Abstract

Purpose – For the past decade, as universities have increased their research commitments, the production of large data sets has become prevalent. Up to this point, the storage and curation of these data sets has been somewhat ad hoc and voluntary. Given recent mandatory stipulations coming from government funding sources regarding the handling of data sets, it is imperative that libraries step into this gap and provision data management services for their institutions. This column aims to explore two primary areas in which libraries can provision services for their parent institutions regarding data management.

Design/methodology/approach – The column is exploratory in nature.

Practical implications – As academic libraries take the lead in data management services, there are many positive implications for their parent institutions. Organizing and preserving important data sets could have a significant impact on the worldwide research community.

Originality/value – All academic libraries, no matter their size, have a level of responsibility regarding the collection and curation of data sets. This is a responsibility not only to the local institution, but also to the wider scope of researchers who may make use of those data sets. This column is an exhortation for academic libraries to take the lead in the area of data management.

Keywords Data management, Data sets, Digital curation

Paper type Conceptual paper

The city of Alexandria, Egypt has one of the most interesting histories in the world. Originally founded by Alexander the Great around the year 331 BC, it quickly became a central hub for learning and research in the ancient world. The library of ancient Alexandria is fairly well known in popular culture, but what is perhaps not as well known is that the library was actually a part of a larger complex known as the Mouseion at Alexandria (in Greek Μουσείον τῆς Ἀλεξάνδρειας).

The word mouseion may sound familiar to English speakers, as it is the source of the modern word “museum”. The Mouseion at Alexandria, though, encompassed far more than the current function of museums as they have been understood for the past few hundred years. Originally founded by one of Alexander the Great’s generals, Ptolemy I, the Mouseion was to be modeled loosely after the Academy of Plato as a center of learning and research. It not only housed a vast collection of ancient texts, but was considered a repository of information where scholars debated, lectured, and collaborated. It was not so much a library as we understand it as it was a container of collected research, observations, statistical information, cultural facts and ephemera. There were certainly complete works located in the library, but the Mouseion as a whole like a contemplative
warehouse for scholars to gather in so that they could detect patterns in the collected manuscripts and draw unexpected insights through serendipity.

The stewards of Ptolemy I's brainchild understood the value of the unexpected discovery. They probably would not have labeled their efforts as curatorial, and they had not formulated the ideas that are common today in modern research regarding the analysis of data. They did, however, know that there was wisdom in preserving extensive quantities of important observations over time that may exist in an unorganized form. The availability of seemingly random collected thoughts, lists and quantitative ephemera could be useful to future scholars. Research activities in the ancient were nascent compared to what we know today following the advent of the scientific method. The principle of curating large quantities of collected data, though, was the same. It was perhaps because their methodologies and level of knowledge were still developing that there was an even greater respect for preserving what would otherwise be considered expendable, efforts that could perhaps be considered bordering on a hoarding mentality. It was not until the unfortunate conflagration of 48 BC when Julius Caesar was in Alexandria defending his forces against an impending siege, that the loss of the Alexandrian Mouseion was most keenly felt. Even today, we suffer from this accident, having lost many important ancient manuscripts to the fire, perhaps most notably many of Aristotle’s unique manuscripts which have never been recovered.

The necessity of data preservation and curation, as a concept, is now rather commonplace. There is a common understanding that as research output is frequently born digital, efforts to collect that output need to be scaled up. Granting institutions are now making it a requirement that a commitment is made on the part of the researcher to digitally preserve all of the output of their research. For some time, the necessity of collaboration between researchers on a global scale has been indicated and written about, both in reference to combined data as well as innovations resulting from an increasingly connected data grid (Hey and Trefethen, 2003). The sciences in particular have been a driving factor because cutting-edge discoveries and the complexity of research projects require the cooperation of geographically dispersed research personnel.

This trend has not gone unnoticed by the research library community. The necessity is clearly acknowledged, and while strides have been made, the need has not been fully recognized by parent institutions that have a competitive interest in supplying academic research faculty and personnel with locally-supported curatorial services, particularly from the standpoint of academic libraries. If the lessons of history that were alluded to earlier are any indication, the criticality of such support cannot be neglected.

The North American based Association of Research Libraries has recently proposed a framework for collaboration between research institutions which is currently being referred to under the acronym SHARE which stands for a “shared access research ecosystem” (see www.arl.org/storage/documents/publications/share-proposal-07june13.pdf). This has been proposed in response to a White House memorandum that outlines resources, which are necessary for the continued success of American as well as global research efforts. The resources outlined include infrastructure, digital repositories, Internet 2 resources, and the developing Data Preservation Network (DPN) (see www.dpn.org). It is highly probable that federal grant agencies will begin to mandate the availability of those resources as a stipulation for the awarding of federally grants. If this comes to fruition, it will give proactive academic libraries an advantage.
Curatorial liaisons
An aspiration of academic libraries is to become a destination point regarding the collection and preservation of data sets, and also to be knowledgeable and reliable consultants for campus partners. A well-known barrier to the adoption of data preservation services is ignorance on the part of research faculty regarding the availability of such services. The time required to learn standards and practices associated with data set curation could be substantial, in many cases because these standards are in a state of flux. In the sciences, the generation of vast data sets outpaces the ability of the government and academic associations to wrap comprehensive data standards around curatorial efforts. From a researcher’s point of view, the basic storage of such data, as well as resources for backup and recovery, are sufficient. The power to generate large quantities of valuable data, though, and the ability to digitally store that data is not even a half measure vis-à-vis making that data available to other interested researchers in a manner that would be productive. Libraries are discovering ways to add value to this type of research productivity (Gabridge, 2009). As was mentioned earlier, the amount of time that must be devoted to keeping abreast of the changing standards about research data stewardship, as well as the related financial implications and pressures precipitated by grant funding organizations, can be significant. Information science professionals are bridging this gap through the analysis of the legal, academic and technical requirements for the various types of data generation and storage. The obtaining of grants, especially those sourced by federal funds in the near future will, if they aren’t already, be contingent upon the adherence to strict data management rules.

While institutional computing centers are supporting the high performance computing needs of scientific and engineering research, rarely do they dedicate personnel for assistance with rigorous data management planning. In fact, the importance of intentional data management for academic purposes has increased over the past few years. Both the private sector and the government sector have used aggregated data sets to their advantage, either for commercial gain or national security interests. There is an obvious advantage in being able to analyze large sets of social science data for academic purposes and for the greater good as an alternative to the purely utilitarian purposes of companies such as Google or the US National Security Agency (Lazer et al., 2009). Non-social science fields such as chemistry, physics and biology have been leverage large data set analysis for some time, but this is an area that has not yet been exploited within the more traditional humanities or social science fields.

For long-term data set curation to be successful, specialists who are versed in the handling and formatting of massive data sets should play a much-needed advisory role even as research projects are in their incubation phases. Library liaisons can act in two important capacities in this regard. On the one hand, increasing numbers of library professionals are becoming experts in this area, from metadata specialists to information technology personnel. On the other, librarians are also building data management into traditional instructional contexts outlining best practices for students who are studying in areas that require data set analysis.

Once large data sets are absorbed into an intentionally structured repository, that data needs to be exposed to the academic community in a couple of important ways. First, the data sets need to be organized in an internally and externally cohesive manner. Information architecture and categorization are an obvious forte within the
information science field. Second, once large data sets have been curated, they need to be exposed in a contextually appropriate manner. This second aspect is necessary for data reuse to be successful. Exposure can be accomplished both at the local institutional level as well as via comprehensive aggregations sponsored by national organizations dedicated to the dissemination of vital research data. The role performed by the library acting in this capacity would not only be a benefit to the larger research community but also provide a benefit to local faculty through timesavings.

The manner in which library liaisons provide data management services to their communities will of course vary depending on the subject area. Each discipline will have contextual needs, which are related to their specific methodologies as well as content. The data needs of a high-energy physicist will vary greatly from those of an archaeologist, historian, or linguist. However, there are no longer academic fields that cannot benefit from data management planning and the analysis of large aggregated data sets. It therefore behooves subject liaisons to become educated in data management practices, or at least to become conduits to specialists in this area.

**Technical bridges**

The core services that libraries offer to research communities have evolved over the last decade to include a greatly enhanced technical capacity. There are several areas that continue to grow in order to support the more demanding needs of researchers at academic institutions from the library perspective. An obvious area is that of information technology itself. In order to fulfill the aspirations of libraries to be at the center of institutional academic and research efforts, a higher percentage of staff positions are being occupied by IT professionals such as systems specialists and programmers. This is a fundamental requirement when possible. The demands of data management for an academic setting require multiple disciplines, however, and thus there are several complimentary areas of concern that require specialists outside of the IT field.

The areas of competency required could be divided into four core sets of skills (see the diagram developed by Martin Donnelly, “Core Skills for Data Management” available at: http://data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html). These are data manager, data creator, data librarian, and data scientist. The one of these four that represents the consumer of library services would be the data creator, unless we are referring to metadata specialists. But, while the responsibility of providing comprehensible data resides with the researcher or producer of the relevant data sets, the library can still play a role in educating the academic community on how to organize and serialize data so that it can be effectively used when curated and aggregated.

In the area of data management, personnel assigned would very likely either need to be IT specialists or those very conversant in IT requirements. This area is actually quite vast in its scope. To mention just a few areas of specialization, competencies in large data networks, redundancy, network architecture, storage scalability, back and recovery, as well as data integrity are absolutely necessary. Having these skills in house are of course a benefit, but more importantly, library liaisons are needed who can act as conduits between faculty engaged in research and technologists who can implement sound data handling procedures. Other areas in which libraries can be involved helping to fill in the blanks include data security, fair use, copyright, risk management as well as access policies.
There are often political entanglements associated with the handling of sensitive research data. The interests involved in the handling of data can sometimes be slightly conflicting, and it isn’t always clear where proprietary jurisdiction lies. Even presumably innocuous data sets may need to be carefully managed in order to avoid institutional entanglements, and it isn’t always safe to assume that the data produced is the property of the researcher. Even so, in the best interest of the broader research community, libraries can assist with the positioning of data sets so that they are of maximal use to the local institution as well as the goals of the individual researcher.

The library can also act as an ambassador between interested institutional parties such as the IT department, the office of research, and students in order to make sure that the highest technical standards are followed. Data set management serves not only contemporary concerns but is also intended to establish the viability of that data for perpetuity. Computational analysis for certain research areas requires a high level of fidelity and consistency. This necessitates that the integrity of data be established at the initial stages of a curatorial project and into the future. It isn’t uncommon for modern research projects to involve terabytes of information that may need to be normalized in order to assure highly accurate computational results.

As data sets are created, merged, and stored, a critical mass will eventually accrete. Data extraction is only as effective as the care that was taken in the initial stages of curation. Information science professionals can assist researchers in planning for the merging, analysis, filtering and serialization of data sets in order to repurpose curated data for future research.

Combined with a rich set of technical skills and capacity, the traditional liaison work of libraries can be applied to specific research areas. Subject specialists tend to have deep knowledge within a certain domain, and this combined with the aforementioned technical assets bridges gaps within the academic research community. There is an acute need for information specialists to keep abreast of the wide array of possibilities for data set curation, and merge that knowledge with domain-specific expertise. In many cases, an unacknowledged need for these services exists in institutions, but it is the job of the library to disabuse academic administrators and skeptical research faculty of this misunderstanding in order to help advance the important research goals of the academy.

References


Corresponding author
Robert Fox can be contacted at: rfox2@nd.edu

To purchase reprints of this article please e-mail: reprints@emeraldinsight.com
Or visit our web site for further details: www.emeraldinsight.com/reprints