

D-Lib Magazine
September/October 2007

Volume 13 Number 9/10

ISSN 1082-9873

Cyberinfrastructure, Data, and Libraries, Part 2

Libraries and the Data Challenge: Roles and Actions for Libraries

[Anna Gold](#)

Massachusetts Institute of Technology

<annagold@MIT.EDU>

2.1 Data librarianship today

Most academic library services support the delivery and stewardship of text-based collections in a variety of print and digital formats. There are exceptions: libraries also manage and deliver images, multimedia, sound, maps, and various other artifacts of research and culture, including data and datasets. Today, libraries' roles in data management and data services tend to relate to a few well-defined categories of data: social science data, geo-referenced data (GIS), and bioinformatics.¹

Social science data services

While social science data services in libraries have existed for decades, the advent of digital data has brought major enhancements to access and manipulability of data. A primary service focus has been support for access to government-produced social science data. However there are many other sources of social science data; in the US, a huge archive of social science data has been collected at the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan. Founded in the 1960's² as part of the Institute for Social Research, ICPSR manages an archive that includes over 50,000 data sets, and provides services that include regular training programs, and online access and analysis of selected studies. The profession of social science data librarianship is well established, with a dedicated international professional association (IASSIST), and specialty training programs either in place or proposed at many library and information schools.³ A workshop held in 1999 by the Digital Library Federation (DLF) on the "state of the art" in social science data librarianship is a useful snapshot both of current accomplishments and future challenges faced in this relatively mature area of data librarianship.⁴

GIS services

"GIS began appearing in academic libraries during the 1980s, but it truly began to be a part of library services when the 1990 census materials were given to [U.S. federal] depository libraries as TIGER (Topologically Integrated Geographic Encoding and Referencing) files. This was followed shortly by a variety of other government data that was distributed by other agencies as part of the depository program⁵ (Abbott and Argentati 1995). Since that time, applications for and use of GIS has blossomed in academic libraries. GIS has moved from being a tool used mostly by map or document

libraries, to being a tool that can be used by reference librarians to meet the needs of any number of disciplines."⁶

GIS services in libraries today are widespread, if not yet universal: a 2006 survey of GIS services in 103 libraries belonging to two academic library consortia reported that 31 had already implemented GIS services and an additional 15 were considering its implementation.⁷ Even among smaller academic libraries, GIS services are not uncommon, though they face challenges that include underfunding and underutilization.⁸

Bioinformatics

While support for social science data and GIS are relatively well-established services in academic libraries, every flavor of bioinformatics services in research libraries, from reference and referral, to training, to expert consultation on data manipulation or management, is in a relatively new stage. An account of the range and diversity of these services was given in a 2006 survey describing services established "in collaboration with biology faculty as part of formal courses, through teaching workshops in the library, through one-on-one consultations, and by other methods. Librarians with backgrounds from art history to doctoral degrees in genetics have worked to establish these programs."⁹ The survey illustrates the growing use and acceptance of bioinformatics support programs in libraries, fulfilling Clifford Lynch's prediction in 1999.¹⁰

What social science data, GIS, and bioinformatics services all have in common is notable: all three services take place in environments that are characterized by large, dedicated data centers; much, though not all data manipulation takes place using widely adopted and well-supported commercial or open source software; and extensive training, from novice to expert levels, is offered on a regular basis by national organizations or enterprises: it is impossible to imagine social science data services without thinking of ICPSR; of GIS without thinking of ESRI (Environmental Systems Research Institute); or of bioinformatics without the National Center for Biotechnology Information (NCBI).

Other data services

Even when they don't offer a full-service data program or hire dedicated staff with data expertise, libraries provide significant data services across many fields. Many libraries acquire data sets through the U.S. federal depository program, or through traditional acquisitions and licensing sources. Indeed, reference data that used to be available in the form of printed handbooks can now be found on the web or as part of commercially produced digital products. Enhanced searching and linking within these sources increase their value to library users. An example of an open source of such data is the National Institute of Standards and Technology (NIST), whose growing family of web-based portals to scientific and technical data is an important library reference source.¹¹ An example of a commercial source is Knovel, an aggregation of data reference works from publishers such as Wiley and Elsevier.¹² Another example is BioBase, whose "Knowledge Library" includes both databases and integrated analysis tools. Libraries that have traditionally purchased book-like tools such as Knovel are finding themselves being asked to acquire data in new forms.

Archival data services

Librarians have also had an important though not unique role in the digitization and curation of historic data sets. At the University of California, San Diego, a collection of fish catch statistics, digitized as part of an NSF-funded project (California Explores the Ocean), found new life and new audiences.¹³ At Cornell University, librarians are involved in creating metadata for an archive of US blackout data.¹⁴ A recent article explored some of the issues for libraries in planning and executing (social science) data digitization projects.¹⁵ While other organizations with research interests or stewardship roles related to historical data have also engaged in major data digitization and archiving, some see libraries as cultural institutions with a uniquely long-term responsibility

for stewardship of the scholarly record. Still, the extent of the roles libraries can play in relationship to data archiving is a question under debate and discussion today.¹⁶

To sum up, there is clearly a wide range of viable roles for libraries and librarians in relationship to the overall production and use of scientific and technical data. Whether libraries can play these roles, and why they might want to, are issues still up for debate and discussion.

2.2 Why data matters to libraries

While data librarianship has developed as a specialty service analogous to services to other "special formats" (e.g., maps, music), some would argue that data librarianship breaks more radically with traditional models of academic librarianship, presaging a greater role for libraries in pre-publication scholarship. Jim Jacobs, now emeritus social science data librarian at the University of California, San Diego, has argued that in the future all librarians may have an expanded role in information services earlier in the research cycle:

"[D]ata-services will at some point be a model for library services in general as libraries deal increasingly with digital materials and become more of a lab than a warehouse....This opens up the possibility of the data librarian working with researchers at the earliest stages of research: helping with the documentation process and ensuring that the data will be preservable, usable, and re-usable for the long-term."¹⁷

The data life cycle begins early in the scientific process (what D. Scott Brandt refers to as "upstreamness"),¹⁸ and new library services may develop around those early research stages. There are also expanding opportunities for libraries downstream for post-production and post-publication services. The digital environment is making it possible to create access to a greatly expanded scientific record – one that is less dependent on papers and articles, and that is increasingly expressed in terms of networks of links and associations among diverse research artifacts. Linking data in rich and robust ways to support data reuse and integration will require understanding and documentation of the data's provenance, the development of ontologies, expert annotation, and analysis. Further downstream, services enabled by these activities will include visualization, simulation, data mining and modeling, and other forms of knowledge representation and extraction.

Commercial publishers are increasingly interested in data, both in creating tools for data management and discovery,¹⁹ and in publishing data sets in peer-reviewed data journals.²⁰ The synergies between data and traditional scientific literature include the potential to mine published scientific literature for interesting data and to expand and transform the traditional notion of what a "publication" is.²¹ While standards and business, distribution and discovery models are well understood for textual publishing, librarians have a deep vested interest in systems of data publishing. Creating such systems will require new skills in including managing datasets as complex objects; it will also require creating standards related to data publication, description, citation, discovery, and reuse; and addressing policies for data disclosure.

At the most fundamental level, engaging the library profession in the problem of data management may lead to reframing the values and practices of the library profession. Where today library practices appear to be rooted in the management and delivery of objects (whether virtual or physical), from another point of view those practices are rooted rather in the management and "delivery" of relationships. And data is, after all, an encoding of relationships in the world, whether those relationships involve instruments, physical phenomena, social entities, measurements, time, place, or other intellectual constructs.²²

Whatever the challenges of scientific data for librarians, the potential for these challenges to be both revitalizing and transformative for librarianship is sufficiently attractive to bring many librarians to the data table.

2.3 Proposed roles for libraries in E-Science

Much of the recent discussion about the roles of libraries and librarians in data services has been inspired, not by the established data librarianship practices such as those described above, but rather by developments in cyberinfrastructure and E-Science.

An important milestone in developing ideas about library roles in cyberinfrastructure was an ARL-NSF workshop in October 2006.²³ Participants in the workshop expressed many different views of the roles that will be most critical as well as most realistic for libraries to undertake, though most agreed that delivering on these roles will require the development of new skills and possibly new career paths within the library workforce. They also agreed that to be effective in the arena of data stewardship, libraries will need to work in partnership with non-library organizations and professionals.

One way of characterizing the range of roles proposed for libraries, both at the workshop, and elsewhere, is to describe them in terms of pre- and post-publication stages of the data life cycle (upstream vs. downstream). This approach does have a disadvantage: it may suggest that the data life cycle is serial and terminal. In fact the opportunities afforded by digital data include more dynamic and fluid linkages throughout a "stream" or life cycle, including the possibility of reusing data. Such reuse is arguably both downstream and upstream at the same time! Still the demarcation may be a useful one to consider, particularly since established library roles and capabilities tend to fall in the "downstream" direction and new library roles may deal more with the "upstream" parts of the research cycle.

Downstream

Within the "downstream" side of the research cycle, librarians can play roles in the selection, acquisition, and licensing of data and data sets; in creating metadata (or metadata standards) for discovery and description of data sets; in creating or organizing documentation related to data; and in offering preservation services for digital data – at least to the extent that data preservation capabilities exist anywhere. A role associated more with archives than with libraries, but common to both, is advising in the appraisal and selection of what data to keep for the long term. Another role libraries are well positioned to play is assisting users with finding data relevant to their research, using third-party high level directories and data discovery sources such as the Global Change Master Directory (<http://gcmd.nasa.gov/>) or the National Space Science Data Center (<http://nssdc.gsfc.nasa.gov/>).

In systems of scholarly publication today, libraries work with scholarly societies and academic publishers to advise and help develop publishing standards and systems. Libraries could play a similar role in developing data publication standards and systems. Work on these standards and systems is still in very early stages.²⁴ Among the many data publishing standards and systems needed are publishing workflows, global identifier schemes, linking schemes, standards for data clean-up and normalization, and also standards for providing credit and recognition to data authors. Similarly, as with other forms of publishing, there is a need for systems, policies, and practices and for standards for rights management. Libraries today play a vital role in advocating for intellectual property policies for scholarly texts that serve the advancement of ideas and knowledge. They can also play a role advocating for the documentation of rights and intellectual property in relation to data, and help make the case for an overarching vision for open science, which might include creating national or trans-national depositories of unpublished and supplementary data.²⁵

Another role libraries can play is to offer long-term repositories of scholarly output. An expanded notion of this role has included creating institutional or disciplinary repositories of digital objects that are part of the scholarly record. While most of these objects to date consist of textual objects (such as articles, technical reports, and theses), creating repositories for data might build on this work. The independence of such repositories from classic publishing workflows suggests that the institutional repository might bridge the downstream/upstream divide. Indeed, some participants in the ARL-NSF workshop offered a vision of a tiered, networked system of data repositories, with distributed (local) repositories for data early in the data life cycle, and centralized depositories for long-term curation of data later in the cycle. While long-term curation and stability of the scholarly record are among the aims of such repositories, libraries also have experience and interest in

contributing value to repositories by building semantics for heterogeneous data, adding functionality for users, and creating incentives to share research output.²⁶

Upstream

Key to libraries or librarians playing more "upstream" roles in data science is their ability to position themselves as partners in research. By collaborating closely, and early, in the research process, librarians may become involved in creating data curation prototypes, or otherwise supporting the use of documentation, practices, or standards that will assure the longevity of the data downstream. Such close collaborations are far from common, but examples do exist, including the work of the Johns Hopkins University Libraries with the National Virtual Observatory;²⁷ and the establishment at Purdue University in 2006 of a Digital Data Curation Center, as an incubator of data collaborations between librarians and faculty. Another potential library role, building on libraries experience with institutional repositories, might be to create more dynamic repositories that support pre-publication workflows, including collaboration environments supporting data integration, analysis, and visualization.

In sum, it is fair to say there is still a substantial amount of uncertainty about the roles libraries can play in scientific data management, reflecting an environment of ongoing experimentation and negotiation (and perhaps some wishful thinking). While libraries can draw on experience with social science data and GIS, and are developing similar experiences in bioinformatics, many of the challenges of cyberinfrastructure and E-Science appear to be of a different order of magnitude altogether.

One reason for this is that cyberinfrastructure and E-Science are often used to mean a pervasive and grand-scale infrastructure serving widely distributed and often large-scale research activities. Not all science takes place at that scale, and even where it does, the infrastructure, community practices, and requirements of broad domains and cross-disciplinary research work vary widely. The role of libraries in such large-scale developments is particularly unclear, given the lack of capacity and expertise in libraries to deal with giga- and peta-scale data storage, high performance computing and data processing, let alone the significant domain-specific expertise needed to support data curation and use. The data infrastructure needs of small and medium sized research, often characterized by local practices and expertise, may be a different story.

2.4 Building capacity and understanding

Building the capacity of libraries and librarians to take up these opportunities, whether upstream or downstream, will present significant challenges, not the least of which is building the skills and knowledge needed to carry out new roles. But what skills are required? In some cases, domain expertise is essential to effectively work with researchers in the upstream phases of research data management and planning. Domain expertise may also be needed to provide credible expert help with data management problems or tools. It has been argued that it makes much more sense to train domain experts in data management and curation skills than it does to try to teach non-scientist librarians to understand the infrastructure and service needs of a domain.

In either case however data librarians will need to forge embedded working relationships with research teams, rather than working through more distal relationships with faculty that are common in larger university research libraries. This is no small challenge, but building fluency across library and scientific cultures will be essential to effective data librarianship. Pauline Simpson of the National Oceanography Centre has made the case that the demands of data curation will make the skills of librarians and information managers more evident to researchers:

"Traditionally, the information and data communities have developed along parallel though not converging lines, but changing attitudes towards open access to the results of scientific research have resulted in new partnerships in which libraries and information managers are working with the data community on new information products. Information management skills: standards, metadata, rights management, discovery services, preservation and particularly service provision are now being accepted as a vital underpinning to the success of the e-science agenda."²⁸

As important as domain knowledge may be for some types of data librarianship, individuals with broad research library skills will bring their own important perspectives on the relationships among different parts of the scholarly record, on patterns of search and use, and on the potential for long-term and multi-disciplinary use of research products that a domain expert might miss. Also, while librarians with domain expertise may be prepared to partner with researchers and faculty, library relationships with laboratory and research center data managers also deserve further attention and exploration. While not the same, data manager and librarian roles are related. The concept of "knowledge provinces," a model of the different roles and relationships among data management, information management, and cyberinfrastructure across a range of research from small to large, and from simple to complex, may be useful in understanding these relationships.²⁹

Whatever their existing preparation and expertise may be, librarians and scientists alike could benefit from additional specialist training and new work environments that better handle emerging practices and issues in data curation and related data services. At least one university has launched a skills-oriented "capacity building" program for data librarianship in the sciences – in 2006, UIUC announced it was developing a master's concentration for librarians working in data curation.³⁰ The curriculum includes required courses in data curation and digital preservation, and electives in topics that include biodiversity and ecoinformatics; information modeling, ontologies, information transfer and collaboration in science; and design of digitally mediated information services.

Libraries struggle today to fund their hybrid programs of bricks and mortar and digital services. If they are to fulfill new roles in data librarianship, libraries will also need adequate resources – staff and funding. Libraries' experience with social science data and GIS underscores the importance of creating productive partnerships with organizations that have complementary strengths and responsibilities to steward data. The broad arena of data stewardship and service is so vast and still so new that libraries will need to choose their roles carefully, to match their own strengths and to fulfill a shared vision or mission with potential partners.

An interesting case has been made that research libraries in fact offer one of the most promising funding models for ensuring the long-term stewardship of scientific data archives. University libraries have a diversified funding base that includes basic institutional support for operations and infrastructure; private funding from gifts and foundations; research funding from foundations and other institutions; as well as other funding streams that may include cost-recovery and government bond issues.³¹ This diversified funding base, together with libraries' core mission to sustain long-term access to the research record, and their culture of standards development and collaboration across institutional boundaries, puts libraries in a unique position relative to other stakeholders, including publishers, funding agencies, scientists, and university technology centers. In considering their capacity to take on new roles in data science, libraries may take heart from this vote of confidence in a model that has demonstrated resilience and longevity.

2.5 Toward an integral role for libraries and librarians

Once libraries and librarians understand the opportunities of integrating data librarianship into their services and perspectives, and are convinced of the value they can bring to the table, they will need to invest time in developing new skills and crossing cultural borders. Once they have forged new partnerships with scientists and data managers, then libraries and librarians will be truly integral to the stewardship of data as a vital part of the scholarly record.

Among the actions that librarians may find it possible and valuable to take are to:

- Participate in the growing number of professional data curation conferences;
- Read key documents relating to data science, and attend, where possible, training related to data curation and data science;³²
- Network with data managers supporting research programs, and work to understand their perspectives,

practices, and culture;

- Expand data acquisition activities beyond social science and GIS data;
- Develop and market data consultancy and referral services;
- Understand and support technologies and systems of data publishing that allow for re-use of data;
- Advise on and advocate for systems and standards of data description that encourage interoperability; and
- Articulate data and informatics issues.

In addition, library leaders and library organizations may:

- Educate and advocate for institutional commitments to require data management plans that support long term access;
- Encourage conceptual dialogue regarding data and informatics efforts;
- Advocate for responsible but open access to data; and
- Cultivate partnerships with library and non-library organizations sharing a common interest in data stewardship of a particular area of research.

Acknowledgements

I wish to thank members of the MIT Data Initiatives Study Group (DIG) in the MIT Engineering and Science Libraries (<http://web.mit.edu/dig>): Anne Graham, Erja Kajosalu, Louisa Worthington Rogers, and Amy Stout; as well as other MIT Libraries staff who consulted regularly with the DIG group: Kate McNeill-Harman and Lisa Sweeney. Many colleagues offered helpful comments and encouragement on drafts of this article, in particular Karen Baker, Steve Gass, and MacKenzie Smith. I also wish to thank Karen Baker, Melissa Cragin, and John Wilbanks for generously sharing their time, insights and research on data librarianship and cyberinfrastructure.

Notes and References

1. For instance at MIT Libraries, one professional position supports use of GIS, with a second position funded at the campus level; and one professional librarian serves social science data needs, with additional technical support provided through a cooperative relationship between Harvard and MIT. In major research libraries at least one dedicated professional for each type of data is a common staffing model; in some cases data services for social science data and GIS are closely integrated. Formal bioinformatics service programs in libraries are less common in academic libraries, with the exception of some larger medical libraries.

2. ICPSR, <<http://www.icpsr.umich.edu/>>. Accessed 9/12/07.

3. A special curriculum for social science data librarianship was proposed by Frank Olke and Fredric Gey in 2006. The content of the curriculum they propose might be offered via a variety of academic departments and would include courses in social science data sets, statistical database management, metadata and data semantics, data library operation, statistical disclosure analysis and networking. "Social science data librarianship – A university curriculum," Jan. 11 2006 v. 8. [sic] <http://hpcrd.lbl.gov/staff/olken/ssdl/iassist_ssdl_curriculum.pdf>. Accessed 9/12/07.

4. *Digital Library Federation Workshop on Social Science Data Archives: Report of the meeting held January 27-28, 1999* (Princeton University), <<http://www.diglib.org/collections/ssda/ssdaresults.htm>>. Accessed 9/12/07.

5. For non-librarians, the US federal depository program may be unfamiliar. One way to think of it is as a pre-digital distributed library of federal information, with well-defined redundancies and archival responsibilities.

6. Camila Gabaldón and John Replinger, "GIS and the academic library: A survey of libraries offering GIS services in two consortia." *Issues in Science and Technology Librarianship*, 2006 n. 48.

<<http://www.istl.org/06-fall/refereed.html>>. Accessed 9/12/07.

7. Ibid.

8. JaNae Kinikin and Keith Hensch, "Follow-up survey of GIS at smaller academic libraries." *Issues in Science and Technology Librarianship*, 2005, n. 43. <<http://www.istl.org/05-summer/article1.html>>. Accessed 9/12/07.

9. David Osterbur et al., "Vignettes: diverse library staff offering diverse bioinformatics services," *Journal of the Medical Library Association*, 2006 July; 94(3): 306, E188-E191.

<<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1525318>>. Accessed 9/12/07.

10. Clifford Lynch, "Medical libraries, bioinformatics, and networked information: a coming convergence?" *Bulletin of the Medical Library Association*, 1999 October; 87(4): 408-414.

<<http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=10550026>>. Accessed 9/12/07.

11. See NIST's Data Gateway <<http://srdata.nist.gov/gateway/>> and index <<http://www.nist.gov/srd/>>. Accessed 9/12/07.

12. Knovel, <<http://www.knovel.com/>>. Accessed 9/12/07.

13. California Explores the Ocean, *Fish Bulletin's Fish Catch Statistics*,

<<http://ceo.ucsd.edu/fishcatchtables/fish-catch-download.html>>. Accessed 9/12/07.

14. See description of Cornell University Metadata Services, <<http://metadata.library.cornell.edu/projects.htm>>. Accessed 9/12/07.

15. Julie Linden and Ann Green, "Don't leave the data in the dark," *D-Lib Magazine*, January 2006, v. 12 n. 1 <[doi:10.1045/january2006-linden](https://doi.org/10.1045/january2006-linden)>. Accessed 9/12/07.

16. See Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, Consultancy Report, UKOLN, University of Bath, 19th June 2007. Lyon's chart (p. 56) of "roles and responsibilities" mentions data centers, scientists, users, and institutions, but not libraries or librarians.

<http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx>.

Accessed 9/12/07. A related statement of principle authored by Stephane Goldstein for the RIN (Research Information Network), *Stewardship of Digital Research Data – Principles and Guidelines*, does mention libraries as key collaborators in data stewardship. <<http://www.rin.ac.uk/data-principles>>. Accessed 9/12/07.

17. James Jacobs and Charles Humphrey, "Preserving research data", *Communications of the ACM*, Vol. 47 (9), pp. 27-29.

18. D. Scott Brandt, "Data, research, metadata, metaresearch," presentation at ACRL/STS, ALA annual meeting, June 2007. <<http://www.ala.org/ala/acrlbucket/stsconferencepro/annual2007programs/brandt.pdf>>. Accessed 9/12/07.

19. Examples include data management tools such as Elsevier's MDL Isentris,

<<http://www.mdl.com/products/framework/isentris>>. Accessed 9/12/07; knowledge management tools such as

ChemDraw from CambridgeSoft <<http://www.cambridgesoft.com/software/ChemDraw/>>. Accessed 9/12/07; and "deep indexing" tools such as CSA Illustrata, offering access to data, variables and other content represented in tables, maps, photographs and other figures in journal literature

<<http://www.csa.com/factsheets/objectscust-nats-set-c.php>>. Accessed 9/12/07.

20. Data journal publishing is not new: a long-standing example is the *Journal of Physical and Chemical Reference Data*, published since 1972 by ACS and later AIP for NIST. New models for data publishing have been discussed recently by Donald Waters of the Mellon Foundation,

<<http://www.sis.pitt.edu/~repwkshop/papers/waters.pdf>>. Accessed 9/12/07; an "overlay journal" approach to

data publishing is being piloted in the UK . See the Overlay Journal Infrastructure for Meteorological Sciences (OJIMS) Project <<http://www.see.leeds.ac.uk/research/ias/dynamics/current/ojims.html>>. Accessed 9/12/07.

21. These efforts are not limited to the commercial world; an exciting recent development is Science Commons' unveiling of the Neurocommons, a data mining tool developed in the commercial sector and released as open source. <<http://sciencecommons.org/projects/data/>>. Accessed 9/12/07. Other innovative approaches to blended and enhanced publication of data, text, and other media are described in "The coming revolution in scholarly communications & cyberinfrastructure," *CT Watch Quarterly*, vol. 3 no. 3, August 2007, <<http://www.ctwatch.org/quarterly/articles/2007/08/>> Accessed 9/12/07.

22. See Brandt (Ibid.) citing the ACRL Roundtable on Technology and Change in Academic Libraries, November 2-3, 2006: "...the Web gives rise to changing conceptions of knowledge production and use. ...Knowledge that is fluid and even imperfect today carries higher value than knowledge perceived as static and intact. Data that can be copied, pasted, mixed, adapted, recast for evolving purposes and new modes of understanding has very strong appeal in today's information environment, particularly for young people. The problem of managing and preserving knowledge produced in these shifting realms of digital proliferation is enormous, and it is one that librarians need to be integral to solving." <<http://www.ala.org/ala/acrl/acrlissues/future/changingroles.htm>>. Accessed 9/12/07.

23. *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>>. Accessed 9/12/07.

24. An example of such a standard is DDI, an XML-based standard for the content, presentation, transport, and preservation of social science data sets. <<http://www.icpsr.umich.edu/DDI/>>. Accessed 9/12/07.

25. For example, in January 2005, CISTI and CODATA issued a report on data access proposing that a national body ("Data Canada") be established as a depository of unpublished and supplementary data. See: *National Consultation on Access to Scientific Research Data* <<http://ncasrd-cnads.scitech.gc.ca/>>. Accessed 9/12/07).

26. See for instance work at MIT Libraries on SIMILE <<http://simile.mit.edu/>>, accessed 9/12/07); and the incentives of increased exposure and citation offered to faculty by librarians marketing their institutional repositories, for which see "Institutional repositories: A review of content recruitment strategies," Timothy Mark and Kathleen Shearer, Canadian Association of Research Libraries, June 27, 2006, <http://www.ifla.org/IV/ifla72/papers/155-Mark_Shearer-en.pdf>. Accessed 9/12/07.

27. For a description of the JHU partnership, see Pamela Higgins, "Virtual Observatory to preserve massive cosmic images online," *Johns Hopkins Gazette*, October 23, 2006, <<http://www.jhu.edu/~gazette/2006/23oct06/23virtua.html>>. Accessed 9/12/07; for the Purdue center, see <<http://d2c2.lib.purdue.edu/d2c2about.html>>. Accessed 9/12/07.

28. Pauline Simpson, National Oceanography Centre, UK, "Libraries supporting E-Science: Combining cultures," EURASLIC 12, Abstract at: <<http://library.ibss.org.ua/Default.aspx?tabid=304>>. Accessed 9/12/07.

29. 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.

30. GSLIS, Master of Science--Concentration in Data Curation, <http://www.lis.uiuc.edu/programs/ms/data_curation.html>. Accessed 9/12/07.

31. Chris Greer, "The digital data universe of the future," Library of Congress webcast, 10/11/06 <http://www.loc.gov/today/cyberlc/feature_wdesc.php?rec=3937>. Accessed 9/12/07.

32. See Amy Stout's account of these activities at the MIT Libraries, "The data dilemma," paper presented at the

American Society of Engineering Education (ASEE) Conference, June 2007, in Honolulu Hawaii.

Copyright © 2007 Anna Gold

doi:10.1045/july20september-gold-pt2