2011

Essential readings in e-Science

Association of College & Research Libraries - Science & Technology Section, Subject & Bibliographic Access to Science Materials Committee

Essential Readings in e-Science

Compiled and annotated by members of the
ACRL-STS Subject & Bibliographic Access to Science Materials Committee

Kathy Szigeti, Co-Chair
Liaison Librarian (Chemistry; Earth & Environmental Sciences)
University of Waterloo
Waterloo, Ontario, Canada
kszigeti@uwaterloo.ca

Kathy Wheeler, Co-Chair
Electronic Services/Reference
University Library
University of South Alabama
Mobile, Alabama
kwheeler@jaguar1.usouthal.edu

Abstract

The amount of data that scientists produce continues to increase every year. People are needed to handle, preserve, describe, and organize that data, and, because many of these tasks are similar to what librarians have done with publications for centuries, it makes sense that librarians would have a role in the emerging task of managing scientific data. It is the purpose of this paper to give librarians a core set of readings to turn to in order to begin learning about this new task in our field; to help us, as individuals and as a profession, understand what our roles will be in the area of "e-Science."

Introduction

According to an International Data Corporation 2007 white paper, "in 2006, the amount of digital information created, captured, and replicated [worldwide] was 1,288 x 1018 bits. In computer parlance, that's 161 exabytes or 161 billion gigabytes. This is about 3 million times the information in all the books ever written" (IDC 2007). Subsequent revisions to this estimate prove that it only continues to get bigger even more quickly (IDC 2009). Much of this growth of digital information is in the form of data related to science.

Fundamentally, e-Science, as it is understood in this bibliography, follows the United Kingdom's National e-Science Centre's definition: "[T]he large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections,

very large scale computing resources and high performance visualisation back to the individual user scientists" (National e-Science Centre). It is a scientific process that uses global networks of computing infrastructures (often referred to as the "cyberinfrasctructure"), data-gathering instruments, analytic software, and of course, scientists themselves.

Over the past several years, science librarians have been hearing more and more about "e-Science." The Science and Technology Section of the Association of College and Research Libraries held a program at the 2009 ALA Annual Conference, called *{Big Science, Little Science, E-Science: the Science Librarian's Role in the Conversation}*. The Association of Research Libraries held a joint forum with the Coalition of Networked Information in 2008 dedicated to the topic (http://www.arl.org/resources/pubs/fallforumproceedings/forum08proceedings.shtml), and the Graduate School of Library and Information Science at the University of Illinois Urbana-Champaign even offers a {specialization in data curation}, often considered a subfield of e-Science.

But what exactly is it that librarians mean when we speak of e-Science and our role in it? Do we mean data curation only, or are we speaking of something larger, involving collection development, reference services, or bibliographic instruction? Is e-Science a nebulous concept, or can we define it in terms that can be easily understood? What roles do librarians have to play in the growing field of e-Science? The typical professional discourse in this respect continues to include more questions than answers.

This annotated bibliography hopes to give the librarian who is interested in e-Science some resources to consult in order to gain an understanding of both this field and the librarian's role in it. Its intent is not to be comprehensive or definitive, but to provide essential readings that can facilitate a shared understanding in science librarianship regarding e-Science, thereby positioning the profession to begin answering our collective questions.

For clarity, this bibliography is organized into four sections that cover different aspects of e-Science: overview (general information and historical background), data curation (managing data so it is accessible over time), intellectual property (copyright issues), and examples of e-Science in action (practical web sites).

Overview

The Association of Research Libraries Joint Task Force on Library Support for E-Science. [Internet]. Washington, D.C.: The Association. c2007. Agenda for developing E-Science in research libraries; November 2007. [cited 2010 June 4]. Available from: http://www.arl.org/bm~doc/ARL_EScience_final.pdf.

This report provides high-level recommendations on the organizational and strategic directions ARL libraries should take in their support of e-Science. It provides background on the issues surrounding a national cyberinfrastructure and network of distributed scientific communities and their effects on scientific practice, all while focusing on the role of research libraries therein. It suggests there are 11 "model principles" to which ARL libraries should focus their efforts,

including involvement in education, policy, practice, and community building. With respect to philosophy and strategy of the library's role in e-Science, this is a must read.

National Science Board. 2005. Long-lived digital data collections enabling research and education in the 21st century. [cited 2010 June 4] Available from: http://www.nsf.gov/pubs/2005/nsb0540/

Produced by the National Science Board Committee on Programs and Plans, this document exhorts the National Science Foundation (NSF) to coordinate a concerted financial and technological effort to sustain and facilitate access to data produced by NSF-funded research projects. It suggests a clear set of policies for data producers and managers, as well as support of their education, could lead to a more sustainable future for access to scientific data. This report helped shape the NSF's vision for its support of cyberinfrastructure and e-Science.

Gold A. 2007. Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. D-Lib Magazine [Internet]. [cited 2010 June 4]; 13 (9/10). Available from: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html

This article is perhaps the clearest, most interesting and accessible summary in the library and information science literature of the development of e-Science. Gold's explanations of the technological developments that underlie e-Science are particularly interesting and provide context and perspective on the change in scientific practice and methods. She covers everything from the rise of grid computing to data preservation, policy challenges to the business models and economics of e-Science. This part-one-of-two papers lays a solid foundation for the second.

Gold A. 2007. Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: roles and actions for libraries. D-Lib Magazine [Internet]. [cited 2010 June 4]; 13 (9/10). Available from: http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

Building on the context of the first part of the two pieces, this article suggests and provides examples of the current role of libraries in e-Science, particularly projects in the social sciences, GIS, and bioinformatics. It is practical yet far-reaching, and expands upon and complements many of the concepts extolled in the ARL's "Agenda for Developing E-Science in Research Libraries." These two articles are probably the best first reads for anyone interested in understanding e-Science and the challenges and opportunities it provides librarianship as a profession.

Borgman C. The role of libraries in e-Science. European Conference of Medical and Health Libraries [Internet]. 2008 June 23-28; Helsinki,Finland. [cited 2010 June 4]. Video available from: http://blip.tv/file/1038783. PowerPoint Presentation available from: http://www.terkko.helsinki.fi/bmf/EAHILppt/Christine_L_Borgman.pdf

<-- unable to connect to blip.tv 12/22/15 -->
This address, available as a video and in presentation slides, is another good overview of both e-Science and libraries' potential role, replete with some good examples from working data curation partnerships.

Hey T. and Hey J. 2006. E-Science and its implications for the library community. Library Hi Tech [Internet]. [cited 2010 June 4]; 24 (4): 515-528. Available from: http://conference.ub.uni-bielefeld.de/2006/proceedings/heyhey_final_web.pdf

This article is an intense, somewhat technical introduction to the developing interrelationship between scientific research and the library community. Although the article is somewhat dated, the authors examine the driving forces behind e-Science: "the imminent data deluge." Focusing on several studies in the U.S. and Europe, the article walks the reader through the beginnings of e-Science and the growing importance of digital data in future library and scholarly research. The bibliography is useful for the beginner to e-Science.

Lynch C. 2007. The shape of the scientific article in the developing cyberinfrastructure. CT Watch Quarterly [Internet]. [cited 2010 June 4]; 3 (3): 5-10. Available from; http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/

A thorough thought piece on how the practice of science as it relates to e-Science will affect the long-standing unit of communication in science: the "article." Lynch suggests a disaggregation of the article is necessary to facilitate the practice of science in distributed, data-intensive research. The piece, however, does not prescribe how libraries and publishers would effectively disaggregate data from its analytical text.

Rhoten D. 2007. The dawn of networked science. The Chronicle of Higher Education. [Internet]. [cited 2010 June 4]; 54 (2): 78. Available from: http://chronicle.com/article/The-Dawn-of-Networked-Science/36125

A succinct overview of the progression of science from Big Science to what the author terms Team Science and Networked Science. Big Science is exemplified by the Manhattan Project and Team Science by the Human Genome Project. Networked Science exploits cyberinfrastructural developments such as high-performance computing, shared data, and remote testing, all of which enable researchers to work collaboratively across time and space.

Microsoft Research. 2006. Towards 2020 Science. [cited 2010 June 4]. Available at: http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/background_overview.htm

A beautifully produced web publication presenting a vision of science towards the year 2020 and the importance of computational science to that vision, especially as it relates to the natural sciences. The report came out of the Science 2020 Group meeting of more than 30 scientists that took place over three days in July 2005. Sections of the report contain contextual information, illustrations and signed articles from members of the Science 2020 Group. Recommendations, a summary, a glossary and extensive references round out this useful document.

Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. 2009. Harnessing the power of digital data for science and

society. [cited 2010 June 4]. Available at: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA501489&Location=U2&doc=GetTRDoc.pdf

This report outlines a vision and a strategy for a partnership among government, academia, and both public and private research entities that can bolster and support emerging cyberinfrasctructure. Discussions of policy and funding issues abound in the document, and the roles of libraries, museums, and archives are delineated. Representatives from the Library of Congress and the Institute for Museum and Library Services formed part of the working group.

Data Curation

Abbot D. What is digital curation? [Internet]. Edinburgh, Scotland: Digital Curation Centre; c2008 [cited 2010 Apr 26]; [about 6 screens]. Available from: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation

This paper provides a brief but high level introduction to data curation. It defines data curation, describes its value, and introduces issues to be considered. This article is one of many in the Digital Curation Centre series: Introduction to Curation.

Digital Curation Centre. [Internet]. Edinburgh, Scotland: Resources for Digital Curators; c2010. Introduction to curation [cited 2010 Apr 26]. Available from: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

This web site pulls together briefing papers written by various authors that provide introductions to various data curation topics that include annotation, selection, curating different data types, data accreditation, data protection, archiving, repositories, classification, interoperability, persistent identifiers, self-audits, using OASIS, and Web 2.0.

Walters T. 2009. Data curation program development in U.S. universities: The Georgia Institute of Technology example. *International Journal of Digital Curation*. [Internet]. [cited 2010 Apr 27]; 4(3): 83-92. Available from: http://www.ijdc.net/index.php/ijdc/article/viewFile/136/153

This article outlines the experience of the Georgia Institute of Technology in developing a data curation program in the neuro- and biosciences. The author suggests their experience represents that of other U.S. research universities and discusses conditions under which programs develop, including policy, funding, regulations, and current library programs. A model is proposed for U.S. research universities to use as guidance.

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., Chiang, K., Corson-Rikert, J., Hirtle, P., Jenkins, K., et al. 2008. Digital research data curation: overview of issues, current activities, and opportunities for the Cornell University Library. eCommons@Cornell [Internet]. [cited 2010 Apr 27]. Available from http://hdl.handle.net/1813/10903

This article makes a case for increased involvement of academic libraries in the research processes of the university. It defines areas in which libraries have the expertise and

infrastructure to apply to data curation; provides an overview of institutions that are at the forefront, including examples of Cornell's considerable involvement with digital data curation and cyberinfrastructure; and provides recommendations for future activities.

Mullins, J.L. 2007. Enabling international access to scientific data sets: creation of the Distributed Data Curation Center. In: *Proceedings of the 28th International Association of Scientific and Technological University Libraries Conference.* [Internet]. 2007 June 11-14; Stockholm, Sweden. [cited 2010 Apr 27]. Available from http://www.iatul.org/doclibrary/public/Conf_Proceedings/2007/Mullins_J_full.pdf

This article discusses how Purdue University Libraries created the Distributed Data Curation Center (D2C2) in response to the needs of researchers and requirements for National Science Foundation funding.

Witt, M. 2009. Institutional repositories and research data curation in a distributed environment. *Library Trends* [Internet]. [cited 2010 Apr 27]; 57(2): 191-201. Available from:http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1126&context=lib_research

This paper delineates both the research data generation process and the flow of scholarly information, and describes their associated challenges. It depicts how Purdue University Libraries are meeting the challenges by using "distributed institutional repositories" as an information technology solution to store data and support interdisciplinary research. The article briefly describes the utilization of librarians in the process by integrating them into the research domains of the University, and their potential role in data curation.

Choudhury, G.S. 2008. Case study in data curation at Johns Hopkins University. *Library Trends* 57(2): 211-220. Link to abstract: { https://doi.org/10.1353/lib.0.0028}

The author explains in detail the problems and benefits of institutional repositories and how they apply to data sets. While limited to a specific institution, this article raises many issues of the roles of institutional repositories and their roles in academic research, as well as raises the question of what is the role of the new library professional.

Garritano, J.R. and Carlson, J.R. 2009. A subject librarian's guide to collaborating on e-Science projects. *Issues in Science and Technology Librarianship.* [Internet]. [cited 2010 May 24]; 57.Available from: http://www.istl.org/09-spring/refereed2.html

Using a case study, the authors describe their work at Purdue University Libraries in supporting the data needs regarding an e-Science project of the chemistry department. From this experience they identified five skill sets librarians should adapt or develop to participate in e-Science projects. The authors suggest that since librarians transcend disciplinary, technical, and other boundaries, they are well-suited to bring people and resources together in order to conduct interdisciplinary research.

Witt, M. and Carlson, J.R. 2007. Conducting a data interview. Libraries Research Publications [Internet]. [cited 2010 June 10]. Available from: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1092&context=lib_research

In this poster presentation by the authors at the 3rd International Digital Curation Conference on December 12-13, 2007, in Washington D.C., the authors share a set of ten questions for librarians that can be used as a starting point for a "data interview" to identify scientific datasets created by the faculty and researchers at any institution. It is not a comprehensive strategy but instead a practical tool for librarians to capture information about a dataset to evaluate its suitability and requirements for the infrastructure and services that might be needed for data curation.

Borgman, C.L., Wallis, J.C., Enyedy, N. 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal of Digital Libraries* [Internet]. [cited 2010 May 25]; 7: 17-30. Available from: http://escholarship.org/uc/item/6fs4559s#

New technology is transforming little science into big science, causing it to face issues of capturing and managing large amounts of data. This is exemplified in the field of habitat ecology, which this study explores. The authors describe their research on the data practices of the researchers which they can then use to help construct systems that capture and manage data, allowing for its use and reuse in a fair manner.

Brandt, D.S. 2007. Librarians as partners in e-research: Purdue University Libraries promote collaboration. College & Research Libraries News [Internet]. [cited 2010 June 22]; 68 (6). Available from: {http://crln.acrl.org/content/68/6/365.full.pdf+html}

This article defines e-Science in a specific context and explains in detail the need for librarians to be active in collaborating with the researcher and the curator of information within the context of Purdue's experiences as a model.

Intellectual Property

Burk, D. 2007. Intellectual property in the context of e-Science. *Journal of Computer-Mediated Communication* [Internet]. [cited 2010 June 10]; 12 (2). Available from: http://jcmc.indiana.edu/vol12/issue2/burk.html

This article looks at the legal aspects of e-Science. The development of e-Science with global collaboration and access to scientific research via computer networks has challenged the intellectual property rights associated with data, its ownership, and networked collaborative activity. The author focuses on identifying the constraints and possibilities of different open science options through intellectual property rights and licensing. The author suggests using "open source licensing" as a possible solution and finally concludes that the applicability of creative commons models, open source licensing, and patenting models on e-Science practices will require considerable adaptation before they can be applied to e-Science.

Burk, D. 2000. Intellectual property issues in electronic collaborations. In: Koslow S H. and Huerta M F. editors. *Electronic collaboration in science*. Minnesota Legal Studies Research Paper No. 06-66. [Internet]. [cited 2010 June 11]. Available from:http://ssrn.com/abstract=938448

Scientific research is increasingly performed via online collaborations where the development and use of intellectual resources gets distributed among many researchers in a variety of physical locations. Often, the ownership and control of research results is affected by a variety of intellectual property issues. Patent, copyright, and trade secrecy and trademark laws have bearings on virtual research collaborations. Moreover, these laws vary by jurisdiction; the requirements of these laws change from country to country, and may generate conflicts among collaborators. This article alerts collaborative researchers about the potential issues related to intellectual property and identifies trends that may affect the conduct of research. It focuses mainly on U.S. law with appropriate references to issues in other parts of the world.

Keeping science open: the effects of intellectual property policy on the conduct of science. 2003. [Internet]. [cited 2010 June 10]. Available from: http://royalsociety.org/WorkArea/DownloadAsset.aspx?id=5764

This Royal Society working group report on intellectual property rights (IPRs) pertains to patents, copyright, and databases; it considers how IPR use and interpretation could affect British science in the year 2003. Findings indicate IPRs can protect creative work and lead to tangible gains in science but also warn that they restrict free exchange of information/ideas in science. Recommendations include: scientists and funding agencies keep their data publicly accessible; patents be rigorously and thoroughly examined before they are granted; or, scientists publish in low-cost peer reviewed journals with open access policies. The report includes a complete list of recommendations and actions.

David, P., den Besten, M., and Schroeder, R. 2006. How open is e-Science? [Internet] In: E-Science '06: proceedings of the second IEEE international conference on e-Science and grid computing. [Internet] 2006 Dec 4-6; Amsterdam, The Netherlands. Washington (DC): IEEE Computer Society. [cited 2010 June 20]. Available from: {10.1109/E-SCIENCE.2006.261117}

This paper considers "openness" in scientific research and its correlation with eScience. The author discusses not only the technical mechanisms of global collaboration but also the norms and practices of participation in open research processes. He describes the initiatives of the U.K.'s Open Middleware Infrastructure Institute in some detail as well as some "current" [the conference was held in 2006] collaborations in the U.K. One conclusion is that not all collaborative eScience research can be called open science, despite the availability of tools that could support it.

Hahn, K. 2009. Achieving the full potential of repository deposit policies. Research Library Issues [Internet] [cited 2010 June 20]; 263: 24-32. Available from: http://www.arl.org/bm~doc/rli-263-repositories.pdf

Institutions are developing repositories at the same time as national and international efforts, such as ArXiv and PubMed Central, are emerging. The paper explores issues relating to the coordination of content and functions of multiple repositories, focusing on how libraries can develop interactions between their local repositories and PubMed Central through the implementation of the NIH Public Access Policy. Of particular interest is the discussion of assigning limited copyrights to institutions to facilitate deposit on behalf of their authors.

Driscoll, C. NIH data and resource sharing, data release and intellectual property policies for genomics community resource projects. In: Fitzgerald, B, editor. *Legal framework for e-research: realising the potential* [Internet]. Sydney (Australia): Sydney University Press, c2008. [cited 2010 June 20]. Available from: http://www.austlii.edu.au/au/journals/SydUPLawBk/2008/40.html

This article, written by the Director of the National Human Genome Research Institute (NHGRI)'s Technology Transfer Office, NIH, discusses the policies supporting data and research sharing of biological information such as DNA sequence and genomic data. It describes research initiatives, considered to be "community resource projects", implemented by NIH-supported groups such as the International Human Genome Sequencing Consortium (IHGSC) and the Trans-NIH Mouse Initiative. It includes a section on intellectual property considerations in genomics research.

Tonge, A. and Morgan, P. 2007. Project SPECTRa: submission, preservation and exposure of chemistry teaching and research data. JISC Final Report. [Internet] [cited 21 June 2010]. Available from: {https://web.archive.org/web/20110825142618/http://www.lib.cam.ac.uk/spectra/FinalReport.html}

This joint project between the libraries and chemistry departments of the University of Cambridge and Imperial College London (both in the U.K.) aimed to input experimental data into institutional repositories for the purpose of storage, dissemination, and reuse. It found challenges which include dealing with proprietary file types, intellectual property concerns, and the use of paper rather than electronic methods to store data. The report does offer recommendations dealing with disciplinary, technological, and policy issues.

Examples of E-Science in Action

ESDIS: Earth Science Data and Information System Project. [Internet]. [cited 2010 Jun 13]. Available from: http://esdis.eosdis.nasa.gov/

As per the web site, "the ESDIS Project manages the science systems of the Earth Observing System Data and Information System (EOSDIS). EOSDIS provides science data to a wide community of users for NASA's Science Mission Directorate." Individual data centers are responsible for specific disciplines and are responsible for archiving, developing, and distributing data.

Sloan Digital Sky Survey. [Internet] [cited 2010 Jun 13]. Available from: http://www.sdss.org/

A 2.5 meter telescope at Apache Point Observatory in New Mexico surveyed a quarter of the sky over a period of eight years (2000-2008) and the data were subsequently released to the public. This web site provides information about the project and provides access to the data.

CERN. Large Hadron Collider [Internet] [cited 2010 Jun13]. Available from: http://public.web.cern.ch/public/en/LHC/LHC-en.html

The Large Hadron Collider (LHC) is a particle accelerator used to study subatomic particles called hadrons; it generates about 15 petabytes (15 million gigabytes) of data annually. CERN is collaborating with institutions in different countries to operate a distributed computing and data storage infrastructure called the Worldwide LHC Computing Grid. This web site provides detail about the project.

Human Genome Project. [Internet]. Oak Ridge (TN): U.S. Department of Energy Human Genome Program [modified 2008 Aug 19; cited 2010 Jun 13] Available from: http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

The project involved labs from around the world sequencing human DNA and storing the information in databases, making it publicly available. This web site provides information about the Human Genome Project.

Biomedical Informatics Research Network (BIRN). [Internet].[cited 2010 Jun 13] Available from: http://www.birncommunity.org/

This web site describes BIRN, a U.S. "initiative to advance biomedical research through data sharing and online collaboration." Geographically dispersed participants can share large amounts of data across incompatible computing tools using this system.

Bibliography

[IDC] International Data Corporation. The expanding digital universe: a forecast of worldwide information growth through 2010. [report on the Internet]. 2007 [cited 2010 July 13]. Available from: {https://web.archive.org/web/20150404044155/http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf}

[IDC] International Data Corporation. Approaches to storage cost management in the new economy. [report on the Internet]. 2009 [cited 2010 July 13]. Available from: http://www.infortrend.com/End-user_epaper/200907/IDC_AP14941S.pdf

National e-Science Centre. "Defining e-Science." [Internet]. [cited 2010 October 8]. Available from: http://www.nesc.ac.uk/nesc/define.html.