Journal of Contemporary Archival Studies

Volume 6 Article 34

2019

Review of Science in the Archives: Pasts, Presents, Futures

Emily A. Graham
Harvard University, egraham2012@gmail.com

Lorraine Daston, ed. *Science in the Archives: Pasts, Presents, Futures.* Chicago: University of Chicago Press, 2017.

The twelve contributors, each a member of the Archives of the Sciences Working Group at the Max Planck Institute for the History of Science, wrote *Science in the Archives: Pasts, Presents, Futures* to demonstrate how scientists employ a variety of methods to preserve data and how those methods impact future research. These essays lead the reader on a fascinating exploration of the history of data preservation, data management, and information organization in the sciences across the *longue durée* and form a well-researched work on the history of science. The volume's contributors consider the sciences inclusive and actively push against the all-too-common perception that this domain of knowledge is uniquely distinct from the humanities; the work of archives and the interdisciplinary nature of primary sources connect practitioners across traditional boundaries of disciplinary knowledge. The essays are organized into four sections that showcase a diversity of materials and methodologies from the past, present, and future. As is true for most archives, archival material produced during scientific study spans a wide variety of material types. In that vein, the contributors strive to uncover the affinities and continuities across archival practices, disciplines, and time, often obscured by the overwhelming noise of material.

The running metaphor of first, second, and third nature, an organizing model established in the introduction, shapes the first section. According to Lorraine Daston, "first nature" refers to natural phenomenon. "Second nature" represents any phenomenon documented or quantified into an information object, like how light through controlled measurement becomes a data point representing an astronomical observation. "Third nature" collates data into an archive that is a precondition for discovery and subject to subsequent recontextualizations as paradigms shift. As paradigms change, so must scientific archives reconfigure: the iterative work of curation and archiving (meaning objects do not experience a single instance of "curation" and then remain static) removes objects from their original contexts and places them in new relational frameworks that "mirror" their natural states. These frameworks shift with the publication of new studies. For example, if a curator organized a natural history collection taxonomically, any phylogenetic reorganization necessitates a mirrored reorganization of specimens. If one agrees with the collection's premise that any aggregate can be an archive—compendia, cuneiform material, astronomical diaries, medical case journals, the earth's crust, the genome—then the authors effectively demonstrate how iterations of organization affect our understanding of the archival material itself. Archivists whose collections do not focus on scientific material may still find this framework useful for considering how our mediation (through finding aids, catalog records, digital exhibits, etc.) can shape how researchers understand objects.

This model continues into chapter 2, David Sepkoski's "The Earth as Archive: Contingency, Narrative, and the History of Life." Here Sepkoski argues that each iteration builds on the previous and offers a linear account of the history of thought, starting with archive_0 (the subscript zero representing its status as the primary or ur-archive), which for paleontologists is the fossil bed itself. Once happened upon by a researcher, the material moves into subsequent archives, from archive_1 through and beyond archive_5, marking epistemological changes in the field. While each change has an epistemological relationship to the previous, each change is also simply a correction or addition to the record (such as reclassifying species) rather than a teleological directive. For

example, Sepkoski writes that *Strata Identified by Organized Fossils* (1916) is an archive$_2$ because, through the process of describing and illustrating the fossils for the 1916 text, the author "transcribed" (archive$_2$) an organized physical archive (archive$_1$) collected from the earth (archive$_0$).[1] These subscripts function as cairns documenting changes in the process of preserving and disseminating scientific thought; the series ultimately provides a framework. According to Sepkoski, archive$_2$ through archive$_5$ (digital databases) are each part of the second nature. In this case, the progression of these subscripts downplays the iterative quality of archival work and risks blurring the line between object and representation. Additionally, the contributors' broad understanding of the word "archive" erases labor—the fossil bed or a book is an archive in metaphor only.

Although she does not use the archive-subscript framework, Florence Hsia's essay better demonstrates the transcription work necessary to preserve data. Astronomy data spans media epochs. Observations began with cuneiform tablets and continue digitally today as part of an attempt to capture all observable celestial phenomena. Caretakers and practitioners learn to transcribe the contents from one medium to another, an activity that covers both migrating the data (whether into a new book or, in the digital age, into a new database) and understanding the calculations behind observations in order to make the data commensurate. These dual acts of transcription restore data to a state of usefulness for astronomers and historians, who can use celestial recordings in manuscripts to date the text based on the determinations from the larger corpus.

As the scope and materials covered in part 1 demonstrate, preservation is an inherently optimistic act. The final essay in this section, "Empiricism in the Library: Medicine's Case Histories" by J. Andrew Mendelsohn, effectively reveals the connective tissue between past, present, and future medical research: publishing case files of unusual symptoms is not an act of naming or curing what ails the patient but an act directed toward the work of future doctors. Some case files remain dormant for decades until a doctor meets a patient with unusual symptoms and searches through the annals for similar occurrences. The initial case files do not name new diseases or provide cures; they are letters to an unknown future, and their value resides in their preservation. The act of placing material in an archive implies a belief that the material has intrinsic value. The essays in section 1 work off that premise and demonstrate that archival material contains data awaiting discovery across disciplines for enterprising researchers. Later essays, like Lorraine Daston's "The Immortal Archive: Nineteenth-Century Science Imagines the Future," carry this theme. In it, she describes how nineteenth-century researchers envisioned archival collections as a precondition for discovery and thus invested time, energy, and resources into large-scale international archives projects.

Daston's essay closes the second section, titled "Spanning the Centuries: Archives from Ancient to Modern." The other two essays from this part, "Archiving Scientific Ideas in Greco-Roman Antiquity" by Liba Taub and "Ancient History in the Age of Archival Research" by Suzanne Marchand, focus on ancient history. Taub describes the development of information storage and retrieval through doxology in Greco-Roman antiquity, while Marchand argues that Leopold von Ranke's scientific history, by positioning source criticism as central to historiographical practices, discredited the study of ancient history as unempirical. Likely meant to demonstrate

[1] William Smith, *Strata Identified by Organized Fossils* (London: W. Arding, 1916).

interdisciplinary affinities in data storage, retrieval, and future use, these essays are fascinating histories but less clearly related to the collection's larger themes.

The volume's third section examines the new challenges faced by data managers in the digital world: the ongoing tension between private and public data, controversy around the human genome, and climate data. While traditional collection development in archives depends heavily on initiation by a donor, public data archives like the Protein Data Bank and GenBank (open-access digital collections of all publicly available protein structures and DNA sequences, respectively) only gained support from researchers when a majority of journals made publication contingent on whether the genetic information had been deposited in such an institution with a corresponding accession number. As Bruno Strasser describes in "The 'Data Deluge': Turning Private Data into Public Archives," three years after GenBank's creation and before journals began to mandate its use, only 19 percent of sequences published the previous year were publicly available in the database. Managers at the Protein Data Bank and GenBank were unable to transform the moral economy and so instead tapped into an existing rewards system. Researchers only began sharing genetic sequences when sharing became a requirement for credit. Considering that acquisition methods do change—for example, some institutions now proactively archive certain ephemeral records (e.g., websites, social media)—perhaps it is worth considering what existing rewards systems are similarly available for our use?

As the first three parts span the centuries and sciences, the final section naturally concludes with the future of data capture, preservation, and storage. While these three essays, written by Rebecca Lemov, Daniel Rosenberg, and Matthew L. Jones, focus on the quantity of data produced and the much-discussed storage, preservation, and access dilemmas, they also throw into sharp relief the data left unmonitored. By deeming such data unimportant detritus (not worth preserving, not worth remembering), we inadvertently reveal aspects of ourselves. In "Archives-of-Self: The Vicissitudes of Time and Self in a Technologically Determinist Future," Lemov discusses how self-archiving and self-tracking applications and projects provide rich data, both in painting a crude picture of the collective psyche and in documenting the temporality of life. Her chapter highlights the uncanny, nostalgic, melancholic fantasy and relentless activity of self-archiving, with every captured data point existing in contrast to what technology cannot capture.

Similarly, in "An Archive of Words," Rosenberg discusses "infraordinary" data through the history of stop lists. Stop lists had their prominence in late twentieth-century digital information systems and are lists of words for a computer to ignore when searching or processing text. Typically, they include articles, pronouns, prepositions, conjunctions, and words so broad as to be useless (e.g., in a medical journal, the words "medicine" and "doctor" might be stop-listed). Considered unimportant, these "infraordinary" lists were typically not saved. The last essay, Jones's "Querying the Archive: Data Mining from Apriori to PageRank," documents the development of data mining at Stanford into the creation of Google and IBM's Almaden research center in San Jose. Using these two examples, Jones demonstrates the initial technical challenge of providing access to huge swaths of large data—a natural outgrowth of the previous chapter's coverage of indexing, with Jones considering association mining and page rank tools. From stop lists to PageRank, the magnitude of digital data necessitates new tools for automation, new skills in data cleaning, and new algorithms for analysis.

The contributors' academic specialties range widely, covering the history of science, comparative literature, natural history, and the life sciences. The predominance of archives users and seeming absence of archivists probably explains why, like the title of this collection, the phrase "science of the archives" remains opaque. It seems to refer to the process of creating and using data for scientific research rather than a scientific methodology or framework for the management and administration of archives. Similarly, the authors present a broad understanding of "archives," generally referring to anything in which data is collected and stored for future research rather than brick-and-mortar establishments. These objects facilitate the retrieval, reconfiguration, and transformation of data, but to call them an archive muddles the archive as a physical place of research and practice, of human governance and human error, and of the biases and value-judgments inherent in acquisition and description.

Science in the Archives: Pasts, Presents, Futures does not focus on archives that collect scientific material or the practice of archivists in those institutions, nor does the book provide guidance for archivists in science-based archives. Perhaps this reflects the nature of a scientific community that has usually circumscribed control of these sources to researchers. While still contending these objects are not archives, they are, indeed, archival in the sense that they have been selected, forgotten, preserved, hidden, and shared, all with a belief in their value for the future. Ultimately, this well-researched collection demonstrates the necessity of archival records to scientific discovery.