Research Data Environment: A View from the Lab Floor

by Stacy Kowalczyk

From Data to Content

Scientific research data can be generated though an experiment or observed via instrumentation; or data can be gathered from existing sources such as government data, vendor data, or web crawls. None of the scientists interviewed was exclusively a creator or a gatherer. Regardless of source, most of the scientists interviewed processed their data by merging data, interpreting and mapping multiple metadata formats, or integrating data with different levels of precision and scale. Data becomes content as value is added via quality control processes, format conversions, contextual data, and structural metadata.

Uniqueness as a Characteristic of Content

Assessment, the process of determining preservation priorities, is a thread through the digital preservation literature. The criterion for assessment generally includes a binary judgment of uniqueness: data is unique and should be preserved; or data is derived, can be recreated, and should not be preserved. The results of this study show that uniqueness is more complicated than previously thought. From the interviews, three levels of uniqueness emerged. The first is the truly unique - no other holdings of this data exist: the preservation worthy data as described in the literature. This type of data is often the results of experiments or observations. The second level is unique to the purpose of the study: while millions of slides of mouse livers exist, none have this specific treatment for this specific research question; or data derived from external sources such as reference collections with this unique analysis. The third level is unique because of the quantity and quality of the data: that is, the level of uniformity and integration of the data, the breadth of data, longitudinal nature of the data, or the added value of metadata. Throughout the literature, the processing of data to create this uniformity or integration is often characterized as simple computation. But to these scientists, the process is very costly in terms of staff, equipment, time, and intellectual effort. It is the processing, both manual and automated, that is unique; thus, all data ultimately becomes unique. It should not be inferred that the scientists wanted all of their data preserved, but it does indicate that uniqueness cannot be the sole assessment criteria.

Formats

Scientific data formats can be proprietary, standard, or pseudo standard. Proprietary formats such as instrumentation data, internal systems data, or vendor data are often migrated to standard formats. Standard formats used by this set of scientist were image standards such as TIFF, JPEG or community XML standards like FITS, BSML, FGDC. Pseudo Standards are generic data standards such as CSV, ascidia files, proprietary SQL or Xpath databases, statistical software formats such as SAS files or Shape and other GIS data formats.

Context

For all most all of the scientists, there is a disconnect between the data and the context of their data – the metadata. Explicit context consists of lab notebooks, data stores in excel or databases, or metadata in community standard formats as discussed above. But much of the contextual data is implicit and is only found in file organization structures or in file naming schemes.

Technical Infrastructure

The largest factor in deciding how scientists dealt with their data was the technical infrastructure of the lab's institution. The scientists in this study were located in three separate universities. The Large Midwest State University had a large computing and storage cloud with no direct cost to researchers. The Medium Midwest State University has a large computing and storage cloud, but costs are allocated by usage to researchers. The Ivy League University required researchers to create the complete computing environment including storage, computing cycle, personnel, space, and electricity. When high quality storage was available at no cost, more data was stored in larger, standard formats. When high quality storage was expensive or had to be created, data was stored in the most economical format, often on removable media like CD-ROMS.

Introduction

Preserving scientific digital data, ensuring its continued access, has emerged as a major initiative for both funding agencies and academic institutions. Providing long-term access to digital data has a number of challenges. Digital data requires constant and perpetual maintenance. Technologies change; equipment ages; software is superseded. Digital data is not fixed and can easily be changed, either intentionally or unintentionally. Much of the research in digital preservation has focused on repositories – systems to manage digital content, to collect and store sufficient technical metadata for preservation, and to manage and initiate preservation actions. This research stream presumes that the data has been either created in a Cyberinfrastructure environment or pushed into a preservation environment and does not address the antecedents to preservation. Yet these antecedents are crucial to the act of preservation. The antecedents to preservation – data management, contextual metadata, and preservation technologies – can also be barriers preventing preservation.

The e-Science: A Data Survey (IU IRB Study Number 06-11593) has been developed to understand more about the issues surrounding the preservation of scientific data. This study has two main goals. The primary goal of this study is to quantify the amount of data to be preserved, the types of data to be preserved and the impact that the loss of this data would have on research. A secondary goal is to understand how scientists perceive the need for data preservation. The data for this study was collected in one hour interviews with 11 researchers in a variety of scientific domains from three different universities using theoretical sampling model of polar examples (Eisenhardt, 1987) : large and small labs, big science and little science (Weinberg, 1961); well funded and poorly funded labs.

Generalizing and Enhancing the Model

A survey instrument has been developed to both generalize and enhance the e-Science data environment model (Study # 1010002804). Constructs to be generalized include data collections, levels of uniqueness, and technical infrastructure. Constructs to be generalized and enhanced are quality control, context, and formats. Additional constructs to be addressed by the model that need to be enhanced are data collections, data management, preservation awareness, and risk assessment.

To generalize the e-Science data environment model, this dissertation proposes to use a broad survey frame of grant awardees of the National Science Foundation (NSF). The Scholarly Database, a collation of data from many sources including journal data such as Medline, Journals of the American Physical Society, and PNAS, as well as funded grants from the National Science Foundation and the National Institutes of Health, will be the source of the sample (LaRowe, Ambre, Burgoon, Ke & Börner, 2009). This database will be queried for all NSF funded Principal Investigators (PIs) from 2007 - 2010. From those PIs with email addresses, a set of approximately 1,200 from each NSF directorate will be selected algorithmically: every 10th name will be chosen. If this results in a sample of less than 1,200, the process will be run again picking every 15th name until 1,200 have been selected. The directorates are the domain specific divisions of NSF each with its own funding initiatives, programs, and management.

References

Atkins, D. E. (2003). *Revolutionizing science and engineering through cyber-infrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*. Directorate for Computer & Information Science & Engineering. Retrieved July 12, 2007 from http://www.cise.nsf.gov/sci/reports/atkins.pdf

Beagrie, N., & Jones, M. (2001). *Preservation Management of Digital Materials: A Handbook*. London: British Library.

Borgman, C.L., Wallis, J.C., Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal of Digital Libraries, 7*.

National Science Board, (2005). *Long-lived digital data collections: enabling research and education in the 21st century*: National Science Foundation.

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of Management Review, 14*, 18.

Gruber, M. (1993). Digital Archaeology. *Wired Magazine, November*.

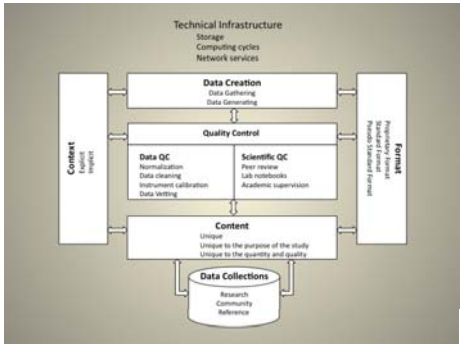
Joint Information Systems Committee. (2004). Data Curation for e-science in the UK. Available from http://www.jisc.ac.uk/whatwedo/programmes/preservation/escience.aspx

Lyon, L. (2007.). *Dealing with Data: Roles, Rights, Responsibilities and Relationships*. Bath, England,: UKOLN.

Ross, S., & Gow, A. (1999). *Digital Archaeology: Rescuing Neglected and Damaged Data Resources*. Glasgow, Scotland: British Library and JISC

Yin, R. K. (2003). *Case Study Research: Design and Methods*. Thousand Oaks, CA: Sage Publications Inc.

The Emerging Model



Data Sources

Data can be created or gathered to be reused. The processes that they use to insure quality of the data depends on the type of data. No scientist is exclusively a creator or a gatherer making the quality process more complicated.

Scientific Quality Control

The scientists had two very distinct understandings of quality. The first was quality of the data. Besides insure that the original data is correct, quality of the data included processes to normalize and "clean" the data to allow accurate merges from disparate sources. The second meaning of quality was of the science. The scientific process has a well established quality control process that we also call "Peer Review. " This process includes vetting by peers, excellent and irrefutable record keeping via lab notebooks, and academic supervision of students.

Data Quality Process

For original Data gathered from equipment, the hardware provides a significant level of quality control. The machines need to be maintained with testing and calibrating. For data gathered from existing sources such as vendors or web crawls, data needs to be manipulated, merged, normalized and "cleaned." The process can introduce errors and needs to have a different level of quality control as described below.

Data Collections

Data developed and vetted for to be shared as Reverence and Community Collections (NSB, 2005) has an extra layer of quality – community vetting. Often these communities are peer, professional academics. But sometimes the communities are ad hoc, vetting data collections via news groups and other social connections or volunteer amateurs contributing to the national database. Trust becomes an important factor in vetting and quality.

November 22, 2010