Transforming Libraries: Big Data

By Mary Ann Mavrinac, Chief Librarian

It is challenging to talk about e-science, e-research and cyberinfrastructure without exploring the topic of ‘big data’ and the implications, opportunities and challenges that surround it for libraries and academia.

As noted in Foreword’s previous article on e-Science, e-research is typically “data intensive, collaborative, geographically dispersed and computationally intense, using large-scale computing resources, infrastructure and the Internet.”¹ ² Digital information is growing at an astonishing rate. There will be five times the amount of bits created in 2012 than in 2008 when 487 exabytes were created. In 2007, for the first time, the amount of bits created exceeded the available storage.³ While not everything is worth saving, preserving research data is a critical issue across national and international spheres.

In their DataShare Project, the Data Information Specialists Committee (DISC-UK) mapped a data sharing continuum of researchers’ data storage methods.⁴ The bottom of the continuum, called “typical status quo,” includes storage devices such as personal hard drives; password protected, networked drives; and networked drives. These data are not openly accessible, not properly stored, and likely not tagged using metadata standards and practices. The “holy grail” at the top of the continuum is high performance computing: analysis tools applied to data over a secure international network; data that

---

are stored, tagged and preserved for long-term access, use and re-use accessible to a broad community of users. The Research Data Strategy Working Group performed a gap analysis of Canada’s publicly-funded research data using the lifecycle of data as a model in relation to 10 indicators,\(^5\) And determined that there were large gaps for six indicators and moderate gaps for four indicators.

The “status quo” storage habits of researchers mean that many data are vulnerable and not easily accessed or shared by anyone other than the researchers involved in a project. Do researchers care about data storage, management, curation and preservation? Do they think about these issues? The answer is likely yes; however, the complexities and cost of managing and preserving large data sets for possible re-use are beyond the interest, available time and skill set of many researchers who focus primarily on researching and publishing their findings. Researchers have many incentives to publish, but few to manage their research data in spite of several funding agency stipulations that researchers include a data management plan in their research proposals.\(^6\)

Why might data management be important to a researcher? An article on the MIT Libraries’ website, “Reasons to Manage and Publish Your Data,” cites several reasons\(^7\), ranging from increasing research visibility to meeting grant requirements to facilitating the re-use of data for knowledge creation. In some cases, there are legal and ethical obligations to preserve data, and most certainly, funders understand the value of data and view it as an asset.\(^8\)


Historically, libraries and librarians have been most active within academia by acquiring, organizing, making accessible and preserving the products of research – that is, post-research publications. These “products” are relatively fixed, easy to locate, text-oriented and typically the outcome of a familiar pattern of publication. Active engagement in the life-cycle of data is another matter. Data are malleable, mobile, easily shared, replicable and at risk of being inaccessible and obsolete if they are not handled using proper metadata standards and practices and are not stored in a trusted digital repository.

The magnitude of the data issue (and the data) requires collaboration by research libraries, researchers, the university, funders, software and hardware technologists, national organizations and policy makers to grapple with the issues of data management, curation, storage and preservation. What skills and abilities do librarians bring to this conversation? Librarians are well placed to discuss with researchers the value of data management and sharing strategies, and to increase their awareness of metadata standards and practices and institutional repositories. Librarians have expertise and experience with institutional repositories, and libraries have a legacy of preserving knowledge for future generations. Librarians advocate open access and, as such, are advocates of the value of open data. Librarians can act as intermediaries, linking faculty with peers in other disciplines to facilitate interdisciplinary research. Librarians are well placed to introduce undergraduate and graduate students to data literacy and data management issues and practices, cultivating in the next generation of researchers the importance of the big issues surrounding ‘big data’.

Many say that the future world will result from the transformative effect of e-research.9 ‘Big data’ underpins e-research, and it must be stored and preserved for future re-use to help solve the big problems facing the world.