JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

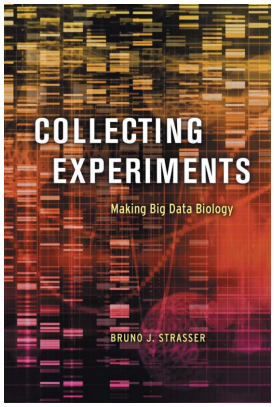
Collecting Experiments. Making Big Data Biology. By Bruno J. Strasser. Chicago University Press, 2019. Pp. 392. Price USD 45.00. ISBN 9780226635040.

John R. Helliwell*

Department of Chemistry, University of Manchester, Manchester, Manchester M13 9PL, United Kingdom.
*Correspondence e-mail: john.helliwell@manchester.ac.uk

I dedicate this book review to John Westbrook (1957–2021), who contributed a great deal to the PDB over many years (Zardecki & Burley, 2021) and to the IUCr's comCIFS, Diffraction Data Deposition Working Group and Committee on Data.

Keywords: book reviews; biology; big data.



I first noticed this book on Twitter, where it was recently highlighted by the Protein Data Bank (PDB). The author, Bruno Strasser, is a 'historian of science, science studies and science education' based at the University of Geneva, who also holds an adjunct professor position at Yale University (https://biologie.unige.ch/fr/la-section/reseaux/didactique-des-sciences/). The book comprises 404 pages in the PocketBook Reader edition. The book spans a very wide range of themes: live organism collections and natural history museums, as well as digital collections of protein and gene sequences and protein crystal structures (the PDB).

Several times in the opening chapter (zero) entitled *Introduction*, the author declares what this book is about, helpfully from different angles. At the end of the chapter, and at its most simple, 'this book is about the hybrid culture of experimenting and of collecting'. In more detail 'This book is about the development and use of data collections in the experimental life sciences from the early twentieth century to the present: their emergence, their development, their meaning and their effects on the production of knowledge and on scientific life.' Thirdly, 'This book builds on the opposition between two ways of knowing: the comparative (via collections) and the experimental (on example cases).' Although focused on the past century, the book delves into the practices of earlier centuries, and compares modern databases with the pre-digital collections in natural history museums or botanical gardens, which continue to this day of course. Several times he thanks the late Professor John Pickstone of the University of Manchester and acknowledges his 2000 book on 'ways of knowing'.

Chapter 1 is entitled *Live Museums* and charts the start of such collections in the USA and in Europe. The chapter commences with a focus on collections of live bacteria, with some mention of botanical gardens and marine stations. It then describes the start and growth of collections of mice of a million or more, which were distributed from one centre alone initially as a for-free science service and then offered for sale as a commodity. Then this was extended to rats and guinea pigs with, for example, the genetically highly homogeneous Wistar rat being trademarked in 1942 in the USA. The chapter moves on to a description of collections of corn and maize and their data, including genetic variations, and the sharing of results by newsletter and then publication. There follow sections on drosophila collections and on viruses and bacteria. At page 61 is an important mission statement for the various collection types, museums of exhibits and live museums: 'The museum is needed to supplement and give substance to the library.' As crystallographers we know of the vital importance of our articles being connected to our digital data, which underpins our studies. As a summary of this chapter, the author states that 'In 2016 there were at least 726 culture collections of microbes alone in 75 countries, and stock collections for all model plant and animal organisms used in research.'

Chapter 2 is entitled *Blood Banks*. 'This chapter focuses specifically on how the classification of species came to be studied in laboratories at the biochemical level.' On page 72 we learn of the 1909 'collection of blood (haemoglobin) crystals from over 100 species' of George Nuttall of the University of Cambridge. [This story goes back even further to 1840 and earlier (Giegé, 2013).] There then follow two sections on the efforts through several decades of the mid-20th century to change the classification of blood (and more general classifications in zoology) via experimental methods of physics (*e.g.* measuring the turbidity of blood) and to achieve 'complete' collections, *i.e.* across

many species. This theme of completeness is an important one for the whole book and links to the obvious dedication of the natural history professional, completeness being as important in this case as precision and accuracy are to a physicist. The latter concerns also apply to the use of physics in studying biology (Helliwell, 2021) and in seeking to know the structural chemistry of the living organism at its temperature and pressure (Helliwell, 2020). The seeking of completeness is a nice reminder of why Max Perutz pursued the X-ray crystal structures of so many haemoglobins from different species, which gave many insights into the biological adaptations of species. The developing emphasis on collecting biochemical samples, such as blood and then egg white from different species of birds' eggs, led to the sending out of mobile laboratories rather than reliance on amateur collectors. A delightful example was the ship 'most probably named Linus Pauling's alpha helix', pictured in 1966 in Fig. 2.9 on page 95. At page 105 emerges the culture clash of the researcher in the field in a shack with a microscope and kerosene light versus a researcher familiar with computer magnetic tapes. The chapter closes with the remark that 'no collection, no data; no data, no knowledge'.

Chapter 3 is entitled *Data Atlases*. The chapter opens with a detailed discussion of the pioneering *Atlas of Protein Sequence and Structure* series of compendia brought together by Margaret Dayhoff (starting in 1965). This opened the door to comparing amino acid sequences, which formed the core of the new discipline of 'comparative biochemistry'. I find this a very interesting approach. I compare it with physics, for example, where there can be an absolute defined by a law and an equation. For a living organism there is the survival of the fittest, but if the organism's environment changes then the laws of its jungle have changed, whereas the laws of physics do not change. Amino acid sequence comparisons across many biological sources of individual proteins also provided a scale of evolution measured via the rates of change of these amino acids. On page 138 comes the question, and a first sketch, of the ownership of data entries and of all the data entries as a collection. [This is a complicated question still exercising our minds. The answer today, for crystallographers, is that ownership of data sets depends on various factors: from country to country, by institution, by funding agency, by facility (X-ray synchrotron or laser, neutron source or electron microscope or NMR installation).] The core point here is that clearly a collection enables new discoveries.

Besides the legal position of a collection of data, there is also (page 142) the question of the new role of the curator of scientific data, who was deemed neither an experimentalist nor a theoretician. The curator role required precision in the sequence data, and in what we would call the metadata associated with the sequence data, and this precision allowed predictions. But the role was often disparaged as clerical, and such instances are carefully documented by Strasser. The US funding agencies (NIH, NIGMS and NASA) were reluctant to fund the *Atlas*. There is a section in this chapter on the evidence of discrimination against Margaret Dayhoff and her all-female team. Eventually the US National Library of Medicine provided modest support. The answer to funding sustainability proved to be subscribers to the *Atlas* and its updates.

Chapter 4 is *Virtual Collections*. This chapter opens with a dramatic statement that 'in the post war decades crystallographers describe their everyday research practices as tedious'. Furthermore that 'The tediousness of the protein crystallographer's job is essential in understanding how and why they developed the Protein Data Bank in 1971'. I rather took umbrage at the author's description: however, it was not his, it was Crick & Kendrew's (1957). This article made a strong impression on me, as did that by Hamilton (1970), who described the marked improvements in crystallographic methodology that had come about since then through the 1960s. There follows a detailed description of the pioneering work with early molecular graphics computer systems, to replace the previous wire models. There is also the wondrous vision of the new discovery power of a collection of crystallographic data instigated by John Desmond Bernal and Olga Kennard. There is a delightful description of the earliest ideas for a protein data bank and the steps taken by Helen Berman and Edgar Meyer with those ideas, and their research. Walter Hamilton as a convinced senior crystallographer decided to join the initiative of Edgar Meyer and Helen Berman at the 1971 Cold Spring Harbor Meeting. The 1971 announcement in *Nature New Biology* made clear that it would be a data bank of 'coordinates, structure factors and electron density maps'. Tragically Walter Hamilton died of cancer at the age of 41. A postdoctoral colleague of Walter Hamilton took over, Tom Koetzle. Distribution of data started in May 1973.

On page 174 we learn that 'the PDB comprised 84 protein structures in 1976, whereas the Dayhoff atlas of protein sequences comprised 767'. Thus 'a gap was evident in the pace of crystal structure determination of proteins versus sequencing of the amino acids in proteins'. Skip forward to 2021 and we have nearly 200 000 PDB entries (now from X-ray crystallography, NMR and electron microscopy) versus billions of sequences. The gap has become a huge gulf, but the AlphaFold2 deep learning from those experimental crystal structures (Jumper *et al.*, 2021; Tunyasuvunakool *et al.*, 2021) allows prediction of 3D structures from those sequences, including measures of lack of confidence on certain stretches of structure prediction. This is a game changer in capabilities. We must pay tribute to the vision and commitment of the early pioneers of the PDB.

At page 181 the book comes back to ownership of data. In 1980, a large fraction of known, *i.e.* published, structures were missing from the PDB (60 missing versus 145 included), and of those included an even greater proportion were missing their processed diffraction data (75% of the 145). By 1990, through the mandates of NIH (NIGMS) and the IUCr, proof from the PDB of data deposition was required before publication.

Chapter 5 is *Public Databases*. 'By 2005 NIH announced that its gene sequences database GenBank had reached 100 billion sequences, with a doubling every five years.' The recounting of the seminal discussions of a gene sequence database, and a workshop held in the USA in 1979, prominently

features Olga Kennard once more (page 201), quoting her point that 'a database must be run by experts but required the distancing of those experts to simply take advantage of shared data from a community for themselves'. Thus 'the practice must include worldwide data sharing and at minimal cost to the user'.

Margaret Dayhoff reenters the book's narrative in establishing a gene sequence database in the early 1980s, interestingly, 'stressing the need for verification of the data entries by specialists and the authors'. This US-based database was parallel to that of the European Molecular Biology Laboratory. On page 209 the author seems to me to reveal a fundamental misunderstanding of the attempts by Dayhoff to secure the finances for her nucleic acid database: 'Despite the modest amounts, the charges made a crucial symbolic difference between a free public good and a commercial product.' That sentence itself should have provoked a question: did she make a profit or was the work done basically on a charitable basis?

On page 223 is the finale of the outcome of competitive bids, whose details are extensively described because of the major turning point in funding this science field. NIH rejected Dayhoff's proposal and instead gave their USD 3.2M grant to her competitor. So, no justice for the pioneer with the track-record of being 'The World's leading sequence collector for twenty years'.

Chapter 6 is *Open Science*. This chapter tackles the links between 'databases, journals and the gatekeepers of scientific knowledge'. The chapter opening declares a *de facto* rather than in principle statement of what is going on between journals, databases and scientists. *De facto* the situation involved traditions determined by a lack of computers 300 years before when the first journals appeared. With the advent of computers, digital data storage capabilities have changed enormously in a short period of time. The author scrutinizes, within the example of gene sequences, whether it should be, or was, the database or journal that validated the sequence data. The practical barrier of a journal not being provided the sequence as a digital data file to accompany the article submission is portrayed by Strasser as fundamental. As crystallographers we know that a general data validation needs to be accompanied by specialist journal referees checking the article and data before approval of these as versions of record (Hackert *et al.*, 2016).

Page 254 contains the important assertion that 'The rise of open databases has transformed how knowledge is produced in the sciences . . . The significance of a history of collections in the experimental sciences lies exactly here: collections have deeply changed the epistemic practices and the moral and political economy of science.' A different slant would be to say that the feasibility of the human genome sequencing effort was due to technology push and that the publicly funded effort could hardly place humanity's genetic heritage behind a paywall. So, it is technology push that deeply changed epistemic practice not collections *per se*.

The book's narrative comes now to a *Conclusion* chapter. The author dissects the advantages, and challenges, of having grouped the modern big data databases into the broader long-time historical theme of museum and other biological collections. A subsection, *The New Politics of Knowledge*, provides a résumé of the thrust of the funding agencies towards open access to data and publications. Let me state first that he is correct that funding agencies wish to account to taxpayers, but therefore it is surely obvious that open access to the public of a publication they funded the research for should be a core principle. That the funding agencies took so many decades to wake up to this seems to me a strange fact. The author misses major points in this modern development though. Firstly, many of the learned societies saw a way to provide low-cost and properly peer-review-vetted publications and, led by *Acta Crystallographica Section C*, proper peer review of an article with its underpinning data. Learned society journals seem to be perceived, however, as the collateral damage in the battle of the funding agencies with commercial (*i.e.* high-profit) publishers. Another major point missed is that with research proposal rates being so low (at best 25%) researchers have been grateful to learned society journals, and their subscribers, for providing zero-cost-to-author publication, with proper peer review, of their articles. Thirdly, the fact that the Cambridge Structure Database has survived, indeed thrived, for more than 50 years and with more than one million crystal structure entries as a not-for-profit charity by providing subscribers with an expert service has not been covered in this book: something for a second edition maybe. Biology collections by contrast have done relatively well out of the funding agencies, but that is the nature of pursuits like the human genome sequence collection, which are seen even by politicians as a public good.

Overall, whilst I disagreed occasionally with some of the emphases of the author, this a very interesting and well researched book of incredibly broad scope. For that reason, though, it is prone to weakness. Scientists philosophizing may well go astray, but the same is true of philosophers 'sciencizing', to coin a word. On the matter of data science technicalities, data as a word is plural, not singular. Datum is the singular version. Less pedantic is to mention that the Protein Data Bank houses derived molecular models from processed diffraction data, not raw diffraction data, except the Protein Data Bank Japan which has launched an accompanying XRDa raw diffraction data archive.

In conclusion, I recommend this book for its novelty in bringing museums and databases together. It also vividly comes alive with splendid pictures of leading players of each genre at their work. Finally, there is an absolute treasure trove of references and footnotes in this book, comprising 100 pages. It is a work of meticulous scholarship.

References

Crick, F. H. C. & Kendrew, J. C. (1957). *Adv. Protein Chem.* **12**, 133–214.
Giegé, R. (2013). *FEBS J.* **280**, 6456–6497.
Hackert, M. L., McMahon, B., van Meervelt, L. & Helliwell, J. R. (2016). *Open Data in a Big Data World: a Position Paper for Crystallography*, https://www.iucr.org/iucr/open-data.
Hamilton, W. A. (1970). *Science*, **169**, 133–141.

Helliwell, J. R. (2020). *Acta Cryst.* D**76**, 87–93.
Helliwell, J. R. (2021). *Acta Cryst.* A**77**, 173–185.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*, **596**, 583–589.
Nature New Biology (1971). **233**, 223.
Pickstone, J. V. (2000). *Ways of Knowing: a New History of Science, Technology and Medicine*. Manchester University Press.
Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J. & Hassabis, D. (2021). *Nature*, **596**, 590–596.
Zardecki, C. & Burley, S. K. (2021). *Acta Cryst.* D**77**, 1475–1476.