INTRODUCTION

Today’s technologies enable new forms of scholarship that are more information-intensive, data-intensive, distributed, collaborative, and multidisciplinary than ever before resulting in large amounts of data in need of organization, archiving, and preservation. As faculty continue to innovate and pursue new avenues in their research, both the technical and human infrastructure will have to evolve with the ever-shifting needs of scholars. These shifting needs and the emerging form of scholarship known as e-science or e-scholarship greatly impact libraries and scholarly communication. The purpose of this FAQ is to help librarians become more aware of the key issues related to the new research that is taking place at UNL. This FAQ is not comprehensive. Furthermore, it is a living document that should change over time.

Although e-science is popularly used throughout the literature, the other commonly used terms, data curation and cyberinfrastructure, include research that is conducted in the social sciences and the humanities. All of these terms have overlapping meanings and there are no universally agreed upon definitions of them. Nonetheless the next section includes the preferred definitions from the Data Working Group.

1. WHAT ARE E-SCIENCE AND DATA CURATION?

E-science is the use of computing infrastructure which supports multidisciplinary and multi-institution collaboration and that allows innovative methods of large-scale research and on-demand access to shared distributed data sources across the sciences.

Data curation is the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purposes and available for discovery and reuse.

E-scholarship is scholarship that is more information-intensive, data-intensive, distributed, collaborative, and multidisciplinary than the scholarship that is not conducted with the support of high performance computing infrastructure known as cyberinfrastructure.

More Definitions can be found at: http://libraries.unl.edu/datacuration#tab2

2. WHAT ARE THE KEY COMPONENTS OF CYBERINFRASTRUCTURE?

Cyberinfrastructure (CI) refers to the integrated hardware, software, services, personnel, and organizations that support research. It is a critical component for a knowledge-based society. The central hardware elements in this infrastructure are high performance computing (HPC) and massive amounts of storage. This can be accomplished through a
variety of means including grid computing where many servers are networked to increase computing capacity and cloud computing where data are stored on a network locally or elsewhere.

On the UNL campus, HPC is supported through the Holland Computing Center, which has significant resources at the Peter Kiewit Institute in Omaha and the Schorr Center at UNL. Personnel are located in each location to assist users, engage students and researchers, and maintain systems. The center provides campus-wide services to researchers who need high performance computing resources. Scientists and engineers who study topics such as nanoscale chemistry, subatomic physics, and meteorology use *PrairieFire*, a supercomputer located in this facility. *PrairieFire* is also used by UNL students for courses. UNL’s Information Services department has identified relatively low cost storage as a priority for researchers with large data sets.

Personnel who are trained to maintain and assist faculty researchers are responsible for developing, deploying, and managing the high performance computation resources, grids, large distributed databases, display and visualization devices, instrumentation, application software, and software that provides security and access control functions. One of the primary roles of administrators on this campus is to bring together stakeholders such as faculty, staff, students, librarians, administrators, and IT professionals to develop a strategic approach to cyberinfrastructure that leverages partnerships across the University in a collaborative environment that facilities cost effective development and support of CI.

3. WHAT IS THE ROLE OF LIBRARIES & LIBRARIANS IN DATA CURATION?

In addition to subject expertise, library staff and faculty possess skills for collecting, preserving, organizing, describing, archiving, and disseminating information and data. In addition, UNL Libraries has the unique ability to build connections between collections and resources that serve as an unbiased central hub for scholarly communication. E-science and the digital humanities are multidisciplinary, by definition, so libraries are an obvious choice for facilitating data curation.

Many liaison librarians have advanced degrees in their respective disciplines in addition to a degree in library science. They have been trained in collecting traditional formats such as monographs and serials. Although building and maintaining collections of datasets present several complexities and challenges, the process of looking at datasets as information assets and exploring what is needed to develop and manage data collections is similar to the traditional collection development practices employed by librarians for decades.

Our role as liaisons between library services and academic departments gives us the opportunity to inform faculty about data curation issues and to assist them in connecting with resources that will help them preserve their data.

4. HOW DOES GOVERNMENT SPONSORED RESEARCH IMPACT DATA POLICY?

Funding agencies, such as the National Science Foundation (NSF) mandate researchers who seek NSF funding to document how they will archive and share data. Therefore, “It is imperative that data be made not only as widely available as possible but also accessible to
the broad scientific communities." ¹ By January 2010, “all [NSF] proposals [must] include a data management plan in the form of a two-page supplementary document.” A data management plan documents the processes for handling the flow of data from collection through analysis. Relevant software and hardware systems along with quality control and validation of these systems are described.

The purpose of the National Institutes of Health (NIH) Public Access Policy is to ensure that the public has access to the published results of NIH funded research. The policy requires scientists to submit final peer-reviewed journal manuscripts, or publications of record that are funded by NIH funds to the digital archive PubMed Central upon acceptance for publication. To help advance science and improve human health, the policy requires that the research is accessible to the public via PubMed Central no later than 12 months after publication. PubMed is a free database accessing the MEDLINE database of citations, abstracts and some full text articles on life sciences and biomedical topics.

5. HOW DO I ENGAGE FACULTY WITH DISCUSSIONS ABOUT DATA?

One of the most effective tactics for learning about the data being produced at UNL is to ask the researcher to tell you about his or her research and then ask him or her to tell you a story about the data behind the research. According to Witt and Carlson of Purdue University Libraries, “It is not a comprehensive strategy but a practical tool to draw out information that needs to be considered in order to understand the data as much as possible...”²

6. WHAT DATA CURATION SERVICES DO THE LIBRARIES PROVIDE?

Repositories for Content

A. Publications & Presentations

Digital Commons <http://digitalcommons.unl.edu>

The Libraries operate the nation’s second largest “institutional repository”—an online, open-access digital archive for research, scholarship, and creative activity by members of the UNL community.

The scholarly communications unit will electronically archive and re-publish the faculty’s published articles, chapters, and reviews, and also any unpublished works, like conference presentations, working papers, seminar papers, teaching materials, and unaccepted manuscripts. Depositing in the “Commons” puts these works online, secures their copyrights, and makes them universally accessible and free to anyone with Internet access.


The Libraries request 1) the author’s permission, and 2) a publication list or vita. The Libraries will clear publishers’ permissions, gather articles, prepare digital versions, and make the deposits. The authors will receive email notification when the works are accessible online as well as monthly email reports that indicate how many times each work is downloaded.

To start participating, email your publication list or vita to the Scholarly Communications Coordinator, Paul Royster: proyster@unl.edu

B. Images

Image & Multimedia Collections <http://0-contentdm.unl.edu.library.unl.edu/>

The Libraries operate Image & Multimedia Collections, a digital collection management software that allows for the upload, description, management, and access to digital collections.

Faculty may use Image & Multimedia Collections images and multimedia for teaching and research; they may also use it to showcase unique collections, when copyright permits. Examples include:

1. The Architecture Program uses images from “African Architecture” for instruction, assignments, and to link images to students from Blackboard and the course syllabus.
2. An art faculty member uses Image & Multimedia Collections to showcase the “Blue Heron Press Collection of Artist Books”
3. An anthropology faculty member uses Image & Multimedia Collections to document his research on the “Cuban Battlefields”

Why Faculty should digitize and add images and multimedia to Image & Multimedia Collections

1. To have image and multimedia collections as well as personal and departmental teaching materials managed.
2. To have content stored, enhanced, and searchable online 24/7
3. To convert analog content into easily accessible digital formats.

How a liaison librarian should jumpstart a project

1. Contact the Image & Multimedia Collections Coordinator. Jolie Graybill: jgraybill3@unl.edu
2. The coordinator will work individually with teaching faculty as well as librarians and staff to design a customized collection to meet the specifications of the faculty member.
3. The coordinator works with library staff to design and establish a collection, and then assist the faculty member as well as the respective department to identify and use
software, and conduct training to manage those collections.

C. Content not appropriate for Digital Commons and the Image & Multimedia Collections

Here are a few examples of data that may not be appropriate:

- Large datasets (e.g. terabytes)
- Data or publications that utilize specialized or proprietary software
- Complex, in flux, or compound data or publications
- Data about human subjects

In such cases, Information Services is responsible for working with researchers on costs as well as methods for storing data and making it accessible. The Libraries may be consulted in the process, particularly in areas related to the organization and description of the data, i.e. metadata support. Contact Ron Roeber, the Director for Research Computing and Data Management roeber1@unl.edu should you have questions or referrals.

D. Archives and Special Collections

The existing hybrid environment entails acquiring analog and digital data worthy of long-term stewardship. The mission of the Archives and Special Collections, University of Nebraska-Lincoln Libraries is to select, preserve, arrange, describe, provide reference assistance for, and promote the use of rare and unique research materials. Contact Mary Ellen Ducey at mducey2@unl.edu should you have questions or referrals.

Metadata Support

The Librarians in the technical services department can add value to the researcher’s data by describing and organizing it for preservation as well as future access. Whenever possible, the metadata specialist should work with the researcher as well as the liaison librarian. For questions about metadata support, contact Elaine Westbrooks: ewestbrooks2@unl.edu.

CONCLUSION

According to D. Scott Brandt, “At the most fundamental level, engaging the library profession in the problem of data management may lead to reframing the values and practices of the library profession. Where today library practices appear to be rooted in the management and delivery of objects (whether virtual or physical), from another point of view those practices are rooted rather in the management and “delivery” of relationships. And data is, after all, an encoding of relationships in the world, whether those relationships involve instruments, physical phenomena, social entities, measurements, time, place, or other intellectual constructs. Whatever the challenges of scientific data for librarians, the potential for these challenges to be both revitalizing and transformative for librarianship is sufficiently attractive to bring many librarians to the data table.

---

3 D. Scott Brandt citing the ACRL Roundtable on Technology and Change in Academic Libraries, November 2-3, 2006.
Conducting a Data Interview  
Reprinted with permission from M. Witt and J. Carlson

1. What is the story of the data? 
Begin the interview with an open-ended question that allows the researcher to talk freely about his or her research, scientific workflow, and community of practice. This lends some insight into the value of the dataset and how it may fit into your collection and be used, and it provides the context for understanding how and why the dataset was created and how it was processed and analyzed.

2. What form and format are the data in? 
What computing environments (e.g., software) are required to use the data? If the data are in proprietary structures, you may consider reformatting them into agnostic formats or ones that can be more easily re-versioned. Is there any existing metadata, either external to the data or description that could be extracted from it? Ideally the data could be described to be discoverable.

3. What is the expected lifespan of the dataset? 
In many cases, there are distinctions in the utility of a dataset as it begins in a raw state and then is analyzed and processed into new forms and versions as a result of different steps in the research workflow. Different entities may have custody of the data and use it for different purposes at different times, affecting its provenance. Funding agencies may require that data be archived for a prescribed period of time or you may forecast its future value and the amount of time it should be retained. The data may be described and archived for effective preservation to ensure its accessibility and integrity over time.

4. How could the data be used, reused, and repurposed? 
This is a primary selection criterion that also impacts how the data are accessed and what policies may be needed to govern its use. As data are archived and shared, new and unintended uses for the data may increase its value. For example, a research dataset may be repurposed as a learning object.

5. How large is the dataset, and what is its rate of growth? 
It is important to quantify the size of the data for storage and network provisioning if you intend to ingest it into a repository. What is its physical (bits) and logical (records) scale? Is the dataset static or dynamic? Ask for a sample of the data to examine.

6. Who are the potential audiences for the data? 
Information regarding potential users of the data and the users’ needs is paramount. Along with potential uses for the data, this is another primary selection criterion. In some cases, the data may need to be embargoed or restricted to a limited group of users who are granted permission to access it.

7. Who owns the data? 
Establishing and maintaining the intellectual property represented by the data should be discussed at the earliest opportunity, and any conflicts should be resolved up-front. Many organizations have a submission policy that asks the contributor to verify that they own the data and have the right to submit it.

8. Does the dataset include any sensitive information?
All data should be reviewed for information that violates confidentiality, such as identification information on human subjects. Data curation activities should be informed by institutional review board requirements.

9. What publications or discoveries have resulted from the data? The researchers may have a bias regarding the importance of their data. The purpose of this question is to establish an objective metric for determining the value of the data for the collection. Different metrics may be more appropriate in determining the selection criteria for different kinds of data and data collections.

10. How should the data be made accessible? There is value in making data accessible using a conventional web-based user interface, but machine-to-machine interfaces should also be evaluated. These methods of access will be informed by the answers to the previous questions, and this question can be asked in an open-ended manner to fill in any g