

Citation for published work:

Greenberg, J. (2010). *Metadata for Scientific Data: Historical Considerations, Current Practice, and Prospects*. *Journal of Library Metadata*, 10 (2/3): 75-78. [DOI: 10.1080/19386389.2010.520262]

[Pre-print]

Metadata for Scientific Data: Historical Considerations, Current Practice, and Prospects

Jane Greenberg

School of Information and Library Science, University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina, USA

Libraries have always served as repositories for scientific output. The Great Library at Alexandria's collection covered mathematics, physics, biology, astronomy, geography, and medicine. Initiated during the early Ptolemaic period (332–30 BCE), the Great Library at Alexandria emerged as a major center for scholarship, illuminating Alexandria's thriving intellectual and scientific climate. Highlighting three selected scientific advancements of the time, Euclid (c.323–283 BCE) wrote the *Elements*; Eratosthenes (c.276–195 BCE), the third librarian of the Great Library at Alexandria, calculated the earth's circumference; and Hipparchus (c.190–120 BCE) invented Trigonometry.¹ It is easy to imagine the Great Library of Alexandria archiving scientific data supporting and germinating from the work of the many great minds of the period, particularly given our knowledge of and experience with today's great libraries. More difficult is confirming this library practice, given the absence of records. Papyrus, the state-of-the-art recording technology of the time, deteriorates rapidly; and the Roman conquest led to eventual total destruction of the Library. Although evidence is limited, the Great Library at Alexandria's connection to the Museum of Alexandria lends credence to the likelihood that scientific artifacts representing data were among collection holdings.

Ptolemy I was the first person to make a permanent endowment of science. He set up a foundation in Alexandria which was formerly dedicated to the Muses, the Museum of Alexandria (Wells, 1922).

Even with the above quoted knowledge, we must consider that the Greeks' perception of "data" may differ from our modern view of this concept. Our understanding of "data" has been both informed and influenced by the scientific revolution and the scientific method—a standardized process allowing for the verification and repeatability of research.

Although ancient library data archiving practices remain vague, library cataloging practices have been explored and contextualized. Strout (1956) presents an articulate account of the evolution of library

catalogs and codes, beginning with simple lists, extending from the Sumerian to the Roman civilizations; following with inventories, from Middle Ages to the early Renaissance; continuing with finding lists, during the 17th and 18th centuries; and ending with the formalization of codes supporting collocation of the 19th century.

Strout's catalog history primarily focuses on traditional library holdings (the codex to the modern monograph); however, her framework likely reflects some aspects relating to the historical cataloging of scientific data. We may turn to the Renaissance for more insight and explore data catalogs describing the works of Leonardo da Vinci (1452–1519), Copernicus (1473–1543), and Galileo (1564–1642). Major libraries sought and continue to collect archival materials generated by these scientists. Chief goals include archiving, preserving, and providing scholarly access to their invaluable scientific work, including “research data.” In this quest, library practices, particularly cataloging and classification, has had appeal to collection managers and curators.

Today's digital data initiatives integrate library practices—particularly standards/metadata developments, archival routines, and life-cycle management frameworks. In fact, national and international agencies such as the U.S. National Science Foundation (NSF) and the U.K. Joint Information Systems Committee (JISC) encourage research integrating varied information practices; and, they promote collaboration among information professionals and scientists in an effort to deal with today's vast and rapidly growing stores of digital data. The innovation in this area is both exciting and extensive: and the magnitude of activity makes difficult to provide a complete picture of scientific data management practices. At the same time, the enormity of activity underscores the importance of capturing aspects of the time. This special issue of the *Journal of Library Metadata* brings together a series of eight articles capturing today's metadata practices relating to the management of scientific data. The articles are organized into two main themes: The first theme emphasizes current practices in active operational digital repositories and initiatives, and the second theme, while including aspects of current projects, places an emphasis on prospective developments requiring more exploration.

Current Practice: Metadata for Scientific Data

The first group of articles in this special issue begins with Dietrich's account of the DataStaR, a data staging repository at Cornell University Library for curating of scientific research data. This article documents key decisions underlying the architecture, notes Semantic Web development, and describes how a user may interact with the system. Next, San Gil, Hutchinson, Frame, and Palanisamysurvey the National Biological Information Infrastructure's (NBII) contributions to informatics and metadata in the biological sciences. NBII's work in this area provides a robust means for standardizing, sharing,

integrating and synthesizing data. Their work is followed by Richesson, Shereff, and Andrews' collaboration on the [RD] PRISM project and metadata describing a library of standardized question and answer sets supporting rare disease research. Their work presents a project case-driven plan, and underscores the importance of metadata registries, a theme present in the work presented by the NBII authors as well. In the last article of this group, Pilsk, Person, deVeer, Mayr, Furfey, and Kalfatovic present an account of the Biodiversity Heritage Library (BHL). The BHL "is an open access digital library of taxonomic literature, forming a single point of access to ... a worldwide audience of professional taxonomists, as well as 'citizen scientists' " (Pilsk et al., 2010). This collaborative piece provides a history of the BHL, reports challenges and priorities, and presents cataloging/metadata context and solutions underlying the project's success.

Projects and Prospects

The articles in this second group draw upon current initiatives, while exploring future prospects. The first article in this group, by White, reports on a study examining scientists' perceptions and practices relating to metadata in the context of personal information management. This work prompts information professionals to consider their role in understanding the personal practices of researchers, whose data we seek to ingest into open and public information systems. White's work is followed by a collaborative piece that I have co-authored with Deshmukh, Huang, Mostafa, La Vange, Carretta, and O'Neal on the COPD Ontology project. The article explores empowerment via ownership and ontological engineering and as a means for addressing ontology development and maintenance costs. The next article, by Qin and D'Ignazio, presents research on metadata coverage in a science data literacy course. As part of an NSF-funded project at the Syracuse University School of Information Studies, these researchers conducted a faculty survey and a course content scan, and developed a framework for digital data management instruction.

In the final article, Andrew Wilson from the Australian National Data Service presents a thoughtful perspective piece on preserving digital data. Wilson states that "the research community can benefit from the work of the archives community to achieve their shared goal of preserving authentic digital data over time and across domains." Wilson introduces developments in scientific use of digital data, reviews archival perspectives relating to core metadata issues, and highlights the relationship between metadata for recordkeeping and preservation. His work encourages us to think about the value of and need for scientific researchers and archivists to collaborate when designing preservation strategies relating to digital data—as well as any form of information.

References

- Pilsk, S., Person, M. deVeer, J., Mayr, E. Furfey, J. F., & Kalfatovic, M. R. (2010). The Biodiversity Heritage Library: Advancing metadata practices in a collaborative digital library. *Journal of Library Metadata, 10*(2–3), 136–155.
- Strout, R.F. (1956). The development of the catalog and cataloging codes. *The Library Quarterly, 26* (4), 254–275.
- Wells, H. G. (1922). *A short history of the world*. New York, NY: Macmillan. Retrieved from www.bartleby.com/86/
- Wilson, A. (2010). How much is enough?: Metadata for preserving digital data. *Journal of Library Metadata, 10*(2–3), 205–217.

Sources Consulted

Casson, L. (2001). *Libraries in the ancient world*. New Haven, CT: Yale University Press.

Library of Alexandria. Wikipedia: http://en.wikipedia.org/wiki/Library_of_Alexandria

ⁱ Greek Scientists: <http://www.livius.org/gi-gr/greeks/scientists.html>.