allen, *The Naturalist in Britain: A Social History* (London: A. Lane, 1976); Jim Endersby, *Imperial Nature: Joseph Hooker and the Practices of Victorian Science* (Chicago: University of Chicago Press, 2008); Kristin Johnson, *Ordering Life: Karl Jordan and the Naturalist Tradition* (Baltimore: Johns Hopkins University Press, 2012). On lay expertise, Steven Epstein, *Impure Science: AIDS, Activism, and the Politics of Knowledge* (Berkeley: University of California Press, 1966).

9. W. Patrick McCray, "Amateur Scientists, the International Geophysical Year, and the Ambitions of Fred Whipple," *Isis* 97, no. 4 (December 2006): 634–58; Dunlop Storm and M. Michèle Gerbaldi, eds., *Stargazers: The Contribution of Amateurs to Astronomy* (Berlin: Springer-Verlag, 1988); Jan Golinski, *British Weather and the Climate of Enlightenment* (Chicago: University of Chicago Press, 2007).

10. Peter Galison and Lorraine Daston, "Scientific Coordination as Ethos and Epistemology," in *Instruments in Art and Science: On the Architectonics of Cultural Boundaries in the 17th Century*, ed. Helmar Schramm, Ludger Schwarte, and Jan Lazardzig (Berlin: Walter de Gruyter, 2008), 296–333.

11. On residential science, see Robert E. Kohler, "Paul Errington, Aldo Leopold, and Wildlife Ecology: Residential Science," *Historical Studies in the Natural Sciences* 41, no. 2 (May 2011): 216–54.

12. Fa-ti Fan, *British Naturalists in Qing China: Science, Empire, and Cultural Encounter* (Cambridge, MA: Harvard University Press, 2003); Mark Barrow, "The Specimen Dealer: Entrepreneurial Natural History in America's Gilded Age," *Journal of the History of Biology* 33 (2000): 493–534.

13. Allen, *The Naturalist in Britain*.

14. Krzysztof Pomian, *Collectors and Curiosities: Paris and Venice, 1500–1800* (Cambridge, England: Polity Press, 1990).

15. Steven Shapin, "The House of Experiment in 17th-Century England," *Isis* 79, no. 298 (September 1988): 373–404.

16. Steven Shapin, *A Social History of Truth: Civility and Science in Seventeenth-Century England* (Chicago: University of Chicago Press, 1995).

17. On sequencing methods, see Miguel Garcia-Sancho, *Computing, and the History of Molecular Sequencing* (New York: Palgrave Macmillan, 2012).

18. Walter C. Hamilton, "The Revolution in Crystallography," *Science* 169, no. 941 (July 10, 1970): 133–41.

19. Strasser, "Collecting Nature"; Bruno J. Strasser, "The Experimenter's Museum:

GenBank, Natural History, and the Moral Economies of Biomedicine," *Isis* 102, no. 1 (March 2011): 60–96.

20. For a review of the notion of moral economy, see Didier Fassin, "Les économies morales revisitées," *Annales: Histoire, Sciences Sociales* 64, no. 6 (2009): 1237–66.

21. On the growing divide between the science and their publics around 1900, Bernadette Bensaude-Vincent, *L'Opinion publique et la science* (Paris: La Découverte, 2013).

22. James D. Watson, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA: Text, Commentary, Reviews, Original Papers* (New York: Touchstone, [1968] 2001), 105

23. Watson, *The Double Helix*; Soraya de Chadarevian, *Designs for Life: Molecular Biology after World War II* (Cambridge: Cambridge University Press, 2002).

24. Interview with Helen Berman, July 17, 2009, New Brunswick, NJ.

25. Thomas Koetzle to Wayne Hendrickson, May 4, 1973, PDB Archives, New Brunswick, NJ (PDB Archives hereafter); anonymous, "Crystallography Protein Data Bank," *Journal of Molecular Biology* 78 (1971): 587.

26. Anonymous, "Crystallography Protein Data Bank," 587.

27. Protein Data Bank Annual Report to ACA, 1974, 1975, 1976, PDB Archives.

28. Wayne Hendrickson to Thomas Koetzle, April 6, 1973, PDB Archives.

29. Thomas Koetzle to Max F. Perutz, May 22, 1973, PDB Archives.

30. Thomas Koetzle to Max F. Perutz, May 22, 1973, PDB Archives.

31. Max F. Perutz, "Refinement of Hemoglobin and Myoglobin," *Acta Crystallographica Section A* 31 Supplement S (1975): 31.

32. Joel L. Sussman, "Protein Data Bank Deposits," *Science News Letter* 282, no. 5396 (December 11, 1998): 1993; Alexander Wlodawer, "Deposition of Macromolecular Coordinates Resulting from Crystallographic and NMR Studies," *Nature Structural Biology* 4, no. 3 (March 1997): 173–74.

33. Lisa Gitelman, ed., *Raw Data Is an Oxymoron* (Cambridge: MIT Press, 2013).

34. Richard E. Dickerson to Charles E. Bugg, July 27, 1987, PDB Archives.

35. Strasser, "The Experimenter's Museum."

36. GenBank Advisors Meeting, Minutes, November 6, 1987, EBI Archives.

37. Richard Lewin, "Proposal to Sequence the Human Genome Stirs Debate," *Science* 232, no. 4758 (June 27, 1986): 1598–1600.

38. Igor B. Dawid, "Editorial Submission of Sequences," *PNAS* 86 (1989): 407.

39. Richard T. Walker, "A Method for the Rapid and Accurate Deposition of Nucleic Acid Sequence Data in an Acceptably-Annotated Form," in *Biomolecular Data: A Resource in Transition*, ed. Rita Colwell (Oxford: Oxford University Press, 1989), 45–51.

40. Richard E. Dickerson to Charles E. Bugg, July 27, 1987, PDB Archives.

41. Abraham M. De Vos et al. "Three-dimensional Structure of an Oncogene Protein: Catalytic Domain of Human c-H-ras p21," *Science* 239 (1988): 888–93; Jim Cassatt to Helen M. Berman, January 4, 1990, PDB Archives.

42. John C. Norvell to principal investigators in NIGMS, April 23, 1990, Protein Data Bank Archive.

43. Patricia Kahn and David Hazledine, "NAR's New Requirement for Data Submission to the EMBL Data Library: Information for Authors," *Nucleic Acids Research* 16, no. 10 (May 25, 1988): I–IV.

44. John Maddox, "Making Authors Toe the Line," *Nature* 342 (1989): 855.

45. GenBank Advisors Meeting, Minutes, November 15–16, 1988, NCBI Archives.

46. Protein Data Bank, Policies & References, available at http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/policies_references.html, accessed July 10, 2014.

47. Anonymous, "More Bang for Your Byte," *Scientific Data* 1 (2014), accessed July 10, 2014, doi:10.1038/sdata.2014.10.

48. Björn Brembs, Katherine Button, and Marcus Munafò, "Deep Impact: Unintended Consequences of Journal Rank," *Frontiers in Human Neuroscience* 7 (2013). doi:10.3389/fnhum.2013.00291.

49. Eugene Garfield, "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas," *Science* 122, no. 3159 (July 15, 1955): 108–11.

50. Bruno J. Strasser, "Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965," *Journal of the History of Biology* 43, no. 4 (2010): 623–60.

51. CODATA-ICSTI Task Group on Data Citation Standards and Practices, "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data," *Data Science Journal* 12, no. 0 (2013), CIDCR1–CIDCR75. See also Paul F. Uhlir, *Board on Research Data and Information, Policy and Global Affairs, National Research Council. For Attribution—Developing Data Attribution and Citation Practices and Standards* (Washington, DC: National Academy Press, 2012).

52. Thomson Reuters, "Thomson Reuters Launches Data Citation Index for Discovering Global Data Sets" (April 2, 2013) accessed November 17, 2014, http://thomsonreuters.com/content/press_room/science/730914

53. Nielsen, *Reinventing Discovery*, 191

54. Mark J. Costello, "Motivating Online Publication of Data," *Bioscience* 59, no. 5 (2009): 418–27.

55. Jacob Glenn, "NSF Data Management Plan" (2013), accessed July 11, 2014, http://www.lib.umich.edu/research-data-services/nsf-data-management-plans

56. The sharing imperative became embedded within the system, as Chris Kelty has described for the creative commons license. See Christopher M. Kelty, *Two Bits* (Durham: Duke University Press, 2008).