Planning for the "Long Term"…..in Library Time

Martha Anderson and Jane Mandelbaum
Office of Strategic Initiatives, Library of Congress
Digital Archive Preservation and Sustainability (DAPS) Workshop
MSST2008 25th IEEE Symposium on Massive Storage Systems and Technologies
Baltimore, Md, September 22, 2008

Libraries commonly interpret the "Long Term" in generations, centuries, and millennia. These measures extend back in time to record the histories and provenance of individual works. They look forward in time to the as-yet-unborn readers with needs that we know we cannot imagine. Decision-making in this context has always been difficult, if not impossible. There is always the concern that an individual item which seems ephemeral today may suddenly acquire new and crucial value. Every institution has its favorite examples, from both the past and today. The scientific community, by comparison, is just beginning to take up the challenges of long term preservation and stewardship.

Preservation of valuable content is a challenge. Historically, individual patrons, the Church, and later universities and then governments provided the resources to preserve this content. More recently, communities or networks of institutions developed to provide distributed services or to take advantage of economies of scale and similarities in objectives. In the past, the realities of economic cycles and fiscal decision-making have had adverse impacts on these long term preservation efforts. Often, collections fell victim to war, pestilence, and the inevitable fire. The very recent advent of large scale digital conversion and preservation technologies seems suddenly to have altered the historic operational assumptions. Librarians and scholars now see the possibility of dramatically raising the survivability rate of precious content. Artifacts will continue to fall victim to the inevitable disasters, but it is now possible to imagine the essential content persisting in a virtual manifestation.

Libraries and scientific communities alike are struggling with a familiar array of challenges: How to determine what to preserve; how to describe the content buried in deep archives so that it can be found again; how to identify the best copy and authenticate it against intentional or accidental alteration; how to ensure against undetected "bit rot" over time; how to prevent and detect the most common and devastating of digital collections threats – simple human error; and how to justify the funding of these collections into the future.

All these challenges, and more, are confronting libraries and the scientific community in the same way. Moreover, these same issues now confront businesses, governments, and the individual. Preservation, once the near-exclusive concern of libraries, is now a universal concern. Questions of what to save, how to best preserve it into the future, and how to finance the effort are now nearly universally asked of a much broader set of stakeholders.

Libraries managed to preserve as much as they have due largely to the dedication and proactive efforts of individual librarians/archivists and the institutions which support them. The success of these activities requires, in each generation, an expert understanding of the practices and standards of previous generations and a view to the needs of future researchers. A continuity of standards, and a commitment to the value of those standards, has enabled the relatively smooth handing down of collections, including the knowledge of how to use them, from generation to generation. It seems likely this model of both forward and backward looking custodial stewardship offers important lessons for the future.

The scientific community, focused as it is on data and the application of analysis and experimentation with the data, has an opportunity to apply some of these lessons. At the physical level, all data is the same. Even at the file level there is seemingly little difference between scientific data and library data. Moving fifty terabytes of scientific data is no more difficult or time-consuming than moving the same amount of library data. However, the communities, practices and standards that libraries have developed for organizing and maintaining data for the long term are generally more advanced in terms of consistency and requirements. For example, the creation and preservation of structural and description metadata are critical to the long term survival and integrity of digital assets. Library communities have developed and refined the standards for the capture, recording, validation and exchange of the essential metadata elements. These types of metadata, accompanied by community practices that support them, will be crucial to the maintenance of the massive data stores being built by scientists. Metadata can pass forward, to future custodians, information about file formats, provenance, access rights, retention periods and scores of other critical elements. Maintaining the accuracy and integrity of this information will be critical to the long term preservation of any datasets.

As data collections become larger and are proactively maintained into the future, the focus of effort shifts from creation and collection to preservation. Scientists, and others, have until very recently been focused primarily on the MIPS and bandwidth. They are interested in moving large datasets from source to storage system or from source to processors. The processors have needed increasing power to execute increasingly complex algorithms against larger and larger datasets. Going forward, increasing attention will need to be focused on preservation of the datasets, and standards in the community on what and how to preserve. This will involve more attention and resources for long term activities such as ongoing monitoring of the integrity of the stored data; migration from legacy file formats to currently supportable formats; and migration from one generation of storage media to the next. At some point the level of effort required to maintain the existing data sets exceeds the effort devoted to ingestion and processing of new data. At that point the scientific computing center becomes an archive.

It is well-known that the Library of Congress was built on the collection of books of Thomas Jefferson. Jefferson considered his library to include all that is "chiefly valuable in science and literature generally." (from a letter of September 21, 1814 to Samuel Harrison Smith). Since that time, the national libraries of the United States have grown

significantly in depth and breadth of scientific materials to include the National Library of Medicine and National Agriculture Library.

One commentator after another has authoritatively announced the death of libraries. Librarians themselves fret about this problem individually and in formal groups. Ironically, digital technologies have led to the creation of innumerable new collections of information, including the staggering stores of scientific data. The science of managing and preserving them for future generations of scientists is library science, and libraries will continue to lead and sustain networks for these future generations.
