NAVIGATE THE NET

Donna Berryman and Matthew B. Hoy, Column Editors

e-Science and Data Management Resources on the Web

Sally A. Gore

ABSTRACT. The way research is conducted has changed over time, from simple experiments to computer modeling and simulation, from individuals working in isolated laboratories to global networks of researchers collaborating on a single topic. Often, this new paradigm results in the generation of staggering amounts of data. The intensive use of data and the existence of networks of researchers characterize e-Science. The role of libraries and librarians in e-Science has been a topic of interest for some time now. This column looks at tools, resources, and projects that demonstrate successful collaborations between libraries and researchers in e-Science.

KEYWORDS. Data curation, data life cycle, data management, e-science, institutional repositories

Author. Sally A. Gore, MS, MS LIS (sally.gore@umassmed.edu) is the Head of Research and Scholarly Communications, Lamar Soutter Library, University of Massachusetts Medical School, 55 Lake Avenue N., Worcester, MA 01655.

Comments and suggestions should be sent to the Column Editors: Donna Berryman (dberryman@URMC.Rochester,edu) and Matthew B. Hoy (hoy.matt@mayo.edu).

INTRODUCTION

e-Science has been defined as a new research methodology that utilizes the technological advances available today in science. Networks and data are the two major features that characterize e-Science. Scientists no longer rely solely upon bench-top research to perform original work, but also use computer modeling and simulation programs to test and produce new theories and experimental techniques, often generating and accumulating vast amounts of data (think of the Large Hadron Collider or the Long Term Ecological Research project). Ideally, that data could be shared with other scientists, for reuse and re-analysis, ultimately speeding up the process of scientific discovery. Technology also allows for the easy development of networks not limited by geographical constraints. "The global sharing of data has fostered an unprecedented level of open access among scientists, promoted interdisciplinary teamwork on complex problems, and has enabled other researchers to use data for different purposes than what the originators of the data had envisioned."[1]

The role of libraries and professional librarians in e-Science has been discussed for some time now. Preservation, security, and accessibility are basic principles understood and practiced by librarians; these pertain to all forms of information, whether they be archival materials, books, electronic resources, or data. The opportunity and the challenge that e-Science presents for librarians is in finding new ways to communicate the value of the skills librarians already possess and in developing roles that were previously not associated with librarians. This column will address some of these, as well as examples of tools, resources, and projects that demonstrate successful collaborations between libraries and researchers in e-Science.

UNDERSTANDING THE PROBLEM

Existing data, in light of new methodologies or new knowledge, can provide researchers with countless new scientific discoveries, but that can't happen if the datasets are hidden. A telling example of one of the major problems presented by e-Science is the number of personal researchers who either make their datasets available on their own websites or build long lists of links to datasets that they happen to know about due to their own research. One may be lucky and discover some of these pages through a website like the National Science Digital Library <http://nsdl.org/> or through random searching of the Internet, but more often than not, these valuable resources remain hidden to all but a select few working in the same field. The two websites listed below are examples of the most prevalent way data is shared today, i.e. personal websites maintained by researchers or individual laboratories.

C. Elegans Differntial Gene Expression Database

<http://edgedb.umassmed.edu/IndexAction.do>

The C. elegans Differential Gene Expression Database is a wonderful part of the website dedicated to the work done in the laboratory of Marion Walhout, PhD, Systems Biologist, UMMS. Here one can feely download the datasets from many of her experiments; however, they are simply put on the site as spreadsheets, lacking the necessary metadata for others to easily find.

Favorite Datasets from Early (and Late) Phases of Drug Research

<http://www.math.iup.edu/~tshort/Bradstreet/>

Thomas Bradstreet at Merck Laboratories has a nice listing of links to datasets he calls,

"Favorite Datasets from Early (and Late) Phases of Drug Research." If one is lucky enough to stumble on to this site, one can find datasets from fifteen years of work in several biomedical research areas.

Librarians assisting researchers with data management and data curation, however, can make a bid difference. Their skills in information management, preservation, security, and most importantly, accessibility, help disseminate this valuable resource to a wider audience of researchers faster and easier. Librarians understand the need to apply standards and proper metadata to make data easier to find and easier to access. Librarians, working with data repositories, also tend to group a number of projects together, thus gathering disparate data from individual projects into a centralized, subject-specific place. Again, this makes for easier sharing amongst a group of researchers within a particular field. Finally, librarians plan for long-term preservation of information that researchers may not be thinking about when they begin collecting and sharing data.

DEFINITIONS OF DATA

There exists a general consensus that there is a difference between information and data, yet the distinction is hard to identify due to the many definitions of "information." While "data" has fewer connotations, it is still open to varying understanding. A commonly accepted definition of data is "[a] reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples include sequence of bits, table of numbers, characters on a page, recording of sounds, geological specimen."[2]

Different types of data include observational, computational, and experimental. Similarly, there are different types of data collections including research data, resource data, community data, and reference data. In the discussion of e-Science, it is important to keep in mind that what many might call data others view as information, and vice versa. Thus the open sharing of both becomes all the more critical for the goal of making the scientific process more efficient and allowing discoveries to occur more quickly.

As funding-bodies such as the National Science Foundation (NSF) and the National Institutes of Health (NIH) begin placing greater emphasis upon data management plans within grant proposals, a number of libraries are developing educational and support programs that address data literacy. Much like information literacy, data literacy instruction includes understanding how to retrieve, select, assess, manipulate, and cite data; it also requires an understanding of the life cycle of data (see Figure 1). While researchers are aware of the need for data management plans, their knowledge and expertise in this area is not on par with their subject expertise. Librarians with a good understanding of the data life cycle and what is involved in it can blend this with their existing information organization and management skills to create a relevant and appreciated role in the research process.

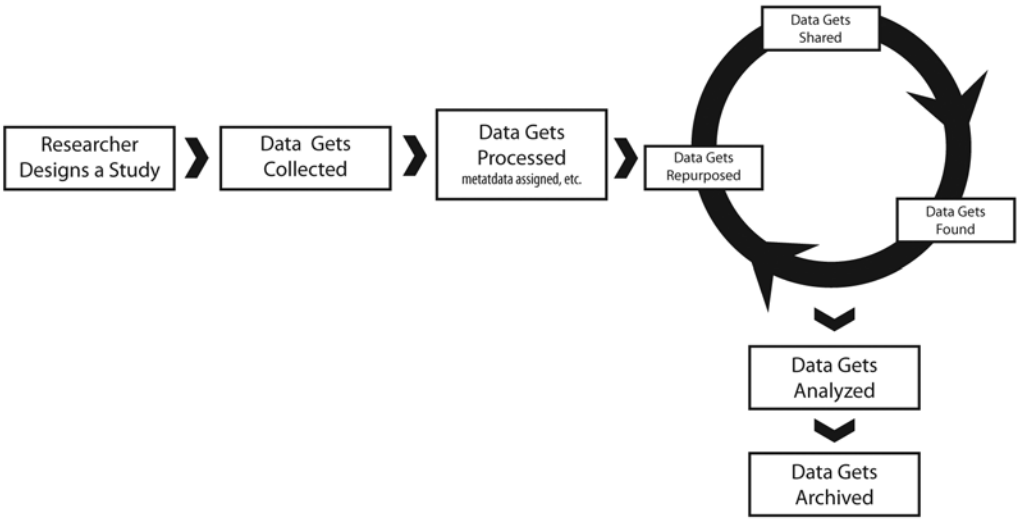


FIGURE 1. Simplified Data Life Cycle

E-SCIENCE AND DATA MANAGEMENT RESOURCES

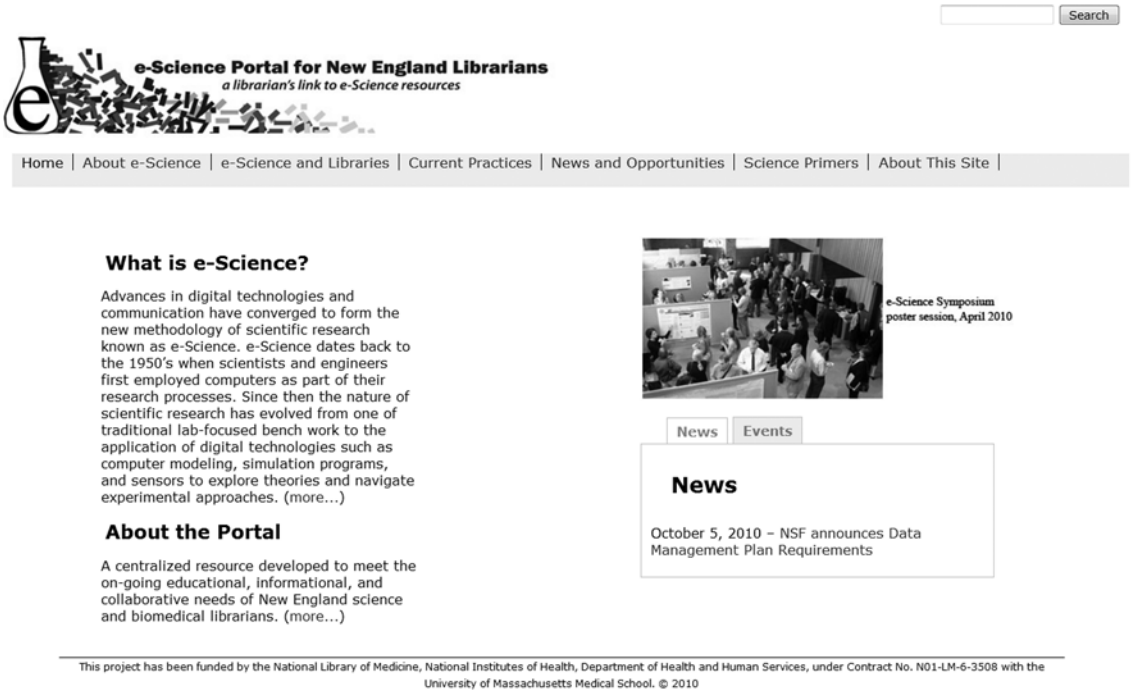
A number of web-based resources exist to help librarians prepare themselves to work in the area of e-Science. These range from sites for continuing and/or basic education in scientific disciplines to those with a focus upon teaching the subject of research methods, including data management principles. Several types of data repository tools are being utilized by libraries to help house data sets both temporarily and as permanent collections. Libraries are active participants in many successful e-Science projects including those associated with Clinical Translational Science Award (CTSA) institutions, federal and state government agencies, and individual laboratories. Finally, e-Science has an impact in scholarly publishing and communication as journals use available technology to enhance what was once a static article. Data sets and other supplemental materials can now be easily embedded in electronic journal articles, adding immense value. Following are examples freely available to view and use as one seeks to become more familiar and skilled in this exciting new role.

Resources for Understanding e-Science

The e-Science Portal for New England Librarians

<http://esciencelibrary.umassmed.edu/>

The Portal is a collaborative project between regional science and medical librarians. It is based at the University of Massachusetts Medical School (UMMS) and funded through a grant from the National Network of Libraries of Medicine, New England Region. The Portal is designed to be a centralized resource "for librarians to lean about and discuss issues related to e-Science, e-Science subject areas, and the impact of e-Science on the profession."[1] It includes resources on education, outreach, best practices, and current events for e-science, as well as basic tutorials on varying scientific disciplines (see Figure 2).



Used with permission, Donna Kafel, Project Coordinator, e-Science Portal for New England Librarians, Lamar Soutter Library, University of Massachusetts Medical Center and NN/LM NER.

FIGURE 2. e-Science Portal screen shot

Library Subject Guides

Lamar Soutter Library, UMMS

<http://libraryguides.umassmed.edu/content.php?pid=77155&sid=571688>

Librarians in the Research and Scholarly Communications Department of the Lamar Soutter Library, UMMS, developed and maintain a subject guide about e-Science. The guide provides researchers with easy access to resources for e-Science in general and data management in particular, including an annotated list of popular, relevant datasets that are available online, news and updates about data management, data sharing, the open data movement, and links to peer-reviewed articles about e-Science and data management.

University of Hawaii at Manoa Library

<http://guides.library.manoa.hawaii.edu/content.php?pid=125160>

The University of Hawaii at Manoa has also developed a subject guide with information and links to resources about data management. This guide provides definitions of data management along with links to major funding-body data management plan requirements (e.g., National Science Foundation and the National Institutes of Health), best practices for research planning, and a detailed description of the data life cycle.

e-Science Graduate Fellows Program

<http://eslib.ischool.syr.edu/>

With support from the Institutes of Museum and Library Services, the Information School at Syracuse University is currently hosting an e-Science Fellows Program in conjunction with its graduate program in Library and Information Science. In addition to their core

curriculum, eight students are taking courses dedicated to data management, database design, and other skills necessary for a practicing e-Science. They also fulfill internship requirements in settings practicing e-Science. A poster presentation, "Educating e-Science Librarians," is also available online at <http://eslib.ischool.syr.edu/presentations/poster151%20educating%20e-science%20librarians.pdf>.

Resources for Understanding Data Management

Online and Classroom Instruction

<http://libraries.mit.edu/guides/subjects/data-management/>

MIT has developed a comprehensive program to educate students and researchers on data management. An outline of the course, "Managing Research Data 101," is available online as well as a subject guide dedicated to the same. The program serves as a model for librarians seeing and taking advantage of an opportunity to provide education in an area previously left untouched, and the resources provided give others a foundation on which to begin.

Data Repository Tools

This section will discuss three different open source options for data repositories: Fedora Commons, DSpace, and REDCap. To search for institutions with open access repositories, using any type of repository software, both open source and third-party hosted, visit the online directory of open access repositories, *Open*DOAR <http://www.opendoar.org/index.html>.

Fedora Commons

<http://www.fedora-commons.org/>

Fedora Commons is open source software that allows users to build customized repositories for a wide range of digital objects. Papers, data sets, images, recordings, and other digital media can be stored, preserved, and disseminated via a well-developed system that links the objects together by the relationships they share. The U.S. National Library of Medicine's Digital Collection of biomedical books and videos <http://collections.nlm.nih.gov/muradora/welcome.jsp> is an example of a digital collection using Fedora.

DSpace

<http://www.dspace.org/>

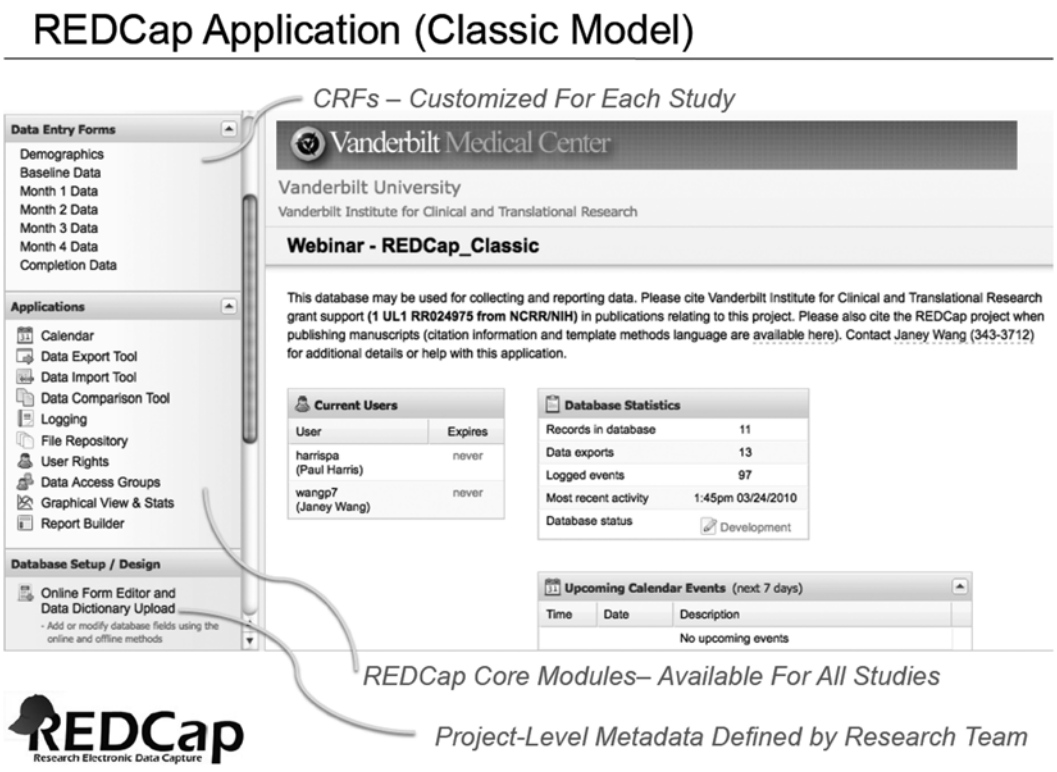
DSpace is another open source product for building digital repositories. Developed by MIT, DSpace has a wide following in academic, non-profit, and commercial organizations. Like Fedora, it allows users to store, preserve, and make accessible all types of digital objects from published papers to data sets. ResearchWorks at the University of Washington <https://digital.lib.washington.edu/researchworks/> and Archie at Kansas University Medical Center <http://archie.kumc.edu/> are two examples of well-developed digital repositories in academic health sciences that use DSpace as their software platform.

REDCap (Research Electronic Data Capture)

<http://project-redcap.org/>

REDCap is one of the newer entries in repository software. Also open source, it was

developed by Vanderbilt University and has as its focus "to support data capture for research studies." [3] Somewhat unique in this regard, REDCap allows users to input experimental data directly into the system, where it can then be analyzed, manipulated, tracked and shared (see Figure 3). REDCap is currently targeted towards institutions specifically working on clinical translational research.



Used with permission by Paul A. Harris, PhD, Office of Research Informatics Operations, Vanderbilt University

FIGURE 3. REDCap Screen Shot

Examples of Successful Data Repositories with Library Connections

GeoMAPP

<http://www.geomapp.net/>

GeoMAPP is a collaborative effort to collect and preserve digital geospatial content. Specifically, GeoMAPP maintains a repository to secure data related to things such as land parcels, zoning, roads, and jurisdictional boundaries. As the site notes, these pieces of information are often subject to change. GeoMAPP makes sure that older versions of this information are preserved for future reference and research. This effort joins together the Library of Congress with three states: North Carolina (the State Archives, the State University Libraries, and the Center for Geographic Information and Analysis); Kentucky (Department of Libraries and Archives, the Division of Geographic Information, and Kentucky State University Library); and Utah (State Archives and Records Services, and the Automated Geographic Reference Center).

Dryad

<http://www.lib.ncsu.edu/dli/projects/dryad/>

Dryad is a joint project of the National Evolutionary Synthesis Center (NESCent) and the University of North Carolina Metadata Research Center, in conjunction with North Carolina State University, the University of New Mexico, and Yale University. The goal of Dryad is to create a repository for data attached to publications in the science disciplines of evolution and ecology. The data is made available so that scientists can both validate findings and use the data for new research.

The Data Staging Repository (DataStaR)

<http://datastar.mannlib.cornell.edu/>

DataStaR is a project of the Albert R. Mann Library at Cornell University. DataStaR is a holding repository, a place where researchers can temporarily store and share data while research is currently underway. Librarians associated with DataStaR also play an integral role in supporting the data management plans of researchers, helping them to think of the entire data life cycle as they begin their work.

The Distributed Data Curation Center (D2C2)

<http://www4.lib.purdue.edu/lcris/edata/>

Purdue University Libraries is a research center devoted to discovering and developing ways to preserve and archive digital data in complex environments. One project within the D2C2 is the eData Repository. In conjunction with its current publications repository, it will allow researchers to find both the data sets and the papers associated with those data sets in the same place.

MIT's GeoData Repository

<http://libraries.mit.edu/gis/data/repository/about.html>

The GeoData Repository allows users to search a vast range of data from census, demographics, property lines, elevations, typographical maps, and more. It also provides downloadable, open source software so that others can use the tool in their own research. Data gathered from geographical information systems (GIS) is extremely valuable in many cross-disciplinary areas of research such as public health and epidemiology.

SUCCESSFUL E-SCIENCE PROJECTS

e-Science for the Public. David McCandless, data miner and information designer, provides consumer health information that illustrates how existing datasets can be repurposed and reused to facilitate discovery. The result is the website "Snake Oil?" <http://www.informationisbeautiful.net/visualizations/snake-oil-supplements/>. By obtaining information from indexed articles in PubMed (an existing dataset), McCandless developed a dynamic, visual representation of the degree of scientific evidence existing for the effectiveness of different dietary supplements on various diseases, providing consumers with a new way to view this information.

e-Science in Government. The United States Department of Health and Human Services recently released a number of public datasets that provide all kinds of health care data. Part of the Open Government Initiative of the Obama administration, this project provides a searchable database containing datasets from multiple government websites <http://www.hhs.gov/open/datasets/index.html>.

e-Science in Clinical and Translational Science. The Institute of Translational Health Sciences (ITHS) is a partnership between several healthcare providers in the Seattle, Washington area, including the University of Washington, Seattle Children's, and the Fred Hutchinson Cancer Research Center <https://www.iths.org/>. ITHS serves as a resource for education and to promote collaboration between biomedical researchers and health practitioners. The program is open to researchers and individuals working in the area of translational research. The ITHS originated and is funded by one of the Clinical and Translational Science Awards of the National Institutes of Health.

The Institute for Health Metrics <http://www.healthmetricsandevaluation.org/> is an independent research center at the University of Washington that also serves as a multipurpose hub of information and services aimed to make access to the necessary tools for research and practice available to a large audience. Researchers can use the resource to both share and find available datasets, plus future research areas and/or partners, through IHME. The results of sharing data at this level can be seen in interactive maps as generated by IMHE.

e-Science in Scholarly Communication

PLoS Computational Biology

<http://www.ploscompbiol.org/home.action>

PLoS Computational Biology couples with RCSB Protein Data Bank <http://www.rcsb.org/pdb/home/home.do> to create two linked databases, one housing publications and the other supporting data. Users can begin discovery through either venue, i.e., searching data and following it to the science from which it originated (published articles) or vice versa.

dbGaP

<http://www.ncbi.nlm.nih.gov/gap/>

NCBI's relatively new resource, the database of Genotypes and Phenotypes (dbGaP) is an archive of research done in the areas of genotype and phenotype. Researchers can upload datasets from their work to dbGaP, to either open or controlled-level access, allowing others to freely search non-sensitive data, as well as summaries of the studies, publications, and other information to aid in further advancing the field.

SUMMARY

Though e-Science is a relatively new method of research, and the role librarians play in it newer still. Evidence exists that those libraries and librarians who develop the necessary skill sets and position themselves strategically can and do find success. Librarians have long been experts in planning, managing, organizing and archiving information. Data is but another form of information and while it is being generated in amounts that exceed our imagination – and often management or storage capabilities – the important roles librarians can play in e-Science are not beyond comprehension. Successful models exist to emulate and explore.

REFERENCES

1. e-Science Portal for New England Librarians. Scope statement. Available <http://library.umassmed.edu/esci_portal_scope.pdf>. Accessed: December 3, 2010.

2. Borgman, C. L. *Scholarship in the Digital Age*: *Information, Infrastructure, and the Internet.* Cambridge, MA: The MIT Press, 2007.

3. REDCap: Research Electronic Data Capture. Introductory Statement. Available: <http://www.project-redcap.org>. Accessed: December 3, 2010.