BOOKS *et al.*

DATA MANAGEMENT

# *Sharing data*

## Scientific collections have long been lightning rods for data ownership concerns

*By* **Dov Greenbaum**[1,2] *and* **Mark Gerstein**[2]

Data sharing and ownership issues have flummoxed curators of biological collections for years.

In *Collecting Experiments*, Bruno Strasser posits that biology's increasing emphasis on large databases, best exemplified by the rise of genomics and bioinformatics, is a return to the venerable world of natural history—of collectors, curators, and museums. Beyond biology, the book connects to the broader context of data science emerging within many academic disciplines and throughout modern life. Many of the trumpeted concepts of data science can be seen simply as a rediscovery of existing concepts from traditional fields such as library science, hybridized with computer science and statistics.

Strasser begins by setting up a dichotomy within traditional biological science. On the one hand, there are observationalists—exemplified by natural historians—who classify and compare a wide range of field specimens. On the other are hypothesis-driven experimentalists—exemplified by molecular biologists—who carry out careful laboratory work, often on model organisms. The central character in Strasser's analysis is a hybrid of the two: an individual for whom both molecular experiments and curated collections play synergistic roles. He traces the emergence of such hybrid scientists from early collectors of living organisms to modern managers of biomolecule databases.


**Collecting Experiments**
*Bruno J. Strasser*
University of Chicago Press, 2019. 420 pp.

In the early 1900s, collections were often headed by strong personalities, including Thomas Hunt Morgan, who cultivated fruit flies at Columbia University. Here, Strasser emphasizes the sense of community among contributors to these collections. Without this, such collections would have been unsustainable, especially given collectors' initial failure to credit individual contributors for their submissions.

The experimentalist community, meanwhile, "considered the elucidation of the structure and function of molecules a key intellectual achievement… [and] felt a sense of ownership over the knowledge they had produced." Without the assurance that one would receive credit for a contribution, this group perceived little incentive to participate in shared scientific endeavors, especially when submitting research to a public collection might allow others to extract publishable information. Viable economic models for credit are thus the key to sustainable collections.

The issue of attribution notwithstanding, the strength of the community was often enough to compel early experimenters to contribute to shared resources, at least as long as the contributors were the primary users of the collections. But as submissions grew, expansion put a strain on early collections. The eventual switch to computerized databases cemented the difference between the new hybrid biologist and their naturalist progenitors by vastly expanding the "extent to which the content circulates, the range of people who have access to the collections, and the comprehensiveness of the comparisons."

One exemplary evolution was the Protein Data Bank (PDB), created by experimental biophysicists in 1971. Many experimentalists were initially reluctant to submit data to PDB, despite promises and threats made by the databank and funding agencies. Eventually, however, researchers began to extract scientific value from the collection, particularly with the creation of fold taxonomies.

Ongoing issues related to credit and knowledge creation eventually culminated in 1982 with the development of GenBank, an open-access database of nucleotide sequences. Here, Strasser focuses on the competing efforts of Margaret Oakley Dayhoff and Walter Goad, contrasting Dayhoff's manual annotation style with Goad's automated submission platform. Goad's philosophy of sharing and openness eventually helped overcome ownership concerns, and subsequent databases mandated open access, providing a forum for citizen-science contributions and crowdsourcing in the process.

Databases are still evolving. The variety of big data in the biosciences now includes much more than molecular sequences and structures. One wonders whether Strasser's observations will apply to new modalities, such as GPS coordinates or high-definition images.

It gets even more complicated when information is derived from human subjects. Using Strasser's economic analogy, the ultimate "ownership" and "credit" for the emerging deluge of human-subject data might not belong to the curator or the experimentalist but to the person from whom the data are extracted.

Such questions are ongoing and necessitate continued discussion. But as Strasser has clearly shown, the advancement of biological science depends on working out such data-ownership kinks. ■

[1]*Zvi Meitar Institute for Legal Implications of Emerging Technologies, Radzyner Law School, Interdisciplinary Center, Herzliya, Israel.* [2]*Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. Email: dov.greenbaum@aya.yale.edu*

10.1126/science.aay3820

PHOTO: FLORIDA MUSEUM/ERIC ZAMORA

# Sharing data

Dov Greenbaum and Mark Gerstein

*Science* **365** (6455), 764.
DOI: 10.1126/science.aay3820