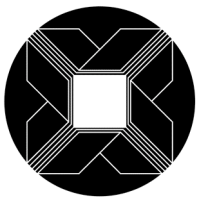
E-Science Talking Points for ARL Deans and Directors

by Elisabeth Jones, University of Washington

with contributions from
Wendy Lougee, University of Minnesota
Neil Rambo, University of Washington
Eric Celeste, Consultant to the ARL E-Science Working Group
and guidance from other members of the ARL E-Science Working Group

October 24, 2008
Association of Research Libraries
http://www.arl.org/rtl/escience/



ARL

TABLE OF CONTENTS

Association of Research Libraries, Washington DC

1. WHAT IS E-SCIENCE (OR E-RESEARCH)?

The term "e-science" is roughly—though not precisely—synonymous with "Cyberinfrastructure;" where the latter term is prevalent in the United States, e-science predominates in the United Kingdom and elsewhere in Europe. Both terms refer to the use of networked computing technologies to enhance collaboration and innovative methods in research. "e-science," however, has a more specific focus on *scientific* research, whereas Cyberinfrastructure is more inclusive of fields outside the sciences and engineering, and incorporates greater emphasis on supercomputing resources and innovation.

Some researchers favor a third term for similar efforts: "e-Research." E-Research is more inclusive of the social sciences and humanities fields, which have also benefited from networked collaboration and investigative resources in recent years.

For e-science in particular, a frequently cited definition appears in a 2006 article by Tony Hey and Jessie Hey:

> e-Science is not a new scientific discipline in its own right: e-Science is shorthand for the set of tools and technologies required to support collaborative, networked science. The entire e-Science infrastructure is intended to empower scientists to do their research in faster, better and different ways.

Further reading

American Council of Learned Societies. *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. 2006. http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf.

Appelbe, Bill, and David Bannon. "Eresearch—Paradigm Shift or Propaganda?" *Journal of Research and Practice in Information Technology* 39, no. 2 (May 2007): 83–90. http://www.jrpit.acs.org.au/jrpit/JRPITVolumes/JRPIT39/JRPIT39.2.83.pdf.

Hey, Tony, and Jessie Hey. "e-Science and Its Implications for the Library Community." *Library Hi Tech* 24, no. 4 (2006): 515–28. http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2380240404.html.

Hey, Tony, and Anne E. Trefethen. "Cyberinfrastructure for e-Science." *Science* 308, no. 5723 (May 6 2005): 817–21. http://www.sciencemag.org/cgi/content/abstract/308/5723/817 (abstract only; full text available with subscription).

Related terms

e-Research, e-Science, Cyberinfrastructure

2. WHAT ARE THE KEY COMPONENTS OF THE DEVELOPING CYBERINFRASTRUCTURE?

Cyberinfrastructure (CI), according to the NSF report that popularized the term, is composed of "hardware, software, services, personnel, [and] organizations" (Atkins, et al, 2003: 13). That is, it incorporates not only physical technologies, but also human processes and social structures; together, these components provide a socio-technical basis for collaboration across geographic, disciplinary, and temporal divides.

A central element on the technical side of this emerging infrastructure is high-performance computing (HPC). HPC involves the use of advanced computing structures with huge amounts of processing power to churn through complex data sets and computational problems. The current state of the art in HPC includes grid computing and cloud computing, the latter built on technical foundations laid by the former.

Still, such computing infrastructure would be useless without the social elements of CI: people to develop useful systems, maintain those systems once built, and work with end-users on employing those systems efficiently in their research work. The idea of CI is to use advanced networking technologies to facilitate collaboration, data management, data analysis, communication, and dissemination across institutional and geographic borders; such technological facilitation will require a significant investment of individual and institutional commitment to system building, maintenance, and support.

Further reading

Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, and Margaret H. Wright. *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* National Science Foundation, January 2003. http://www.nsf.gov/od/oci/reports/toc.jsp. [Note: Appendix A of this report contains an excellent listing of the components of Cyberinfrastructure.]

National Science Foundation, Cyberinfrastructure Council. *NSF's Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, July 20, 2006. http://cyberinfrastructure.us/resources.htm.

Related terms

Cyberinfrastructure, Distributed Computing, Grid Computing, Cloud Computing, List of Distributed Computing Projects

3. WHAT ARE THE MOST RELEVANT AREAS FOR LIBRARY INVOLVEMENT IN E-SCIENCE PROJECTS?

Perhaps unsurprisingly, many of the relevant areas for library involvement in CI projects have to do with managing the large amounts of information that such projects produce. CI projects often reside in departments or institutes that lack any specific data management expertise. An important library role moving forward could be to help such departments and institutes with efficient storage, preservation, metadata creation, and access provision for the data they generate. Beyond this, libraries can develop methods for maintaining increasingly important chains of connection between publications and their data and between data and scientific workflows.

Libraries can provide researchers with valuable policy and content-management consulting services. Librarians increasingly will need to develop expertise in the areas of open access/open data issues, licensing, and data policy management in order to address challenges the we face; this expertise will in turn become a valuable resource for researchers and research teams with questions in these areas. This would build on the expertise librarians have already developed in the area of content management and the implementation of robust models for long-term data preservation such as the Open Archival Information System (OAIS). This base of expertise could help to make librarians an excellent resource for researchers in need of centralized data support for distributed, multi-institutional teams.

The emerging cyberinfrastructure also provides excellent opportunities to build partnerships between libraries and other university units, including science research teams, campus IT, offices of sponsored research, and offices for copyright or rights management. Depending on the particular structures in place at a given university, the library could even come to play a bridging role between different stakeholders in this area, as has occurred in the context of Cornell's VIVO project. Since science is often practiced by teams that cross disciplinary and institutional boundaries, it will also be important for libraries to help their institutions meet the needs of interdisciplinary and multi-institutional research teams.

Further reading

Hey, Tony, and Jessie Hey. "e-Science and Its Implications for the Library Community." *Library Hi Tech* 24, no. 4 (2006): 515–28. http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2380240404.html.

Luce, Richard. "A New Value Equation Challenge: The Emergence of eResearch and Roles for Research Libraries." In *No Brief Candle: Reconceiving Research Libraries for the 21st Century*. Washington, DC: Council on Library and Information Resources, 2008. http://www.clir.org/pubs/reports/pub142/luce.html.

Related terms

Collaborative Working Environment (CWE), Computer-Supported Collaboration (CSC), Computer-Supported Cooperative Work (CSCW), Digital Preservation, Metadata

4. WHAT ARE THE KEY ISSUES SURROUNDING DATA?

Key issues surrounding data and e-science include:

- **Discovery and Identification**: What data exist? Where are the data and how can they be accessed?
- **Access**: Who has access? How will the privacy of both users and research subjects be protected? What kinds of rights management structures need to be established, if any?
- **Interoperability**: In what formats will data be stored and presented? What kinds of metadata will be applied? How will variables be described? What data models apply?
- **Retention Criteria**: Is the data likely to be reused? Will another researcher be able to reasonably replicate or build upon the original results using this data? What is the cost of metadata creation, and how does that compare to the expected value of the data to other researchers?
- **Migration/Preservation**: Will data need to be converted or migrated in order to be usable? Will legacy system configurations need to be preserved or emulated in order to ensure long-term usability of this data?
- **Idiosyncratic practices for data management**: How was the data managed in the laboratory environment? If researchers developed their own ad hoc systems, what impact will this have on how the data will need to be stored for future usability?
- **Culture of "data as private good"**: On what grounds do researchers and institutions object to data sharing? Is there a sense that the data is personally or institutionally owned? Is this the case legally or ethically?

Further reading

Hey, Tony, and Anne E. Trefethen. "The Data Deluge: An e-Science Perspective." In *Grid Computing—Making the Global Infrastructure a Reality*. New York: John Wiley, 2003. http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf.

Lyon, Liz. *Dealing with Data: Roles, Rights, Responsibilities and Relationships (Consultancy Report)*. JISC, 2007. http://www.jisc.ac.uk/publications/publications/dealingwithdatareportfinal.aspx.

To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering. Report of the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, Arlington VA, September 26–27, 2006. http://www.arl.org/pp/access/nsfworkshop.shtml.

Related terms

Data Access, Data Management, Data Sharing, Scientific Data Archiving

5. WHAT ARE SOME EXAMPLES OF LIBRARY INVOLVEMENT IN THE DATA ARENA? WHAT ROLES ARE LIBRARIANS AND LIBRARY STAFF FULFILLING?

Roles that libraries are already playing in the data arena:

- Data management, including collection, organization, description, curation, archiving, and dissemination.
- Creation of new data- and scholarship-based electronic resources for university and/or public use.
- Development of new models, standards, and architectures for various aspects of data management, description, etc.
- Building accessible linkages between all the components and stages of research, from data to researchers to publications.
- Bridging institutional hierarchies and departmental divisions in service of interdisciplinary initiatives.

This is by no means an exhaustive list. A growing body of work assessing possible library roles in e-science and data initiatives, as well as the professional skill base that will be necessary to successfully perform these roles, continues to emerge from libraries and library organizations worldwide, especially in Canada, the UK, and Australia; a few recent exemplars are cited below in greater detail and in the further readings. The NSF sponsored Science Data Literacy project at Syracuse University provides one list of opportunities.

The VIVO project emerged at Cornell in 2003, born out of a set of initiatives geared toward increasing interdisciplinary work at the university. The library, became a leader in this collaborative effort, acting as a bridge between Cornell's strongly hierarchical administration, academic departments, and research centers. In the spirit of this leadership role, the library's Life Sciences Working Group developed VIVO as a discovery tool for both resources and potential collaborators; that is, VIVO includes not only traditional library materials like journal articles, but links from those materials to pages for the faculty who produced them, other materials produced by the same researchers, and events related to the topic area the materials cover. To bring all of these resources together, the architects of VIVO scoured the university for datasets that they could mine and cross-reference. For example, grants information from Cornell's Office of Sponsored Programs, journal citations from BioSis and PubMed, and researcher department and contact information from Cornell's PeopleSoft human resources database all became part of VIVO.

The Distributed Data Curation Center (D2C2) at Purdue had a somewhat different genesis and development. Purdue University has a strong institutional orientation toward science, technology, and engineering disciplines. The D2C2 initiative sprang out of a recognition that the university's librarians were well positioned to help such researchers and interdisciplinary groups manage their data needs. Purdue librarians are tenure track faculty, and this not only gained them credibility among the departmental faculty, but also made it reasonable for them to do things like sign on as co–Principal Investigators for grant proposals requiring a data sharing component. The D2C2 initiative has also led to the creation of tangible technical products such as the distributed institutional repository (DIR) framework, which "supports discovery and access to digital objects of e-research, including data and documents in various forms,

formats and locations," interoperating with other information systems and repositories through an OAI-based architecture. An especially visible output of the D2C2 efforts, Purdue e-Scholar, was built on this DIR framework; it acts as an umbrella service, including a document repository, a special collections repository, and a federation of data repositories.

Further reading

Brandt, D. Scott "Librarians as Partners in e-Research: Purdue University Libraries Promote Collaboration." *College & Research Libraries News* 68, no. 6 (June 2007): 365–7, 396. http://vnweb.hwwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e43f188a0944caaa9d6dbc413a9170af4bc0bf95ef528b3d3322c22330a20c41e&fmt=H.

Devare, Medha, Jon Corson-Rikert, Brian Caruso, Brian Lowe, Kathy Chiang, and Janet McCue. "VIVO: Connecting People, Creating a Virtual Life Sciences Community." *D-Lib Magazine* (July/August 2007). http://www.dlib.org/dlib/july07/devare/07devare.html.

Henty, Margaret. "Developing the Capability and Skills to Support eResearch." *Ariadne,* no. 55 (April 2008). http://www.ariadne.ac.uk/issue55/henty/.

Swan, Alma, and Sheridan Brown. *Skills, Role and Career Structure of Data Scientists and Curators: Assessment of Current Practice and Future Needs.* Report to JISC. Truro, UK: Key Perspectives, 2008. http://www.jisc.ac.uk/publications/publications/dataskillscareersfinalreport.aspx.

Syracuse University School of Information Studies. *The Science Data Literacy Project, Example Job Descriptions.* http://sdl.syr.edu/careers.html.

6. What impact might the rise of "virtual organizations" such as those championed by NSF have on the provision of library services?

Scientists have begun to work across institutional boundaries through inter-institutional or even international "collaboratories," which provide network-enabled environments for executing particular kinds of research. A few examples:

- The Southern California Earthquake Center (SCEC) gathers seismic data from hundreds of scientists at 46 institutions, and provides shared resources like a community modeling environment for visualizing quake impacts.
- The NSF's nanoHUB project provides a venue for sharing nanotechnology research resources, including simulations, presentations, and teaching tools, freely over the TeraGrid, and for communally filtering these resources so that the most useful will "rise to the top."
- The Humanities, Arts, Science, and Technology Advanced Collaboratory (HASTAC) links together a diverse set of more than 80 institutions—from supercomputing centers and grid infrastructure groups to museums and humanities institutes—to support education, archiving, and collaboration among those interested in the historical, social, and humanistic implications of digital technology use.

When research projects are composed of hundreds of researchers from dozens of universities, as many projects supported by virtual organizations are, librarians must work to establish services that are untethered from location, accessible broadly to researchers collaborating over the Internet. Libraries can establish their own presence in the virtual organizations relevant to their institutions (perhaps embedding chat reference services or data or repository linkages on collaboratory sites like nanoHUB), or establish "reference desks" in virtual worlds like Second Life. We can also continue promoting researcher participation in open access repositories, since these help to remove the institutional subscription barriers to electronic resource access, providing a common literature on which multi-institutional collaborations can draw.

More examples of virtual organizations

- Southern California Earthquake Center (SCEC)—http://www.scec.org
- The Cancer Biomedical Informatics Grid (caBIG)—http://cabig.nci.nih.gov
- The Earth System Grid (ESG)—http://earthsystemgrid.org
- The Large Hadron Collider (LHC)—http://lhc.web.cern.ch/lhc/
- nanoHUB—http://www.nanohub.org
- Biomedical Informatics Research Network (BIRN)—http://www.nbirn.net
- Humanities, Arts, Science, and Technology Advanced Collaboratory (HASTAC)—http://www.hastac.org
- The Sloan Digital Sky Survey (SDSS)—http://www.sdss.org
- Second Life (SL)—http://secondlife.com

Further reading

Cummings, Jonathon, Thomas Finholt, Ian Foster, Carl Kesselman, and Katherine A. Lawrence. *Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations*. National Science Foundation, 2008. http://www.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf.

7. WHAT ARE THE DATA POLICIES OF THE MAJOR FUNDING AGENCIES?

Funding agency data policies, especially in the United States, are highly dispersed, variable in their scope and specificity, and in many cases difficult to even locate. Some policies mandate data archiving, while others call only for data sharing; some exist at the highest agency level, while others are specific to particular departments or even specific projects. In May 2008 the president's Office of Science and Technology Policy promulgated "Principles for the Release of Scientific Research Results" that may, in time, drive agencies to develop clearer policy.

The NIH policy is quite detailed for a US agency, but has raised political objections from scholarly science publishers who feel that it tramples their publication rights. The NSF's policy has been less controversial, but remains extremely general, and lacks any specifics on archiving, metadata, or policy enforcement. The earth sciences have a reasonably well-established protocol for data sharing, thanks in part to an existing global system of data centers for this kind of information (and, one suspects, in part to the fact that geospatial data tends not to implicate human subjects issues).

Human subjects issues and proprietary data sets create larger roadblocks to data sharing in other research disciplines, particularly health and social sciences. Nevertheless, the major US federal supporters of these types of research, NIH and NSF, continue to push forward in developing data sharing policies.

Abroad, the situation is quite different. In some countries, data policies have become national priorities: Australia, for example, recently implemented a nationwide mandate for data sharing within state funded research.

Further reading

ANDS Technical Working Group. *Towards the Australian Data Commons: A Proposal for an Australian National Data Service*. Canberra: Australian Government, Department of Education, Science, and Training, October 2007. http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf

National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Foundation, September 2005. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

Office of Science and Technology Policy, Executive Office of the President. "Principles for the Release of Scientific Research Results." May 2008. http://www.arl.org/bm~doc/ostp-scientific-research-28may08.pdf.

A selection of data policies and similar documentation

United States

- Department of Energy
- Department of Justice
- Environmental Protection Agency (for Geospatial Data)
- National Science Foundation (Article 36)
- National Institutes of Health

Europe/Australia/International

- Listing available at SHERPA JULIET, a project of Research Libraries UK (formerly CURL)

8. What is the connection between Open Access and Open Data?

Open Access and Open Data share strong ideological ties, but diverge in the content being shared and the arguments for and against such sharing.

In Open Access, the object of sharing is generally scholarly literature, conventionally defined: that is, journal articles, conference presentations, and other more or less "finished" scholarship. In Open Data the focus is different; as the name suggests, open data policies and initiatives focus on increasing access to *data*—that is, the underlying geospatial codes, laboratory measurements, and other "raw" information produced in the course of conducting research—so that others can review, repurpose, and/or aggregate that information to improve the quality, utility, and reach of the underlying research, or to build it into something new.

Like Open Access, Open Data has proven controversial, yet the sources of controversy differ between the two movements. For Open Access, the most forceful objections have been raised by the existing scholarly publishing industry, who object to policies that they see as a challenge to their business model. For Open Data, the complaints emerge not from the publishing industry, but from researchers and research institutions. The objections raised against Open Data are quite distinct from those leveled against Open Access, among them:

- Having to share data before the individual researcher/research group/institution has fully exploited it might reduce the incentive to produce the data in the first place.

- Different legal systems afford different protections for databases and datasets; effective sharing creates thorny international intellectual property issues, and in some cases may directly clash with particular pieces of database protection legislation.

- Particularly in medical fields and others dealing with human subjects, data sharing creates complicated confidentiality issues.

- The formats of research datasets are insufficiently standardized to enable their integration, and attempting to increase standardization might create a disincentive for healthy variation in methodological choices.

Though the two movements arise from a common desire to broaden access to scientific work, the obstacles that they face—and the parties raising concerns about them—could hardly be more different.

Further reading

Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical, and Medical Publishers (STM). "Databases, Data Sets, and Data Accessibility—Views and Practices of Scholarly Publishers." June 2006. http://www.alpsp.org/ForceDownload.asp?id=129.

European Research Council, Scientific Council. "ERC Scientific Council Guidelines for Open Access." December 17, 2007. http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf.

Freese, Jeremy. "Overcoming Objections to Open-Source Social Science." *Sociological Methods Research* 36 (2007): 220–26. Accessible with subscription from Sage Publications.

Peek, Robin. "Fair Copyright in Research Works Act Challenges Federal Funding." Information Today, September 22, 2008. http://newsbreaks.infotoday.com/nbReader.asp?ArticleId=50849.

Science Commons. "Protocol for Implementing Open Access Data." [Memo.] http://sciencecommons.org/projects/publishing/open-access-data-protocol/.

Related terms

CODATA, Open Data

ARL E-SCIENCE WORKING GROUP (2008-09)

Wendy Pradt Lougee, Chair, University Minnesota

Pam Bjornson, Canada Institute for Scientific and Technical Information (CISTI)

Clifford Lynch, Coalition for Networked Information

Becky Lyon, National Library of Medicine

Carol Mandel, New York University

James Mullins, Purdue University

Gary Strong, University of California, Los Angeles

Betsy Wilson, University of Washington

Eric Celeste, Consultant to the Working Group

ARL Staff Liaisons Crit Stuart & Julia Blixrud