

D-Lib Magazine
September/October 2007

Volume 13 Number 9/10

ISSN 1082-9873

Cyberinfrastructure, Data, and Libraries, Part 1

A Cyberinfrastructure Primer for Librarians

[Anna Gold](#)

Massachusetts Institute of Technology

<annagold@MIT.EDU>

Introduction

E-Science, cyberinfrastructure – these ideas are at the heart of the great ambitions and promise of science in the new century. The last several decades of network- and computer-enabled work in science have produced untold amounts of data, leading to the challenge of developing practices to manage and provide access to this data. Along with oceans of data and technology, changes in the conduct and nature of science – notably new collaborative and computational science practices – present both novel requirements and exciting opportunities to succeed in meeting this challenge.¹ A global effort is emerging to take collective responsibility for a growing yet still vulnerable investment in scientific data as a permanent part of scientific research communications and practices.

Today, everyone with a role in the traditional infrastructure of scientific research and communication is jockeying for a role in the emerging landscape of scientific data: national libraries; research funding agencies; universities and research libraries; and giants of the software and publishing industries. As roles and responsibilities get sorted out, librarians are testing the waters to identify what present and future roles they may have in these developments; but these are early days, and it is still unclear what those roles may be.

For information professionals, getting to know the concepts and issues raised in the name of cyberinfrastructure and E-Science is no mean feat. There has been an outpouring of lengthy reports addressing and framing their challenges, and the result is a ferment of ideas, vocabularies, organizations and strategies. Becoming literate in cyberinfrastructure means understanding cyberinfrastructure, E-Science, collaboratories, collaboration science, computational and grid science, data curation, the Semantic Web, open data, data archiving, digital preservation, and data management, and how they relate to each other.

While this learning curve is one factor, a second challenge adds to the difficulty that librarians have engaging in these issues: data science and data management are an awkward fit with the text-oriented constructs and systems that still dominate library relationships with science communication and publishing. Most librarians are much less familiar with the data-generating research phases of the scientific research cycle than with post-research phases of reporting, communication and publication.

Finally, add to these challenges the technical background needed to understand, even at a general level, the technologies that may be used in data management. Language itself becomes a barrier to understanding. When

used in discussions of E-Science, a vocabulary familiar to librarians (archival, curation, stewardship, provenance) takes on new or specialized meaning.

Despite these challenges, the library profession has been eager to grapple with the new issues presented by scientific data as a part of the scholarly record. Major professional organizations, among them the American Society for Information Science & Technology (ASIS&T), the American Library Association (ALA), and the Association of Research Libraries (ARL), are sponsoring or leading activities with a focus on cyberinfrastructure and scientific data. A major milestone in this growing interest is the initiative launched in 2006 by ARL, a Task Force on E-Science.² The purposes of this ongoing E-Science initiative include exploring and addressing gaps in awareness and understanding within the profession; as well as positioning research libraries alongside research funding agencies and research universities and organizations, as contributors and partners in dealing with the immense challenges of managing and preserving access to digital scientific data.

Shortly after the Task Force was established, ARL and the National Science Foundation (NSF) co-sponsored an invitational workshop in October 2006 on the topic of digital data stewardship. An executive summary of the workshop was published³ reaching a broad audience of library leaders and professionals. Yet even this brief and relatively accessible article, focused on the roles of libraries and librarians in E-Science, presumes a fairly high level of general understanding of the relevant issues and vocabulary as well as familiarity with a host of stakeholders engaged in the field of digital data management.⁴

The following two-part article is offered to help open up the discussion with library practitioners working directly with research faculty and graduate students, advising on issues of scholarly communication, and concerned with providing relevant data services in the context of relatively well-established library-based data support programs in GIS, social science data, and bioinformatics. Part 1 provides a primer for librarians on cyberinfrastructure, including an overview of major issues and readings to help locate the issues in the larger national and global framework, as well as an introduction to emerging critiques of global cyberinfrastructure theory. [Part 2](#) offers an overview and analysis of current theories about the roles libraries and librarians can have associated with the multiple dimensions of cyberinfrastructure.

1. A cyberinfrastructure primer for librarians

At a research university, a professor collects data on protein crystal measurements. This data is very valuable because getting protein to crystallize is difficult: it may take a week, or it may take fifteen years. Repeating experiments is non-trivial. The resulting raw data files are huge: they include 360 images – one from each angle. Data formats change as instrumentation changes – constantly. The professor saves the raw data in the latest file format, on the latest storage media – currently CD-ROM. She provides detailed instructions to graduating students on how to pass the data back to her, but there is no central storage service for this kind of data product. Once the data is captured, the raw data is processed (analyzed) into a form that is widely used by researchers (although this form, too, changes over time). Submitting data to an international data bank requires one such conversion. But once submitted, there is no easy way to reverse engineer that format back to the working format that makes it usable for other researchers. Also, it becomes important to publish data quickly because otherwise researchers will not be able to use the original data as different detectors do not talk each other's language, and these programs are constantly changing.

The sketch above illustrates several of the problems facing scientists in data-rich, data-driven science. Instruments and arrays yield vast amounts of data, and increasingly the data itself becomes the subject of research and re-use through computational manipulations. In many fields these activities take place in local settings that are subject to rapid and continual changes in instrumentation and digital media standards. Support is lacking for long-term storage, for retrieval of usable formats and derivatives, and for sharing data in forms suitable for reuse or for use in different research settings. Where standards and practices for publishing research

results in textual form are well-established and supported, managing access to data has the feel of a vast frontier with pockets of homesteaders and small settlements, and a few well-supplied and well-guarded nodes. And yet, data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself.

1.1 Setting the stage: supercomputers, informatics and grid science, and the rise of cyberinfrastructure

By the late 1990's, the use of computers to handle manipulations of very large data sets was well-established. National computational science infrastructure took three primary forms: local high performance computing (HPC) centers – located at most research universities; supercomputing capabilities located in national laboratories; and a small number of distributed supercomputing nodes that provided supercomputing capabilities on demand to scientists around the country. At about the same time, the rise of genomic science offered a template for a new approach to problems in life science, using collaborative efforts to gather pieces of data in standard and well-understood ways, thereby building up shared data resources – the genomes of several organisms – culminating in the sequencing of the human genome in April 2003.⁵

At the end of the last century this inter-networked and distributed infrastructure began to evolve towards what became known as "grid computing." The grid offered a new way of accessing centralized supercomputing capabilities; at the same time it provided a more distributed computational architecture that reduced reliance on supercomputing centers.⁶ Grid computing – sharing computational resources across a decentralized and heterogeneous computing and institutional network – had its origins in the 1970's, but it became a major element of scientific computing infrastructure with the emergence of the Globus Toolkit in 2002. The Globus community includes such major computational science initiatives as BIRN (Biomedical Informatics Research Network, <http://www.nbirn.net>); the U.S. National Virtual Observatory (<http://www.us-vo.org/>); the Sloan Digital Sky Survey (<http://www.sdss.org/>); the Particle Physics Data Grid (<http://www.ppdg.net/>); and GEON, the Geosciences Network (<http://www.geongrid.org/>).

At about the same time, a new kind of data science initiative emerged that didn't require high performance or parallel computing, but helped provide an important model for data sharing: the PDB (Protein Data Bank, <http://www.wwpdb.org/>), a shared data bank of protein structures.

Gradually these and similar developments came to be referred to as "cyberinfrastructure" in the U.S., and elsewhere as "E-Science" – mostly in Europe and the U.K. A focus on creating an infrastructure of computing hardware, software, networks, and services began gradually to give way to a new interest in data reuse, followed by a concern for the long-term stewardship of the data being produced. Public research funding organizations were worried about their investment in data produced through publicly funded research, especially when it had long-term, authoritative and reference value within a research community.

Similar concerns were emerging among scientists and engineers who had begun to understand that data management over time and across heterogeneous data sources posed an increasingly complex and unaddressed research problem. Meanwhile, the growing dependence of all social, academic, and cultural sectors on digital data created a new awareness of the vulnerability of digital data, and the need to create both technical and social models to assure the persistence and integrity of important digital data over time.

1.2 Cyberinfrastructure and E-Science

After 2000, a number of workshops, reports, and related publications dealing with cyberinfrastructure began to appear from the NSF. Of these perhaps the best known is *Revolutionizing Science and Engineering Through Cyberinfrastructure* (also referred to as the Atkins report), published in January 2003.⁷ As defined by the authors of the Atkins report, "cyberinfrastructure" refers to a comprehensive and integrated system of hardware,

networks, software, and middleware, designed to support a range of advanced data acquisition, storage, management, integration, mining, and visualization over the Internet:

"The term...cyberinfrastructure refers to infrastructure based upon distributed computer, information, and communication technology. If *infrastructure* is required for an *industrial* economy, then we could say that *cyberinfrastructure* is required for a *knowledge* economy."⁸

An important outcome of the Atkins report was the creation by the NSF of a new Office for Cyberinfrastructure (OCI). Since its establishment, the OCI has already issued a number of calls to action, vision, and other planning documents. The Atkins report also marked recognition of a broader change in the role of computing in science at all scales. For one thing, the emerging cyberinfrastructure was clearly something with human and social dimensions as well as technical ones. Both the Atkins report and another draft NSF report that followed in early 2006, *Cyberinfrastructure Vision for 21st Century Discovery*,⁹ called for steps to develop the workforce that would be needed to create and sustain cyberinfrastructure. This workforce would include data scientists, a new type of professional with both domain science expertise and high-level computational skills; and also social scientists trained to understand the human factors relevant to the use of technology-mediated collaborative tools.

The broadening implications of cyberinfrastructure were echoed in a publication by Microsoft Research issued later the same year, *Towards 2020 Science*.¹⁰ Written by a panel of scientists, this report argued that computer science concepts, tools, and theorems have the potential for having "a profound impact on science. It is a leap from the application of computing to support scientists to 'do' science (i.e., 'computational science') to the integration of computer science...into the very fabric of science" (p. 8). The Microsoft report notes that the new paradigm will bring with it the challenge of "end-to-end scientific data management, from data acquisition and data integration, to data treatment, provenance and persistence." The Microsoft report claims that a new paradigm will bring with it the challenge of "end-to-end scientific data management, from data acquisition and data integration, to data treatment, provenance and persistence." It also predicts a paradigm shift in scientific publishing, from the "traditional sequence of 'experiment > analysis > publication'" to "'experiment > data organisation > analysis > publication'" (p. 16).

1.3 Data archiving and preservation

While these ideas about cyberinfrastructure were forming, a different, if overlapping array of reports began to chart another challenge of digitally based research: the problem of preserving digital content. Given the staggering output of digital data and digital records, the problem had several daunting aspects. On the one hand, there was the problem of what "preservation" means in the digital realm. It seemed clear already that in the absence of practices for refreshing and migrating both data and relevant retrieval software, digital records of all kinds were threatened with loss. Some began to refer to the threat of a "digital dark age." Early prophets of this threat had begun to warn of this a decade earlier.¹¹ While keeping digital objects usable was one concern of the digital preservationists, there were others, including capture (archiving) and selection, as well as providing sufficient metadata or other descriptive or administrative information to ensure adequate (as well as legal) access and retrieval over the long term.¹²

Several years ago, major government organizations worldwide began to outline the scale of the digital archiving problem, while calling for the development of needed technical, social, and business models. Internationally, these included the National Library of Australia in the 1990's;¹³ UNESCO, with its 2003 *Charter on the Preservation of Digital Heritage*;¹⁴ JISC in the UK, with its 2003 *E-Science Curation Report*;¹⁵ and, in Canada, the National Consultation on Access to Scientific Research Data 2005 report.¹⁶

In the US, the Library of Congress and the National Archives both began to convene stakeholders and develop

research programs to create "national infrastructure for long-term preservation of digital information."¹⁷ In 2004 the Library of Congress began to fund research through its DIGARCH program, in partnership with the NSF, as part of its new National Digital Information Preservation Program (NDIPP). After pursuing related research beginning in 1998, in 2005 the National Archives (NARA) awarded a major contract to Lockheed Martin to build a permanent archive of federal electronic records.¹⁸ With these and other partner organizations, the NSF continues to play a key role in shaping U.S. discourse and strategies for digital preservation. The National Science Board (NSB) issued an important report in 2005 on *Long-Lived Data Collections*,¹⁹ and in late 2006 the NSF funded a workshop with the Association of Research Libraries, documented in its report, *To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering*.²⁰ For libraries, this report was significant for the explicit case it made for the role of U.S. research libraries in future cyberinfrastructure.

As diverse as these reports have been in their focus and provenance, common threads can be found across all of them. With varying emphasis, each has called for more education on the issues; for collaborative action by cross-sector partnerships of academe, government, industry, and others; and for enabling public policies (e.g., requirements to provide data management plans as a condition of receiving government grants), as well as research funding to support, encourage, and improve preservation tools, standards, and practices: in short, to fund infrastructure.

1.4 Curation, access, and interoperability

If preservation in the narrowest sense means keeping the bits safe, and safely reusable, the promise as well as the challenge of digital *curation* is a more encompassing concept that embraces preservation as one element of digital archiving. Curation, according to the UK-based Digital Curation Centre, means "maintaining and adding value to a trusted body of digital information for current and future use; specifically....the active management and appraisal of data over the life-cycle of scholarly and scientific materials."²¹

A major effort in the arena of curation is to develop metadata practices and standards that will do several important things: first, they must make data understandable by computers; second, they must support discovery across heterogeneous data collections; and third, they must manage all of this across data scales, from the small to the immense.

One of the keys to making all this possible, many believe, is expressing data using a data model called RDF (Resource Description Format). The promise of RDF is to encode data in a standard that is highly flexible and yields significant economies of scale. Adopting RDF in turn supports the development of the Semantic Web, an effort incubated within the W3C (World Wide Web Consortium).²²

The promise of the RDF standard in combination with the Semantic Web is to make it possible for meaningful data in documents to be processed and extracted by computers. That in turn makes it easier to exchange data within and across research domains (e.g., between geology and chemistry). Indeed, the huge scale of data production makes it essential that metadata be generated and captured by automated means, even if human expertise is needed to develop cross-walks, annotations, or other descriptive information that will contribute to its value and usability. Ideally metadata will be captured as a by-product of the natural data production life cycle, e.g., at the point the data is actually created. But to do this means to understand, at a fairly finely-grained and socially nuanced level, what the actual life cycle of data production is like.

1.5 Services and the data life cycle

Librarians have long been familiar with canonical representations of the life cycle of scientific activities that result in publications, as well as (separate) representations of the various stages of scholarly communication and publishing. Given these two somewhat stable frameworks, it was relatively easy to describe the flow from

experiment to results, with the outcome being the generation of primary and eventually secondary literature such as indexes, abstracts, handbooks, encyclopedias, and textbooks.

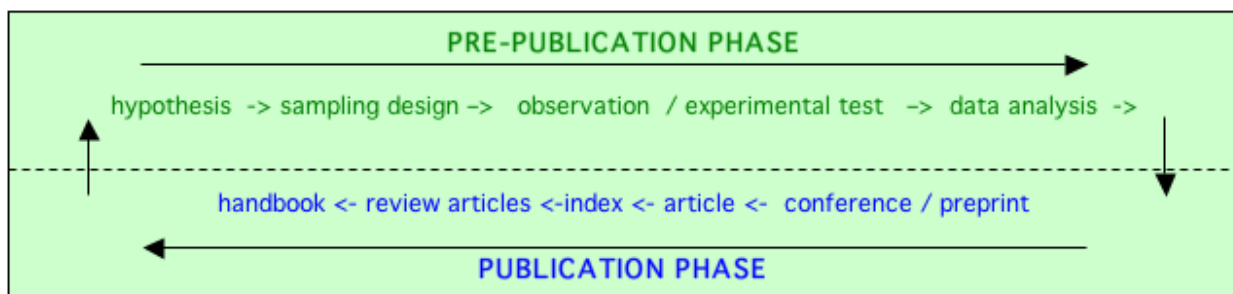


Figure 1: Data and publication life cycles

While some library services, including GIS and social science data services, operate in the top half of this framework in the pre-publication phase, for the most part libraries have assembled, managed, and served artifacts of various genres mostly from the "published" or bottom half of this framework.

Scientific data exists mostly in the pre-publication phase, and only a small sub-set of data is transferred across the divide into the publication phase. Here data – usually in highly reduced form – appears as tables or graphs embedded within a textual report of the scientific work. The demand for systems to capture, steward, and reuse scientific data clearly violates this old demarcation. Librarians in particular have not traditionally been involved in the production of scientific information prior to the publication of results. Yet if data produced through the conduct of science is to become a target of derivative works (for data mining, discovery and other reuse), then standards such as RDF, and systems and ontologies developed as part of the Semantic Web and other annotation systems, will be necessary.

To develop these systems will require a more complete and nuanced understanding of the data life cycle. One well-developed model for understanding this cycle is the "life cycle model" developed by Hunter,²³ which follows the workflow of data from conception to final product. Several strengths of this model are that it allows for the possibility that a data set will grow; it supports the ability to track the evolution of data over time; and it includes the data source (provenance) as an integral aspect of the data.²⁴

To some extent, generalizations about data production in science can be useful. As can be seen in the story at the beginning of this article, data production in science may be characterized in several phases and flavors. To begin with there is "raw data" – data as it comes directly from the scientific instrument. There may be stages of validating and selecting the data. Then the data may be subjected to any number of standard or ad hoc processes that calibrate it – that translate it into a more general schema (e.g., sky maps, catalogs). Another way of abstracting the nature of data is to contrast particular knowledge with "properties in general" – data abstracted or generalized to a level of understanding.

The 2005 National Science Board (NSB) report entitled *Long-Lived Digital Data Collections* suggests that data can be differentiated based on its nature, e.g., numbers, images, video or audio streams, software and software versioning information, algorithms, equations, animations, or models/simulations; and also by their origins: whether they are observational, computational, or experimental. These latter distinctions are crucial when it comes to making choices for archiving and preservation. Observational data (e.g., observations of ocean temperature on a specific date) are essentially historical and cannot be reproduced, and so may be primary candidates for indefinite archiving. Computational data may require archiving complete information about the computer model and the execution (e.g., hardware, software, input data), but not of the data results themselves – which can in theory be reproduced. Experimental data may not be easily reproduced, given both cost considerations and the complexity of all the experimental variables. These factors (cost and reproducibility) will be relevant to preservation policies for experimental data. If data is processed for a variety of purposes (as

in our illustration of the chemist's work), a variety of derivative products may also merit preservation.

In the life sciences, scientists have been encouraged to share their data with others – yet no one expects them to share all their data, in all its pre-processed and post-processed forms. The National Institutes of Health (NIH) has developed data sharing guidelines²⁵ that offer some important distinctions between different categories of scientific data in relation to whether it is important that they be shared. The NIH requires that their grantees share "final research data," consisting of records "necessary to document and support research findings." This category therefore falls in between the summary tables and graphs normally found in research articles, and the preliminary records such as laboratory notebooks, partial dataset, and other artifacts of the research process. The NSB, following the NIH, has also distinguished between data collected by individual researchers ("research collections"), or in the context of collaborative research ("resource collections"), or research that produces a shared, curated resource ("reference collections").

While the data life cycle is essential to understanding the needs and services for data curation and stewardship, the fact that most data is digital adds another dimension to the consideration of life cycle: the digital life cycle. Digital curation – the activities that assure that digital resources are managed and preserved for discovery, access, and reuse – is a concern that overlaps widely with data curation, since it is essential to understand both the data life cycle and the digital life cycle if the investment and opportunities represented by digital scientific data are not to be lost.²⁶

In short, models of data life cycles suggest the need for a wide range of supportive services, including repository services (for data that is stable and where sharing is a goal); a variety of metadata production services appropriate to the anticipated use and reuse of the data within and across the originating and interested research domains; discovery and data-mining services that provide essential linkages and pattern finding tools; and preservation services that assure long-term access for reuse.

1.6 Policy tools

Not all services and tools supporting cyberinfrastructure are technological. The social nature of science and the network of interested stakeholders in the future of access to scientific data make it essential to develop social and policy tools to support this future. One of the most significant policy tools to be developed to aid in the effort to ensure access to digital scientific data is the attachment of data management and data sharing requirements to government research funding.

Since 2003 the NIH has required applicants seeking \$500,000 or more in direct costs to include a plan, in their application, for sharing their final research data.²⁷ In *Long Lived Data Collections*, the NSB recommended requiring that applicants for NSF funds include a data management plan that would be subject to peer review.²⁸ In addition, specific NSF units, such as the Division of Ocean Sciences, and the NSF Arctic System Science program have for years had their own published policies on data sharing and management. In addition, some research communities have addressed this issue, creating data policies and guidelines.²⁹

The establishment in 2007 by the National Science and Technology Council of an Interagency Working Group (IWG) on Digital Data promises to bring about a great increase in sharing data policies and practices, as well as aiming at creating an "open interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering".³⁰

This is very good progress, because in the absence of such requirements, as we have learned at least in the social sciences, very little data survives over even short periods of time. The primary exception is data that has been deposited in well-understood community-based data repositories.³¹

Another policy need is for standards and practices for *citing* data. Standard citation practices in scientific

communication have developed for every conceivable type of textual publication, and these, along with web-based linking standards, are what make immensely popular and effective services such as SFX possible. Similar standards do not exist for most scientific data, although there are a few exceptions, such as the data publishing practices of the American Geophysical Union³² and the well-established social-science data archiving practices of the ICPSR.³³ In the UK, the eBank project (funded by JISC and led in part by UKOLN, the UK Office for Library and Information Networking), has been working on the problem of creating links between publications and datasets referred to in those publications, using citation and citation schemes (URIs) such as DOIs, PURLs, and handles.³⁴

A different approach is being adopted in Germany, where the German Science Foundation has been sponsoring a general-purpose data repository and registration system since 2005, as well as a project on the "Publication and Citation of Scientific Primary Data (STD-DOI)".³⁵ A goal of this system is to make data citable as publications using DOIs assigned by one of several data publishing agencies, while the German National Library of Science and Technology (TIB) acts as a DOI registration agency for primary scientific information.

Finally, another aspect of social policy regarding data is the development of practices and mechanisms to provide appropriate access control mechanisms. While data sharing and open data practices may serve the public interest, there are cases either where privacy requires some access limitation, or where there are intellectual property restrictions on its sharing. Funding agencies address these issues in relationship to data sharing requirements. For example, the NIH notes that applicants are expected to address if, or how, they will exercise intellectual property rights while making data available to a broader research community.³⁶

1.7 Business models

A growing concern of research funding agencies is finding workable business models for supporting the massive and evidently complicated new line of business that digital scientific data management represents. The business model for distributing, curating, and preserving textual representations of scientific findings, including published articles, papers, technical reports, and the like – is well understood and reasonably stable. While far from perfect, that system comprises an ecology of institutional players, from societies and commercial publishers to national and university libraries and archives. These partners maintain a system that can be reasonably counted on to sustain usable access to the published record of science over the long term.

The same can't be said, yet, for digital data. As can be seen from the review of issues above, national funding agencies and national libraries have played a very large role in framing the problem and bringing parties together to create the social and economic infrastructure that is as clearly needed as the technical one. But the problem is immense and, like the data itself, heterogeneous and distributed across a system at many scales. It seems only logical that a business model for data will likewise be both distributed and federated, a new ecology of institutional arrangements. Those will almost certainly include research libraries, whose funding structure is tied to even longer-lived institutions – universities – yet who have collaborated well and diversified their revenue streams to include private and grant funding. At the same time, research libraries have urged the NSF to involve economic and social science experts in developing models for sustainable digital data stewardship.³⁷ A danger is that, given what some refer to as the "Google effect" (the expectation that good infrastructure can be free), there will be a reluctance to devote scarce public or research dollars to create and sustain programs of long-term data stewardship.³⁸ There is therefore an urgent need to articulate a strong case for investing in data commons controlled and shaped by research communities, not by commercial interests.

1.8 A rebalancing act: local and global cyberinfrastructure

Accounts of cyberinfrastructure that prevail today are strongly influenced by their technological roots in high performance computing, and tend to idealize distintermediated solutions for optimizing data flow and transfer to general end-users. These solutions express assumptions that knowledge is a direct result of more and better

access to data; many of these solutions will naturally be centered on large-scale data centers, not libraries. However these solutions tend to emphasize global, large-scale data-taking and data-management enterprises, and these neither exhaust nor address the full extent of science data production and use in knowledge-making. Complementing these solutions are others that focus on local knowledge-making and local data centers:

"Paying attention to the growth of local knowledge-making provinces is a strategy for changing how we think about generalizations and network federation. Considering local and global arenas as distinct yet interdependent, allows a local data center to be transformed, to be seen as a new-age information environment where design and communication efforts are understood as knowledge work... A key information age challenge is creating a modular system of decentralized, heterogeneous information environments that function as learning arenas across the digital landscape."³⁹

Librarians interested in what their roles will be in the future cyberinfrastructures of science may find their skills and capabilities give them a natural and even crucial role in building and supporting the information infrastructures of local data centers – a role that draws on and adapts the mediating practices of the library profession to the world of data. In the second part of this article, the roles of libraries and librarians in cyberinfrastructures in light of current developments are explored further.

Notes and References

1. "2020 Future of computing," *Nature*, March 23, 2006, <<http://www.nature.com/nature/focus/futurecomputing/index.html>>. In this article I use the terms E-Science and cyberinfrastructure more or less interchangeably. Ann Zimmerman suggests that a more exact equivalent of "cyberinfrastructure" would be "eScience applications" (Ann Zimmerman, "Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse," *International Journal on Digital Libraries*, May 1, 2007 <<http://www.springerlink.com/content/p42u8177421u1477/fulltext.pdf>>. Accessed 9/12/07). Zimmerman acknowledges Clifford Lynch for pointing out that the UK usage of the term "eScience" encompasses not only infrastructure, but also related changes in scientific practice. See, Clifford Lynch, "Research libraries engage the digital world: A US-UK comparative examination of recent history and future prospects," *Ariadne*, issue 46, February 2006 <<http://www.ariadne.ac.uk/issue46/lynch/>> Accessed 9/12/07. My thanks to Karen Baker for pointing this out. Current usage of the two terms overlaps broadly, and may have evolved from the definitions for E-Science and cyberinfrastructure as articulated by John Taylor and Dan Atkins, respectively. See "Defining eScience" <<http://www.nesc.ac.uk/nesc/define.html>> for the former (Accessed 9/12/07), and *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, <<http://www.nsf.gov/od/oci/reports/atkins.pdf>> for the latter (Accessed 9/12/07). Still another variant, used by Peter Murray-Rust, is "cyberscience." <<http://wwwmm.ch.cam.ac.uk/blogs/murrayrust/?cat=16>> Accessed 9/12/07.
2. *ARL Joint Task Force on Library Support for E-Science: Charge & Roster*, <<http://www.arl.org/rtl/escience/escicharge.shtml>>. Accessed 9/12/07.
3. *ARL Bimonthly Report* No. 249, December 2006, <<http://www.arl.org/bm~doc/arlbr249datasteward.pdf>>. Accessed 9/12/07.
4. Another important document on the subject was published by the UK Library community (JISC) in 2007 – a "snapshot" on rights, roles and responsibilities of various stakeholders regarding scientific data. This 65-page report, by Liz Lyon, *Dealing with Data*, offers useful insights particularly into the UK perspective on these issues. <http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf>. Accessed 9/12/07.
5. *The Human Genome Project Completion: Frequently Asked Questions*, National Human Genome Research

Institute (NIH), <<http://www.genome.gov/11006943>>. Accessed 9/12/07.

6. See *GridCafe*, <<http://gridcafe.web.cern.ch/gridcafe/>> for an amusing and accessible introduction. Accessed 9/12/07.

7. *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, <<http://www.nsf.gov/od/oci/reports/atkins.pdf>>. Accessed 9/12/07.

8. *Ibid*, p. 5.

9. The final version of this report was issued in March 2007. See <<http://www.nsf.gov/pubs/2007/nsf0728/index.jsp>>. Accessed 9/12/07.

10. *Towards 2020 Science*, 2020 Science Group (Microsoft Research). <http://research.microsoft.com/towards2020science/background_overview.htm>. Accessed 9/12/07.

11. Brand, "Escaping the digital dark age," *Library Journal* v. 124. n. 2, p46-49, <<http://www.rense.com/general38/escap.htm>>. Accessed 9/12/07; see also Danny Hillis in MacLean and Davis, *Time and Bits: Managing Digital Continuity*, Getty, 1998; and Steward Brand, *The Clock of the Long Now: Time and Responsibility*, 1999.

12. Whether digital preservation is as expensive or difficult as is widely believed, is questioned by Chris Rusbridge of the UK Digital Curation Center in a 2006 article published in *Ariadne*, "Excuse me....Some digital preservation fallacies," <<http://www.ariadne.ac.uk/issue46/rusbridge/>>. Accessed 9/12/07.

13. *Preserving Access to Digital Information*, National Library of Australia, <<http://www.nla.gov.au/padi/>>. Accessed 9/12/07.

14. *UNESCO Charter on the Preservation of the Digital Heritage*, <http://portal.unesco.org/ci/en/ev.php-URL_ID=13366&URL_DO=DO_TOPIC&URL_SECTION=201.html>. Accessed 9/12/07.

15. Philip Lord and Alison Macdonald, *E-Science Curation Report, JISC Committee for the Support of Research, 2003*, <http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf>. Accessed 9/12/07.

16. National Consultation on Access to Scientific Research Data (NCASRD), *Final Report*, <http://ncasrd-cnads.scitech.gc.ca/home_e.shtml>. Accessed 9/12/07.

17. *Workshop on Research Challenges in Digital Archiving: Towards a National Infrastructure for Long-Term Preservation of Digital Information*, April 12-13, 2002, <<http://www.si.umich.edu/digarch/>>. Accessed 9/12/07.

18. ERA, or Electronic Records Archive, <<http://www.archives.gov/era>>. Accessed 9/12/07.

19. National Science Board, *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*, <<http://www.nsf.gov/pubs/2005/nsb0540/>>. Accessed 9/12/07.

20. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe*, September 26-27, 2006, Arlington, VA, Association of Research Libraries, <<http://www.arl.org/bm~doc/digdatarpt.pdf>>. Accessed 9/12/07.

21. Digital Curation Center, *About the DCC*, <<http://www.dcc.ac.uk/about/>>. Accessed 9/12/07.

22. Eric Miller, "The Semantic Web,"

- <<http://www.w3.org/2002/Talks/www2002-w3ct-swintro-em/Overview-6.html>>. Accessed 9/12/07.
23. Jane Hunter, University of Queensland, 2006, "Scientific models – A user-oriented approach to the integration of scientific data and digital libraries."
<http://www.valaconf.org.au/vala2006/papers2006/55_Hunter_Final.pdf>. Accessed 9/12/07.
24. See also Figure 7, "e-Research Life Cycle and data curation," in Liz Lyon, 2007, *Dealing with Data*, op. cit.
25. NIH, Office of Extramural Research, *NIH Data Sharing Policy and Implementation Guidance*, <http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm>. Accessed 9/12/07.
26. "Conceptualizing the digital life cycle," *IASSIST Communique*, <<http://iassistblog.org/?p=26>>. Accessed 9/12/07.
27. NIH, Office of Extramural Research, *NIH Data Sharing Policy*, <http://grants.nih.gov/grants/policy/data_sharing/>. Accessed 9/12/07.
28. *Data and Sample Policy*, <<http://www.nsf.gov/pubs/2004/nsf04004/start.htm>>. Accessed 9/12/07. Also see <<http://nsidc.org/arcss/protocol/protocol.html>>. Accessed 9/12/07.
29. See *ILTER Data Policy* <<http://www.lternet.edu/data/netpolicy.html>>. Accessed 9/12/07; Porter and Callahan, 1994, "Circumventing a dilemma: Historical approaches to data sharing in ecological research," in *Environmental Information Management and Analysis: Ecosystem to Global Scales*, Michener, Brunt, and Stafford (eds), p193-202. See also Ann Zimmerman (forthcoming). "New knowledge from old data: The role of standards in the sharing and reuse of ecological data." *Science, Technology, and Human Values*. <http://www-personal.si.umich.edu/~asz/zimmerman_sthv_manuscript_accepted_april_2007.pdf>. Accessed 9/12/07; and Peter Arzberger et al., "Promoting access to public research data for scientific, economic, and social development," *Science* 3: 135-152. My thanks to Karen Baker for providing these references.
30. *IWG Terms of Reference*, January 2007, <<http://iwg.cfa.harvard.edu/twiki4/pub/IWGDD/IwgddTermsOfReference/>>. Accessed 9/12/07.
31. Chuck Humphrey, "Secondary data analysis," <<http://datalib.library.ualberta.ca/nphs/SecondaryDataAnalysis.ppt>>. Accessed 9/12/07.
32. AGU, *Policy on Referencing Data in and Archiving Data for AGU Publications*, <http://www.agu.org/pubs/data_policy.html>. Accessed 9/12/07.
33. Inter-University Consortium for Political and Social Research, founded 1962, <<http://www.icpsr.umich.edu/>>. Accessed 9/12/07.
34. E-Bank UK, *Data Citation*, <<http://www.ukoln.ac.uk/projects/ebank-uk/data-citation/>>. Accessed 9/12/07.
35. Publication and Citation of Scientific Primary Data (STD-DOI) <http://www.std-doi.de/front_content.php>. Accessed 9/12/07.
36. See for example the NIH policy posted at <<http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-042.html>> on sharing of model organisms. Accessed 9/12/07.
37. ARL *Bimonthly Report* No. 249, December 2006, p. 5, <<http://www.arl.org/bm~doc/arlb249datasteward.pdf>>. Accessed 9/12/07.
38. Conversation with MacKenzie Smith, June 2007. Indeed Google has begun to get into the data act, with its beta database repository and indexing service, Google Base. <<http://base.google.com/>>. Accessed 9/12/07.

39. Karen S. Baker and Florence Millerand, "Scientific infrastructure design: Information environments and knowledge processes, to appear October 2007 in *Proceedings of the American Society of Information Science and Technology*. Preprint at http://interoperability.ucsd.edu/docs/07BakerMillerand_07asist_KnowledgeProvinces.pdf. Accessed 9/12/07.

Copyright © 2007 Anna Gold

doi:10.1045/september20september-gold-pt1