Agenda for Developing E-Science in Research Libraries

Final Report and Recommendations to the Scholarly Communication Steering Committee, the Public Policies Affecting Research Libraries Steering Committee, and the Research, Teaching, and Learning Steering Committee

November 2007

Prepared by the Joint Task Force on Library Support for E-Science

Task Force Members:
Wendy Lougee, Chair (Minnesota)
Sayeed Choudhury (Johns Hopkins)
Anna Gold (MIT)
Chuck Humphrey (Alberta)
Betsy Humphreys (NLM)
Rick Luce (Emory)
Clifford Lynch (CNI)
James Mullins (Purdue)
Sarah Pritchard (Northwestern)
Peter Young (NAL)

Staff Liaisons:
Julia Blixrud (ARL)
Neil Rambo (ARL/Washington)

Table of Contents

Executive Summary

E-science has the potential to be transformational within research libraries by impacting their operations, functions, and possibly even their mission. Recognizing this potential, the ARL Steering Committees for Scholarly Communication and for Research, Teaching, and Learning jointly appointed a task force in 2006 to address the emergent domain of e-science. The Joint Task Force on Library Support for E-Science focused its attention on the implications of trends in e-science for research libraries, exploring the dimensions that impact collections, services, research infrastructure, and professional development. Priorities of government funding agencies further shaped the task force's work.

The task force recommends that ARL establish dedicated capacity within the Association to develop a program agenda over time and to build a shared understanding among the membership of the component issues and challenges for library engagement. In addition to the recently appointed program officers (one permanent and another temporary part-time), the report proposes a working group with an initial charge to develop principles that will inform program development. Anticipated programmatic efforts would emphasize: education of the research library community about scientific trends, the emergent role of data curation, characteristics of virtual organizations, relevant policy for data and research dissemination, and tools and infrastructure systems. While the task force focused on e-science, it was mindful of the broader e-research trends that are shaping research and scholarship in all disciplines.

The task force believes that ARL's engagement in the issues of e-science is best focused on educational and policy roles, while partnering with other relevant organizations to contribute in strategic areas of technology development and new genres of publication. These types of strategic collaborations will also provide opportunities to re-envision the research library's role and contribution as 21st-century science takes shape.

Introduction

In 2006, ARL's Steering Committees for Scholarly Communication and for Research, Teaching, and Learning jointly appointed a task force to address e-science. The charge of the Joint Task Force on Library Support for E-Science focused on raising awareness and positioning research libraries to be players in this new arena. There was a growing sense that e-science trends (and more broadly, e-research trends) were evolving rapidly and libraries could miss opportunities for contribution and engagement as this form of research evolved.

In developing an agenda for the ARL membership, the task force initially concentrated on defining and bounding the arena of e-science. While there was a recognition that there were more general themes of e-research, the group was cognizant of another ARL task force responding to

the ACLS cyberinfrastructure report and its likely emphasis on humanities and social sciences. Consequently, the focus of this task force has remained on the sciences.[1]

The task force's charge provides a mandate to shape an early agenda for developing e-science capacity within research libraries that includes these elements:

- *Develop an understanding within the research library community* about the issues and needs associated with e-science and cyberinfrastructure and the associated needs of scientists and researchers.
- *Recommend approaches to addressing issues related to the curation of long-lived digital data*, including the handling of simulations and storage of massive data sets.
- *Engage ARL members in the development of new roles* for libraries as e-science infrastructure and service needs emerge at research institutions.
- *Identify the skills needed* as information professionals move into the emerging e-science landscape and encourage the development of information professionals prepared to assume new roles.
- *Identify opportunities and recommend strategies for developing relationships* with various government science agencies and other stakeholders such as scientific societies.

The task force did not create its own definition for e-science, a term that has been used to encompass a variety of concepts. The UK National e-Science Centre defines it as "the 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 visualization back to the individual user scientists." For purposes of this report, however, e-science is a more general term taking into consideration not only big computational science, but also team science, and networked science as described in a recent article in the *Chronicle of Higher Education* (Rhoten 2007).

Again in the context of this report, e-science includes all of the natural and physical sciences, related applied and technological disciplines, as well as biomedicine and social sciences that share research approaches with the natural sciences.

The e-science environment is dynamic, multi-sector, and highly heterogeneous. While some core issues are known, there are many models and priorities within different disciplines and interdisciplinary groups. The task force could not reasonably expect to address such a diversity of topics, so it concentrated on teasing out the key issues that will be most critical in helping ARL members to understand this rapidly evolving landscape and in identifying those areas that show potential for research library engagement and collaboration.

The Task Force has identified a list of desired outcomes that will position the research library community as a significant partner in the development of e-science. We encourage ARL to partner with allied organizations and stakeholders to explore and shape a broad vision of the

[1] Consultations with ARL steering committees and with the recently formed (at the time of writing) Special Collections Working Group suggested that there are parallels with the issues associated with special collections. Data curation has some similarities with the issues of archives management.

research library's role in an ideal e-science infrastructure both at the institutional and cross-institutional levels.

The focus of this report, in keeping with the charge of the task force, is on education and policy and communicating with the ARL community about e-science and associated issues. First, we need to build member awareness of the sea change that we see coming as a result of the emergence of e-science, and ensure that members are knowledgeable about the implications for research libraries.

At the same time, however, we are deeply aware that approaching these transformational issues from the constrained perspective of current conditions (e.g., organizational structure, staffing, funding, etc.) will ultimately be insufficient. We need to engage the broader community in a fundamental reassessment of the research library's role and structure, in effect, in redefining the research library for a new era.

Whither Science Libraries?

What are contemporary library uses and needs of the scientific community? Are subscriptions to e-journals sufficient? The emergence of e-science raises important questions about the services and infrastructure within ARL libraries in support of science.

Libraries may, in fact, be creating obstacles to emerging interdisciplinary models of science. Branch libraries based on separate collections in related areas of the sciences are cited as a hindrance to multidisciplinary research at a time when online access transcends discipline-based collections. [2] Other recent behavioral assessments suggest that libraries are often not perceived as part of the evolving research infrastructure in support of interdisciplinary, team science.[3]

There is a perception that science librarians, more than ever before, need to be actively engaged with their user communities. They need to understand not only the concepts of the domain, but also the methodologies and norms of scholarly exchange. This level of understanding and engagement goes well beyond knowledge of the literature. It requires being a trusted member of the community with recognized authority in information related matters. This new paradigm suggests a shift in focus from managing specialized collections (the "branch library" model) to one that emphasizes outreach and engagement.

Many science librarians, of course, are already doing this. There are examples of science and health science librarians working with faculty in teaching courses, participating in research projects, and publishing. Are these models extensible? Can we re-conceive the science library for e-science?

[2] University of Washington Libraries internal study of bioscience faculty and student information needs and uses (unpublished, 2006)

[3] See, for example, University of Minnesota assessment (http://www.lib.umn.edu/about/scieval/) and recent UK report *Researchers' Use of Academic Libraries and their Services* from the Research Information Network and Consortium of Research Libraries (2007), (http://www.rin.ac.uk/files/libraries-report-2007.pdf)

E-Science: Implications for Research Practice

Tony Hey and Jessie Hey (Hey and Hey, 2006) have described e-science as a new research methodology, fueled by networked capabilities and vast amounts of data. E-science departs from well-established experimental and theoretical methodologies with its large-scale, data-driven, and computationally intense characteristics. E-science fundamentally alters the ways in which scientists carry out their work, the tools they use, the types of problems they address, and the nature of the documentation and publication that results from their research. E-science requires new strategies for research support and significant development of infrastructure.

Nearly all aspects of the research library's classic functions and roles are influenced by these new methodologies. Research libraries have traditionally been structured and staffed around disciplines. E-science embraces inter- and multi-disciplinary approaches with significant dependence on computation and computer science. Although technology capacity in libraries has grown considerably in recent decades, it is not of the scale or complexity of the e-science environment. E-science is data-intensive. While research libraries have broadened the scope of information formats they manage and preserve, most have not been responsible for scientific data. E-science is frequently conducted in a team context, with members of the team distributed across multiple institutions and often on a global scale. For libraries, the primary user constituency is generally comprised of those affiliated with the local institution. Further, licenses for electronic content are typically restricted to a particular institutional user community, and the infrastructure to rationalize institutional licenses in a multi-institutional context of team-science is not well developed. E-science challenges these well-established paradigms for library collections and services.

It is worth noting that e-science is not a singular model. Emergent projects suggest there are highly discipline-specific characteristics. Thus, while the ARL task force can provide guidance on the overall attributes of e-science and general trends, it cannot fully prepare libraries to address the challenges that are likely to emerge in any particular project or program.

The task force noted a number of relevant trends at play in the environment. For example, the evolution of distributed and collaborative forces has prompted libraries to be far more engaged in the *processes* of research, integrating content, tools, and services more intimately within scholarly communication workflows. These same forces have enabled library involvement in new forms of scholarly communication, including management of e-print repositories and other new genres of publishing. Clearly, the investments in developing capacity and services for digital libraries are relevant to e-science as well. Just as libraries have contributed to the development of digital content and associated services, there is now the potential for our organizations to contribute essential value and structure to the e-science enterprise by supporting needs for digital content, by creating discovery and management systems for digital data, and realizing new models of support for distributed teams. Library expertise in developing systems and standards for digital content is relevant, as are library roles for stewardship and preservation of content. In short, research libraries are potential partners in e-research, yet our existing expertise and infrastructures will be seriously stretched by the new, more complex demands of e-science.

Research computing organizations are also challenged by these contemporary trends. At an operational level, they are dealing with questions about the appropriate distribution and sustainability of support for research technologies. The focus of these questions is typically on facilities, utilities, and the centralization/decentralization of management. As a result, we see a number of institutions engaged in a review of research technology infrastructure. These reviews typically are confronting the emergent demands of e-science as well, particularly as institutions compete for the new funding made available for cyberinfrastructure and large-scale science projects. The timing of these review processes creates an opportunity for libraries to raise issues related to e-science that will draw on library infrastructure and expertise.

National and International Context for E-Science

US Context & Priorities

The 2003 National Science Foundation (NSF) report, *Revolutionizing Science and Engineering through Cyberinfrastructure,* provided critical documentation of the challenges and opportunities of e-research. It proposed to "use cyberinfrastructure to build more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools, and instruments that operate at unprecedented levels of computational, storage, and data transfer capacity." Further, the NSF report noted that the emerging e-science projects "require effective federation of both distributed resources (data and facilities) and distributed, multidisciplinary expertise, and that cyberinfrastructure is a key to making this possible." (NSF, 2003, p. ES3). Libraries have potential roles to play in both the development of the technology and organizational infrastructure.

A more recent NSF report, *Cyberinfrastructure Vision for 21st Century Discovery,* lays out a more detailed plan of action (NSF 2007). Four interdependent areas of investment are specified:

- **High-Performance Computing**: investments in petascale capabilities for science and engineering.
- **Data, Data Analysis, and Visualization**: investments in data/metadata/ontologies, data collections, and the development of a national digital data framework.
- **Virtual Organizations for Distributed Communities**: investments in tools and technology systems for collaboration as well as evaluative research on the social and organizational dimensions of virtual communities.
- **Learning and Workforce Development**: investments to prepare professionals who will support, deploy, develop, and design cyberinfrastructure.

The research library community has obvious interest in all these areas; the last three offer particular opportunity for library involvement.

The National Institutes of Health (NIH) strategic planning also incorporates themes relevant to the concept and methodologies of e-science. Three priority areas highlighted in the NIH roadmap for medical research capture issues associated with team science, integration of data, and clinical systems (NIH Sept. 2007):

- **New Pathways to Discovery**: this theme addresses issues related to the complexity of biological systems. "Future progress in medicine will require a quantitative understanding of the many interconnected networks of molecules that comprise our cells and tissues, their interactions, and their regulation…. New Pathways to Discovery also sets out to build a better 'toolbox' for medical research in the 21st century." This theme also recognizes the imperative of technology applications, e.g., "libraries of chemical molecules that may provide: probes of biological networks; imaging probes for molecular and cellular events; improved computational infrastructure for biomedical research; nanotechnology devices capable of viewing and interacting with basic life processes; and potential targets for new therapies."

- **Research Teams of the Future**: this theme stresses the imperative of interdisciplinary research and the exploration of new organizational models for team science, including skills and disciplines in both the physical and biological sciences.

- **Re-engineering the Clinical Research Enterprise**: This theme captures the importance of translational research. The challenges here include the development of integrated networks of academic research centers and clinical providers. NIH notes this "vision will require new paradigms in how clinical research information is recorded, new standards for clinical research protocols, modern information technology platforms for research, new models of cooperation between NIH and patient advocates, and new strategies to re-energize our clinical research workforce."

In addition, the Genome Wide Association Studies (GWAS) is a current, prominent, and scientifically promising example of the more centralized approach to creating e-science resources. The data from many of these studies will be entered in the NLM/National Center for Biotechnology Information database, dbGaP (Genotype and Phenotype), which is linked to many other NLM data and literature resources.

Canadian Context & Priorities

Recent activity in Canada has centered on the National Consultation on Access to Scientific Research Data (NCASRD), a partnership initiative of the National Research Council Canada, the Natural Sciences and Engineering Research Council, and the Canadian Institutes of Health Research. The NCASRD final report recognizes the role played by the Canada Institute for Scientific and Technical Information (CISTI), the Canadian Association of Research Libraries (CARL), and the Canadian National Committee for CODATA (CNC/CODATA) in leading efforts for the preservation and access of scientific data. Recommendations encompass a set of 18 actions for digital data that can be taken to begin to address a national strategy for managing research data (http://ncasrd-cnadrs.scitech.gc.ca/).

In other developments, Library and Archives Canada is scheduled to release a Canadian Digital Information Strategy in the fall of 2007 (http://www.collectionscanada.ca/cdis/index-e.html). Resulting from an extensive consultation process, this strategy addresses Canada's need to create, manage, and preserve digital information from its cultural, scientific, and government sectors. Scientific data are recognized as an important element of this national strategy, which

will serve as a foundation for discipline-specific initiatives to preserve and provide access to data.

International Developments

E-science developments and plans in the UK were recently articulated in a major report of an e-science working group, *Developing the UK's e-Infrastructure for Science and Innovation* (National e-Science Centre 2007). Notably, the working group brought together a number of relevant communities: scientists, technologists, librarians, and government representatives. The working group included senior representatives from the Research Councils, Joint Information Systems Committee (JISC), Research Information Network (RIN), and the British Library. JISC's programs, and in particular the Digital Curation Centre (DCC), provide important models for the North American research library community in addressing issues of infrastructure, professional development, and collaboration with the scientific community.

CODATA, the Committee on Data for Science and Technology, is an interdisciplinary Scientific Committee of the International Council for Science (ICSU) working to improve the quality, reliability, management, and accessibility of data of importance to all fields of science and technology. It is concerned with all types of data resulting from experimental measurements, observations, and calculations in every field of science and technology, including the physical sciences, biology, geology, astronomy, engineering, environmental science, ecology, and others. Particular emphasis is given to data-management problems common to different disciplines and to data used outside the field in which they were generated. Established in 1966 by ICSU, CODATA has a history of collaboration on a worldwide basis. (See http://www.codata.org/about/who.html.)

Critical Areas for Research Library Engagement

National funding priorities and research trends suggest several focused areas where the attention of the research library community is warranted.

Data Issues and New Genres of Scholarly Communication

Given e-science's exploitation and generation of a significant volume of digital data, there has been important support by NSF and other agencies to better understand this dimension of cyberinfrastructure. NSF sponsored a workshop in 2005 devoted to data (*Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*, NSB-05-40*)* and in 2006 it supported a workshop convened by ARL to address issues associated with roles and responsibilities of data stewardship. The ARL workshop brought together librarians and scientists, including representatives from several data-intensive programs. The resulting report, *To Stand the Test of Time*: *Long-Term Stewardship of Digital Data Sets in Science and Engineering* (http://www.arl.org/bm~doc/digdatarpt.pdf), explores the potential roles of libraries in managing and preserving data. It also articulates the types of data resources (e.g., reference data sets, data associated with defined discipline communities) and explores potential models for support. For example, the social science research community has a long-standing model in the Inter-University Consortium for Political and Social Research (ICPSR).

The ARL workshop also underscored the complexities introduced by the unprecedented size of data sets, the discipline-specific data models that are required, the desire for re-purpose-able data, and difficulties of identifying a sustainable economic model for long-term preservation of data. This latter point is being explored through a more recent NSF Office of Cyberinfrastructure committee to examine economic models and sustainability for long-term management and stewardship of data. This Blue Ribbon Committee, funded by NSF, the Andrew W. Mellon Foundation, and JISC, will be chaired by Fran Berman (San Diego Supercomputer Center) and Brian Lavoie (OCLC Research).

Data curation brings together roles in preservation as well as active management for access (see Appendix A for a more complete description of data curation). Given the core library mission of preservation and the well-established role in the provision of access to information sources, there is considerable potential for library involvement in data curation. An example, borrowing from archival practices, is the issue of determining policies for selection, access, and length of preservation: What is collected? What is preserved long-term, and what does "long-term" mean? Who can have access to the collection and under what terms of use?

The relationship of data to traditional processes of scholarly communication is another important area for library attention. There is already ample evidence of electronic journals that provide links to the underlying and associated data. There are also emerging examples where the associated data can be accessed and manipulated. An impressive overview of e-science issues, *Towards 2020 Science,* articulates a future where new genres of scholarly publications are born from the relationships between data and the published research record:

> Yet, far from limiting themselves to merely *linking* to databases, scientific journals will in some senses need to *become* databases. Initially this will manifest itself in the way that papers handle accompanying data sets. In the longer term, though, hybrid publications will emerge that combine the strengths of traditional journals with those of databases. We are likely to see a new breed of scientific publication emerge on a timescale of about 10 years that will cater primarily to researchers who wish to publish valuable scientific data for others to analyse. The data will be peer-reviewed and the author will get credit for having published a paper even if the information contained does not explicitly present any new scientific insights. (*Towards 2020 Science,* 2006, p.19)

Support will be needed for text and data mining and computation upon the literature of various disciplines. The integration of this literature into the research and collaboration environment has broad-reaching implications for the nature of licensing agreements that libraries will need to negotiate in the future.

These new genres will challenge existing library descriptive methods and challenge our existing access systems. (See, for example, Hunter [2006] for a description of "scientific publications packages," an approach to encapsulate data, derived products, algorithms, software, and textual publications.) A recent overview by Cliff Lynch (2007) underscores the transformation of the scientific journal underway as well as the critical role these new resources play in virtual communities.

> The implications of this shift are extensive and complex…. First, there will be greater demand for the availability of scientific literature corpora as part of the cyberinfrastructure, and for those corpora to be available in such a way—both technically and in terms of licensing, legal and economic arrangements—so as to facilitate ongoing computation on the literature by groups of collaborating researchers, including groups ("virtual organizations") assembled often fairly casually from across multiple institutions. The barriers here are formidable….

Research libraries have an opportunity to move into publishing roles for new genres of scientific and scholarly communication, given that other traditional players are not doing so.

Virtual Organizations

One of NSF's priority areas for investment is in the development of virtual organizations—that is, online environments that bring distributed researchers together and provide relevant content, data, tools, and services to enable collaborative work. We can anticipate that such communities will require access to relevant information sources, including licensed resources. This new multi-institutional service model will challenge existing approaches to licensing, to access systems, and to user support.

Library involvement with digital repositories provides one area that might lend itself to the evolution of virtual organizations. Discipline-focused repositories could provide an important, core component for the virtual organization to which relevant social tools for collaboration could be added. In addition, libraries may have a role to play in tools for information access, management, and use. Cliff Lynch notes:

> From one perspective, these [virtual] environments are natural extensions of digital library environments, but at least some sectors of the digital library community have always found active work environments to be an uncomfortable fit with the rather passive tradition of libraries; perhaps here the baggage of "digital libraries" as the disciplinary frame is less than helpful. But there is a rich research agenda that connects literatures and evidence with authoring, analysis and re-use in a much more comprehensive way than we have done to date; this would consider, for example, the interactions between the practices of scholarly authoring and communication on one hand, and on the other, the shifting practices of scholarship that are being recognized and accelerated by investments in e-science and e-research. (Lynch 2005)

A sampling of recent NSF requests for proposals indicates the emphasis on the development of cyberinfrastructure for discipline/interdisciplinary communities. For example, NSF is investing in a Plant Science Cyberinfrastructure Collaborative, Environmental Cyberinfrastructure, and an Arctic Cyberinfrastructure and Sensor Program. The community-based Data Interoperability Networks (INTEROP) Crosscutting Program supports community efforts to provide for broad interoperability through the development of mechanisms such as robust data and metadata conventions, ontologies, and taxonomies. Support is provided for Data Interoperability Networks that will be responsible for consensus-building activities and for providing the expertise necessary to turn the consensus into technical standards with associated implementation tools

and resources. Examples of the former are community workshops, Web resources such as community interaction sites, and task groups. Numerous research projects are already underway, targeting specific communities. Notably, a similar diversity of project and program titles, representing all areas of science, emerges when focusing on data management or data curation.

Policy Development

Government agencies such as NSF and NIH play a key role in setting policy. One area of particular resonance for the research library community relates to data policies. For example, NSF's current position on data indicates "all science and engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected and preserved. Through a suite of coherent policies designed to recognize different data needs and requirements within communities, NSF will promote open access to well-managed data…. In addition to addressing the technological challenges inherent in the creation of a national data framework, NSF's data policies will be designed as necessary to mitigate existing sociological and cultural barriers to data sharing and access…." (NSF 2007).

Regarding the data-intensive and data-driven aspects of e-science, NIH policy supports the concept of sharing data that is produced as a result of NIH-funded projects. NIH policy states that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. In the realm of e-science, the lack of access to primary research data impedes the progress of replicating, and possibly extending, research results. NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers and NIH-funded projects with direct costs of $500,000 or more in any single year are expected to include a plan for data sharing and dissemination.

Similarly, the policy for sharing of data obtained in NIH-supported or NIH-conducted Genome-Wide Association Studies (GWAS) indicates that "the NIH expects that NIH-supported genotype-phenotype data made available through the NIH GWAS data repository and all conclusions derived directly from them will remain freely available, without any licensing requirements, for uses such as, but not necessarily limited to, markers for developing assays and guides for identifying new potential targets for drugs, therapeutics, and diagnostics. The intent is to discourage the use of patents to prevent the use of or block access to any genotype-phenotype data developed with NIH support." (See NIH Aug. 2007 for the recently released NIH-wide policy for data access and GWAS.)

Into this mix, the US House of Representatives in July 2007 approved language supporting public access to the results of research funded by NIH. The House FY2008 Labor, HHS and Education Appropriations Bill directs NIH to require its funded researchers to deposit copies of eligible manuscripts into PubMed Central immediately upon acceptance by a peer-reviewed journal, to be made freely accessible to the public no later than 12 months after publication. The Senate will be voting on the appropriations legislation in the fall. The discussions and lobbying surrounding the inclusion of that language has raised awareness among researchers, publishers, libraries, and others in the scientific community about the management and control of the results of scientific research.

Just as the open access movement has prompted libraries to engage in policy discussions about open and sustainable access to scientific journal literature, so too the open data movement will prompt libraries to understand the implications and advantages of models that encourage unfettered access to data, where appropriate. SPARC has been a leader in raising awareness of the need for open access to support the sharing, review, and publication of research results. Efforts are also taking shape through programs such as the Science Commons (http://sciencecommons.org) to provide licensing models that remove barriers to the sharing of information, tools, and data within the scientific research cycle. Related issues emerge with respect to authentication and authorization issues necessary for data that require protection or filtered access.

Current Library Capability to Support E-Science

A case can be made that research libraries already have existing capacity and expertise that they can bring to bear to support e-science. By virtue of their experience in service and data management, and, for many, their mission, they are capable of advising and developing infrastructure to support the needs of scientists working in a cyberinfrastructure-enabled environment. For example, research libraries have:

- Expert understanding of the policies and principles related to open exchange of scholarly information, as well as the roles that can be played by institutional repositories in assuring that exchange, and a demonstrated ability to offer and support both institutional repositories and domain repositories (e.g., arXiv)

- Experience with developing and supporting integration and interoperability tools for information distribution and discovery, both within and across disciplines (e.g., SFX, metasearch, metadata standards, SIMILE)

- Experience with developing and supporting both business and technical strategies for long-term archiving (e.g., archival support generally, Portico, grant research with NARA, NDIIPP)

- Understanding of archival and life-cycle aspects of scientific information, including the importance of assuring access and usability over the long term (preservation, metadata)

E-Science Task Force Recommendations: Desired Outcomes, Strategies, Actions

In an early inquiry to the ARL membership by the task force Chair about ARL library roles in e-science, the responses indicated a limited amount of current activity in this area. During the time of the task force's work, however, there have been rapid developments in new technologies, research protocols, and discipline-specific projects. While libraries may have little immediate engagement in these processes, clearly e-science has the potential to be transformational within research libraries by impacting their operations, functions, and possibly even mission.

In order to build knowledge and engagement within the research library community, the e-science task force has identified desired outcomes for a focused ARL program and potential strategies and initial actions to achieve these outcomes. These recommended steps focus on critical education to build an informed research library community and on ARL involvement in the policy dimensions of e-science. These arenas are clearly aligned with ARL's primary roles. However, the task force believes ARL will need to build on existing relationships with other stakeholders and pursue new alliances to develop a more robust network in support of a broadly conceived e-science agenda.

Note: Priority actions are designated with a ♦

OUTCOME 1: **An ongoing capacity and process within ARL to develop, coordinate, and evaluate an e-science program agenda.**

STRATEGIES:

- Develop structure and processes for carrying out the ARL e-science program agenda, including a robust education and communication program.

- Coordinate and monitor progress on the outcomes, strategies, and action plans outlined below.

ACTIONS:

1.1 ♦Establish an ARL e-science working group with responsibility for program development (including education and communication programming), and associated resource development, and coordination across the three ARL strategic direction steering committees. The recently appointed program officer(s) provide critical focus, but will need such a group to coalesce interests and carry out a broader agenda. The working group's charge and accomplishments should be reviewed annually to ensure that it continues to address ARL's strategic priorities.

1.2 ♦Create a structure to ensure communication channels and coordination between the e-science agenda, the three ARL steering committees, the ARL membership, and external stakeholders.

1.3 Identify key library and technology organizations and scholarly societies/associations with whom to establish communication linkages and a potential structured liaison program.

OUTCOME 2: A widely shared understanding both within research libraries and among other stakeholders in the e-science support community of how libraries can contribute to the development and ongoing evolution of cyberinfrastructure and e-science.

STRATEGIES:

- Build understanding at the level of library leadership of the potential for e-science to transform the process and conduct of research.

- Develop, in collaboration with other stakeholders and experts, a set of principles for research libraries support of e-science. (See example, Appendix B.)

- Articulate, both within ARL and with key education and research societies, a vision following from these principles and from existing models and exemplars, for research libraries roles in stewardship of research assets and as a consultant/partner in the full life cycle of scientific data.

- Build understanding at the practitioner level in the library profession of e-science support practices and needs.

ACTIONS:

2.1 ♦The ARL e-science working group (identified in Action 1.1) will propose a process for developing, vetting, and sharing a set of principles for research library engagement in e-science. (See model in Appendix B.)

2.2 ♦Sponsor a program for library directors at an ARL membership meeting to discuss principles, with invited participants outside the library community (e.g., Science Commons, NSF Office of Cyberinfrastructure (OCI)).

2.3 ♦Develop talking points on key issues for use in communicating with campus stakeholders about library roles and engagement for e-science, e.g., dealing with data persistence, methods, etc. Develop separate talking points for different sectors, including:

- University librarians/library directors, to have language to use in talking with campus leadership
- Individual librarians, to use when talking with their disciplinary communities

2.4 Plan programs to explore e-science issues for the ARL membership and consider ways to disseminate these events to a broader audience, e.g., staff at research libraries and interested faculty and administrators. Program types may include:

- **Panel of Significant Players in E-Science Projects.** This might highlight case studies illustrating exemplary library roles (e.g., CERN, Cornell, Illinois, Johns Hopkins, Purdue, Queensland University of Technology, Woods Hole Oceanographic Institute) or discipline-specific projects. Part of the intent of this programming would be to identify and showcase the major federally funded research centers.

- **Workshop for Self-Selected Teams** with a particular motivation or an emergent e-science project. This could take the form of institutional teams or representatives from multi-institutional projects. Focus on exploring the roles and challenges of team science. Using knowledge gained through team workshops, design and deliver programs for library service practitioners with a focus on building understanding and strategies to engage with faculty to support their research agenda.

2.5 ♦Initiate conversations within the Association of American Universities, EDUCAUSE, the National Association of State Universities and Land-Grant Colleges (NASULGC) and the I-Schools Project to promote the role of research libraries as significant players in e-science. Seek to establish a formal liaison network with these organizations.

2.6 Develop a communications agenda related to data preservation directed to scientists and researchers similar to the successful Council on Library Resources' "Slow Fires" campaign for issues of preservation. This should involve other partners with preservation interests.

OUTCOME 3: Knowledgeable and skilled research library professionals with capacity to contribute to e-science and to shape new roles and models of service.

STRATEGIES:

- Highlight exemplary programs and lessons learned.
- Identify gaps in library services and recommend steps that research libraries can take to address support needs of team science, including inter-institutional team science.
- Build a library workforce with relevant new skills and knowledge about emergent forms of documentation and research dissemination.

ACTIONS:

3.1 ♦Establish a process within ARL to develop and sustain e-science education and communication resources to include:

  - **Glossary**—a glossary defining key e-science terms to foster a common understanding.
  - **Resource bibliography**—an annotated bibliography to identify and link to the major reports on e-science and cyberinfrastructure in various disciplines.
  - **Inventory**—an inventory of important discipline-based e-science centers and large-scale projects. Identify the major groups dealing with data issues (e.g.,

Committee on Data for Science and Technology/ CODATA and CENDI) at national and international levels.

- **Wiki**—an ARL-hosted wiki to gather and build a knowledgebase of resources surrounding e-science topics. Use to gather together information about relevant projects and to document emerging, innovative types of library staff positions. This tool will be used initially among working group and interested members to communicate findings. Eventually, interest and momentum may build to the extent that it may become a useful tool for member communication.

3.2 ♦Pursue programs to develop science librarian skills to meet the needs of e-science. Collaboration with IMLS may be one strategy.

OUTCOME 4: Research libraries as active participants in the conceptualization and development of research infrastructure, including systems and services to support the processes of research and the full life cycle of research assets.

STRATEGIES:

- Actively monitor, understand, and engage in activity around emergent models in publishing, particularly publication with associated primary research data.
- Monitor developments in research tools and systems, e.g., electronic laboratory notebook systems.
- Monitor and document development of collaboration environments (e.g., through requests for proposals) to identify logical points in which research librarians and research libraries might play a role.
- Document development of discipline-based repositories.
- Support new forms of scientific data publication.
- Support long-term access to scientific data as part of the scientific record.

ACTIONS:

4.1 ♦Work with relevant professional organizations and disciplinary societies to explore issues associated with new forms of publication "packages" and genres that include data. The National Academy may be relevant here.

4.2 ♦Partner with CNI to pursue potential "Friday Forum" or executive roundtable opportunities to explore research infrastructure associated with data and related applications.

4.3 Partner with CNI to conduct an analysis of the unmet infrastructure needs related to research collaboration that might be met by libraries.

OUTCOME 5: **Influence on policy, standards, and resource allocation decisions that support ARL principles.**

STRATEGIES:

- Promote research library involvement in shaping policy and protocols with respect to emerging scholarly communication models that integrate data and publications.

- Develop mechanisms to be an active participant in the open data movement.

ACTIONS:

5.1 ♦Inventory and document policies of government agencies, foundations, and other organizations funding e-science projects.

5.2 ♦Develop an education and communication program for ARL libraries to assist university officials with new research council regulations about data deposit and access and to support compliance officials on local campuses regarding these policies.

5.3 Identify and share models of library support to assist scientists in complying with data-management policies of relevant funding agencies (e.g., the "concierge" model practiced at University of California, San Francisco).

5.4 Align the policy apparatus and resources of both ARL and SPARC to work in concert in support of open data principles and policies.

Appendix A: Data Curation[4]

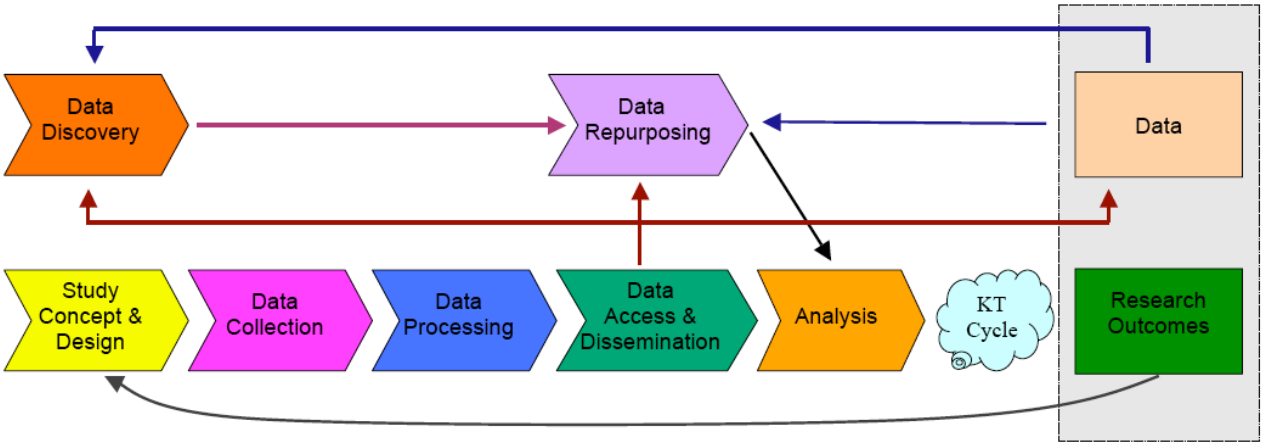
Stewardship of digital resources involves both preservation and curation. What is the distinction between these terms? Are these concepts describing the same thing? If not, how do curation and preservation differ? Both concepts have been borrowed from other fields and applied to the realm of research data. Curation has its roots in museum management, while preservation traces its origins to archivists. The Digital Curation Centre (DCC) in the UK defines digital curation as "maintaining and adding value to a trusted body of digital information."[5] DCC documents frequently make reference to "curation and preservation." That is, they treat these concepts as functionally different.

What are the functions that one would attribute to curation that differ from preservation? Preservation consists of (a) the management practices based on standards that guide and build metadata and data throughout the research life cycle and of (b) the subsequent long-term care for these digital products. The outputs of (a) are copies of the metadata and data in discipline-acknowledged standards best suited for (b), their long-term care, access, migration and refreshment.

Curation involves ways of organizing, displaying, and repurposing preserved data collections. Along the lines of the DCC definition, curation functions add value to a collection of preserved data by organizing and displaying the data through analyses of the collection's metadata or through the creation of new data from the preserved collection.

From the perspective of the life cycle of research data, preservation occurs through the stages of data production and the creation of research outputs represented on the bottom row. Long-term preservation in this model consists of the practices followed in caring for the data, which is represented by the box on the right side of the figure. Data curation is characterized on the top row by the stages of data discovery and data repurposing, which make use of the preserved data. The activities of these two functions bring new value to the collection through analyses of the metadata, which display aspects of the collection in new light, and the creation of new data from the existing data collection.

[4] Prepared by Chuck Humphrey for *To Stand the Test of Time: Long–Term Stewardship of Digital Data Sets in Science and Engineering* (Washington, DC: ARL, 2006 (http://www.arl.org/bm~doc/digdatarpt.pdf).
[5] See "What is Digital Curation" on the Digital Curation Centre Web site at http://www.dcc.ac.uk/about/.



The "KT Cycle" in the diagram represents the processes of knowledge transfer. This life cycle diagram comes from Charles Humphrey, "E-Science and the Life Cycle of Research" (2006) available online at http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc.

Appendix B: Model Principles

Model Principles for Research Library Roles in E-Science*
draft by Chuck Humphrey, with edits from task force members

Research libraries, as partners in the knowledge creation process, have significant roles to play in e-science, including the areas of cyber-infrastructure, data, metadata, and tools underlying this newly emerging research methodology.

Research libraries are both physical and virtual places with rich resources, including human capital, and services that support many of the stages of the life cycle of e-science research.

Principles:

1. ***Open Access:*** Research libraries will support open access policies and practices regarding scientific knowledge and e-science. Barriers will be removed that impede or prevent open access to research outputs, and consequently that restrict the potential linkage of outputs to the data upon which research findings are based.

2. ***Open Data***: Access to open data is a movement supported by research libraries, taking into consideration the ethical treatment of human-subject data.

3. ***Collaboration:*** Research libraries will collaborate with multi-institutional, interdisciplinary research projects by developing and supporting digital repositories for their research outputs, data, and metadata.

4. ***Digital Stewardship & Preservation:*** Research libraries will have institutional repositories that meet international preservation and interoperability standards and practices.

5. ***Equitable Service and Support:*** Research libraries will work collectively to ensure that gaps do not develop in the levels of support provided across e-sciences. Frequent environmental scans will be conducted to assess the comprehensive coverage of library support for e-sciences.

6. ***Professional Development & Investment:*** Research libraries will develop the human capital to provide the range of knowledge management skills at the appropriate level needed by e-sciences. This includes professional development initiatives through library organizations, including ARL, and new professional positions dedicated to the support of e-science resources and services, such as employing data scientists/librarians/archivists. Research libraries will employ a mix of professional graduates in filling these new positions, including MLIS graduates, but not exclusively.

7. ***Metadata Standards & Metadata Creation:*** Research libraries will spearhead initiatives to develop metadata standards supportive of scientific data. Research libraries, working closely with the different e-science communities, will take the lead in creating metadata essential to the operation of e-science cyber-infrastructure.

9. ***Virtual Communities:*** Research libraries will contribute to the establishment of and participate in virtual laboratories or organizations developed across e-sciences.

10. ***Sustainable Models:*** Research libraries will participate in the development of and contribute to sustainable business models for the resources and services essential to e-sciences.

11. ***Communication:*** Research libraries will participate in initiatives to increase wider professional and public understanding of e-science contributions to knowledge and its infrastructural requirements.

* The June 2008 OECD Ministerial Meeting on the Future of the Internet Economy will be considering a set of digital information principles, indicating a broad international interest in developing principles for cyberinformation. (See http://www.oecd.org/department/0,3355,en_2649_34223_1_1_1_1_1,00.html.)

Appendix C: Readings and References

Gold, Anna. 2007. Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine*, 13 (9/10), September/October. http://www.dlib.org/dlib/september07/gold/09gold-pt1.html.

Gold, Anna. 2007. Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries. *D-Lib Magazine*, 13 (9/10), September/October. http://www.dlib.org/dlib/september07/gold/09gold-pt2.html.

Hey, Tony and Jessie Hey. 2006. E-science and Its Implications for the Library Community. *Library Hi Tech*, 24 (4), 515-528.

Hunter, Jane. 2006. Scientific Publication Packages—A Selective Approach to the Communication and Archival of Scientific Output. *The International Journal of Digital Curation*, 1 (1). http://www.ijdc.net/./ijdc/article/view/8/7.

Lynch, Clifford. 2005. Where Do We Go from Here? The Next Decade for Digital Libraries. *D-Lib Magazine*, 11 (7/8), July/August. http://www.dlib.org/dlib/july05/lynch/07lynch.html.

Lynch, Clifford. 2007. The Shape of the Scientific Article in the Developing Cyberinfrastructure. *CT Watch Quarterly*, 3 (3), 5-10. http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/.

National Consultation on Access to Scientific Research Data. Final report, January 31, 2005. http://ncasrd-cnadrs.scitech.gc.ca/NCASRDReport_e.pdf.

National Institutes for Health. Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS). Released August 28, 2007. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.

National Institutes for Health. NIH Roadmap for Medical Research. Last reviewed September 18, 2007. http://nihroadmap.nih.gov/.

National Science Foundation, Blue Ribbon Advisory Panel on Cyberinfrastructure. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure*. http://www.communitytechnology.org/nsf_ci_report/
Arlington, VA, May 1-2, 2006.

National Science Foundation, Cyberinfrastructure Council. 2007. "Cyberinfrastructure Vision for 21st Century Discovery." http://www.nsf.gov/od/oci/CI_Vision_March07.pdf.

National e-Science Centre, Office of Science and Innovation e-Infrastructure Working Group. *Developing the UK's e-Infrastructure for Science and Innovation.* 2007 http://www.nesc.ac.uk/documents/OSI/report.pdf.

Rhoten, Diana. 2007. The Dawn of Networked Science. *The Chronicle of Higher Education*. 54(2), B12. http://chronicle.com/weekly/v54/i02/02b01201.htm.

Research Information Network and Consortium of Research Libraries [UK]. 2007. *Researchers' Use of Academic Libraries and their Services*. http://www.rin.ac.uk/researchers-use-libraries

Towards 2020 Science. 2006. Microsoft Research. http://research.microsoft.com/towards2020science/downloads.htm

Appendix D: About the Joint Task Force

Charge for the ARL Joint Task Force on Library Support for E-Science
Prepared Jointly by the Research, Teaching, and Learning Steering Committee and the Scholarly Communication Steering Committee

Background:
The UK National E-Science Centre defines e-science as "the 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 visualization back to the individual user scientists." E-science is defined broadly to include all of the natural and physical sciences, related applied and technological disciplines, as well as biomedicine and social sciences sharing research approaches with the natural sciences.

Charge:
The task force was jointly organized by the Research, Teaching, and Learning Steering Committee and the Scholarly Communication Steering Committee to recommend and initiate strategies to address emerging issues in the development of e-science for ARL libraries.

The work of the task force included:

- Informing the membership about e-science issues
- Identifying opportunities and recommending strategies for developing relationships with various government scientific agencies and other key stakeholders such as scientific societies
- Building an understanding of the needs and experiences of scientists and researchers in various disciplines who are using electronic collections of scientific data
- Recommending approaches to addressing issues related to the curation of long-lived digital data, including the handling of simulations and storage of massive data sets
- Engaging ARL members in the development of new roles for libraries as e-science infrastructure and service needs emerge at research institutions and promoting the contributions of research libraries in this arena
- Identifying the skills needed as information professionals move into the emerging e-science landscape and encouraging the development of information professionals prepared to assume new roles

The task force monitored closely activities at the agencies and organizations that fund science research. In addition it sought points of intersection with key initiatives such as Science Commons.

The task force consulted freely with ARL steering committees. Recommendations are to be made to the ARL Board (copying the steering committees).

Timeframe:
The task force made an interim report in October 2006 to the full ARL membership, provided updates to the steering committees in May 2007, and resented a draft final report to all three steering committees in October 2007.

Task Force Process
An eight-member task force, co-chaired by Wendy Lougee (Minnesota) and Bernard Dumouchel (CISTI), was formed in 2006. The task force brought together directors and senior administrators from ARL member libraries with expertise in library support for research in the sciences:

- Bernard Dumouchel, Co-Chair (CISTI)
- Wendy Lougee, Co-Chair (Minnesota)
- Sayeed Choudhury (Johns Hopkins)
- Chuck Humphrey (Alberta)
- Betsy Humphreys (NLM)
- Clifford Lynch (CNI)
- James Mullins (Purdue)
- Sarah Pritchard (Northwestern)
- Julia Blixrud (ARL staff liaison)

The task force met face-to-face at the fall 2006 Membership Meeting and again at the spring 2007 Membership Meeting, supplemented by conference calls and e-mail exchanges. A brief survey of ARL directors was conducted, soliciting information about innovative library programs and services in support of e-science projects, but little input emerged from this investigation.

Mr. Dumouchel retired from service effective in January 2007 and Ms. Lougee assumed the role of Chair. Three additional members were added in early 2007:

- Anna Gold (MIT)
- Richard Luce (Emory)
- Peter Young (NAL)

A Visiting Program Officer for Library Support for Research and E-Science, Neil Rambo (Washington), was appointed in March 2007, to provide additional support to the task force and to advance ARL's role in e-science and research.

Sub-topics emerged during task force deliberations and small-group conference calls were conducted in advance of the spring 2007 meeting to explore these issues more fully for the benefit of the full task force. The desired outcomes, strategic themes, and recommended first steps were developed during the spring 2007 meeting.

Wendy Lougee, Neil Rambo, and Julia Blixrud collaborated in the drafting of the final report.