Research Libraries and Digital Library Research for Cyberinfrastructure-enabled Knowledge Communities (CKCs)

Daniel E. Atkins
The University of Michigan, U.S.A.
atkins@umich.edu

Introduction

Research libraries have stood as the primary institutions for access to and effective use of recorded knowledge in support of teaching, research, scholarship, and in some cases community service. This mission is not strictly reserved for institutions calling themselves "libraries" and it generally includes archives, historical societies, museums, and other institutions devoted to improving access to information that supports research and learning[1]. Most research libraries are now well into the reality of the digital information age and moving at various rates and levels of success to cope with the challenges and opportunities it presents. There is increasing recognition that the ultimate impact of IT on libraries is not just automating what they have always done, or even limited to digitizing the paper collections. The concept of "digital libraries" while useful in the interim, is part of a moving target of vision, aspirations, and expectations of the funders and clients of research libraries. The digital library concept is becoming a part of an even larger integrated concept driven by information technology application that truly revolutionizes how information and knowledge are created, disseminated, and preserved. There are no standard names for this concept. In this paper I will use the term *cyberinfrastructure-enabled knowledge communities* (CKCs for short). Examples of CKCs go by a variety of names including *co-laboratories, collaboratories, grid communities, e-science communities, and virtual communities*. This paper will review the recent digital library movement; the current emergence of the "cyberinfrastructure movement," and conclude with some of its implications for the future form and function of research libraries.

The Digital Library Movement

Over the past decade, the exponential change in price-to-performance curves of digital technology has crossed thresholds that now enable rapid and broad adoption of electro-optical digital representation of information. It has brought into focus the concept of "digital libraries." This term has been applied to information collections and services as well as to research and development activities made possible and necessary by the

[1] I do not intend to downplay the importance of special or public type libraries, but I assume these type libraries are functionally contained within research libraries and that lessons learned there can be transferred to the world of public and special libraries.

exponential growth and adoption of digital technology for the representation, storage, retrieval, and preservation of multimedia information. The stakeholders in these activities include:

1. the leaders and professionals of existing libraries (and other cultural institutions including archives and museums) that have become a hybrid of traditional and digital collections and services;

2. the computer and information technologies research and development communities that pursue digital libraries as an advanced application of distributed computing systems and in particular as services within the global internet-world-wide-web infrastructure; and

3. a wide range of knowledge-based communities and institutions (information users), that depend upon organized information for commerce, entertainment, learning and discovery.

My experience has been that the best research and development in digital libraries occurs in projects that align mutual self-interest and create joint learning between these three stakeholder groups. The cyberinfrastructure movement I will briefly describe has revolutionary implications for all three groups, although I will give more emphasis in this paper on the implications for libraries as institutions.

The growing ubiquity of digital infrastructure means that the bulk of new information is "born digital." Government and private funding have also funded retrospective conversion to digital representation of important collections of physical-analog materials including text, sound, images, motion pictures, and even 3-dimensional objects. The latest version of the Berkeley project How Much Information Project 2003 www.sims.berkeley.edu/research/projects/how-much-info-2003 estimates that print, film, magnetic, and optical storage media produced about 5 exabytes of new information in 2002. An exabyte is 10^{18} bytes[2].

The core mission of libraries, *access to information,* has three dimensions: physical access, intellectual access, and long-term (preservation) access. The general lack of guaranteed long term access to (or preservation of) digital objects is finally being widely recognized as a major impediment to the adoption of exclusively digital libraries, but much remains to be done to create the technology and institutions to do so. A significant web resource on this topic is at www.digitalpreservation.gov/index.php and a new report on research challenges in digital archiving and long-term preservation is available at www.si.umich.edu/digarch. Furthermore, the distinction between the mission and operations of differently named cultural institutions is increasingly blurred in the digital age. Libraries, museums, and archives may become themes and variations on distributed but federated "digital collectives."

[2] *How big is five exabytes?* If digitized with full formatting, the seventeen million books in the Library of Congress contain about 136 terabytes of information; five exabytes of information is equivalent in size to the information contained in 37,000 new libraries the size of the Library of Congress book collections.

We are indeed in a digital information age and appropriately, government and private sponsors in the past decade initiated research and development to explore the interdisciplinary topic of function, form and use of digital libraries. In the United States the primary leader sponsoring digital library research and development has been the National Science Foundation (NSF) through two major digital library initiatives (DLI-1 and DLI-2.) These initiatives helped establish an international digital library research community with many projects including collaborations between research libraries/librarians, technologist, and specialized information user communities. The larger projects were focused and informed by the creation, use, and evaluation of experimental prototype systems, with some of these becoming operational systems. Google, for example, was in part an outgrowth of NSF sponsored work at Stanford University. The NSF DLI also led to the birth of D-Lib Magazine www.dlib.org an important repository of the results of digital library R&D.

The NSF has also coordinated with digital library R&D in Europe, Australia, and Asia and as this audience knows well, digital libraries are the topic of numerous international conferences and journals. Digital library research at NSF is not continuing as a special initiative but is rather subsumed in regular programs and likely to be relevant to emerging cyberinfrastructure research and development programs. A report from a recent NSF workshop on the topics for future digital library research is available at www.sis.pitt.edu/~dlwkshop. As articulated well by Borgman[3] digital libraries continue to be important as part of the bigger dream of a global information infrastructure serving the entire world in a seamless way.

The agenda for digital library research has been broad and multidisciplinary including topics such as distributed systems architecture; information federation, access and retrieval of multi-media objects; information economics, agent technologies, metadata structures and their automated creation, human-factors, evaluation of use, and multi-lingual issues. Libraries as institutions are also interested in the training/re-training of human resources, business models, evaluation, and contemporary metrics for library capacity and effectiveness. Research libraries are noticing, as reported recently in the New York Times[4], that increasingly patrons, even faculty, are making a trip to a physical library only as a last resort if online digital resources cannot be found. The primary entrance to the library is becoming a web portal and thus libraries are paying more and more attention to the quality of their web portals and issues of human-computer interaction.

Concurrent with the emergence of digital libraries has been the emergence of important new academic programs to educate the appropriate professionals to create and manage

[3] Borgman, C.L., From Gutenberg to the global information infrastructure : access to information in the networked world. Digital libraries and electronic publishing. 2000, Cambridge, Mass.: MIT Press. xviii, 324.

[4] Katie Hafner, "Old search engine, the library, tried to fit into a Google world," *New York Times*, June 21, 2004. Available for fee at http://query.nytimes.com/gst/abstract.html?res=F00F12FB395D0C728EDDAF0894DC404482

such environments from a socio-technical perspective – digital library architects and digital librarians. In some cases traditional library schools have been transformed into much broader entities and become part of a new *information schools* movement. See for example www.ils.unc.edu/ils/releases/RELEASE_deansPanel.html and the website for the University of Michigan School of Information at www.si.umich.edu.

The transformation of libraries from print-on-paper collections to a hybrid of digital and paper is well underway. But there is still much to be done in the appropriate institutional change, especially in the area of national and international library cooperation (or eventually mergers) in federating collections and services that are fundamental to anytime and anyplace information access. In theory digital libraries enable common holdings to be held and managed in common, and individual libraries to differentiate themselves, not by hosting redundant collections but rather through the creation, curation, and stewardship of unique digital collections that they share with the rest of the world. In libraries as well as many other institutions in the academic and commercial world, information technology is causing re-evaluation of the incentives and policies for the balance between when to cooperate and when to compete.

Cyberinfrastructure also enables functional disaggregation -- institutions other than libraries may assume library functions and libraries may assume some of the traditional functions of publishers or book stores. HighWire Press highwire.stanford.edu , for example, is a division of the Stanford University Libraries, which produces the online versions of high-impact, peer-reviewed journals and other scholarly content. The JSTOR Project www.jstor.org is a successful example of a non-profit organization providing international access to scholarly journal collections. JSTOR is now part of a constellation of similar services including ARTstor under the non-profit umbrella, ITHAKA www.ithaka.org. There are rumors that the for-profit Google web searching service intends to expand into the business of hosting substantial collections of scholarly holdings.

Transformation of Scholarly Communication

Even though the digital transformation of libraries is still very much unfinished, it is now part of an even bigger picture that includes the transformation of the processes and formats of scholarly communication. Digital publishing through the web eliminates the upfront, fixed-cost of printing and distributing ink on paper. It also enables new multimedia formats that are born digital; that may include audio, video, data sets, and interactive programs that have no print-on-paper equivalent. It blurs the distinction between libraries, clients, and publishers and potentially disaggregates the stages in the life cycle of information creation, access, use, and re-use. Several trends are noteworthy:

1. Serious exploration of open (free) access to well-credentialed publications in which authors do not give away their copyright to commercial publishers who sell them back to libraries at high profit margins. A good example is the Public Library of Science. [www.publiclibraryofscience.org]. Recently Springer, the world's second largest scientific publisher announced adoption of Open Access (OA) publishing. Springer Open Choice allows authors to choose to

pay $3000 for OA print and online publishing. [Information World Review http://www.iwr.co.uk/iwreview/1156517].

2. Evolution of alternate licensing models for digital objects that help re-establish a public domain of resources and encourage their creative use in derivative products. Creative Commons creativecommons.org is at the forefront of this movement. It is "devoted to expanding the range of creative work available for others to build upon and share."

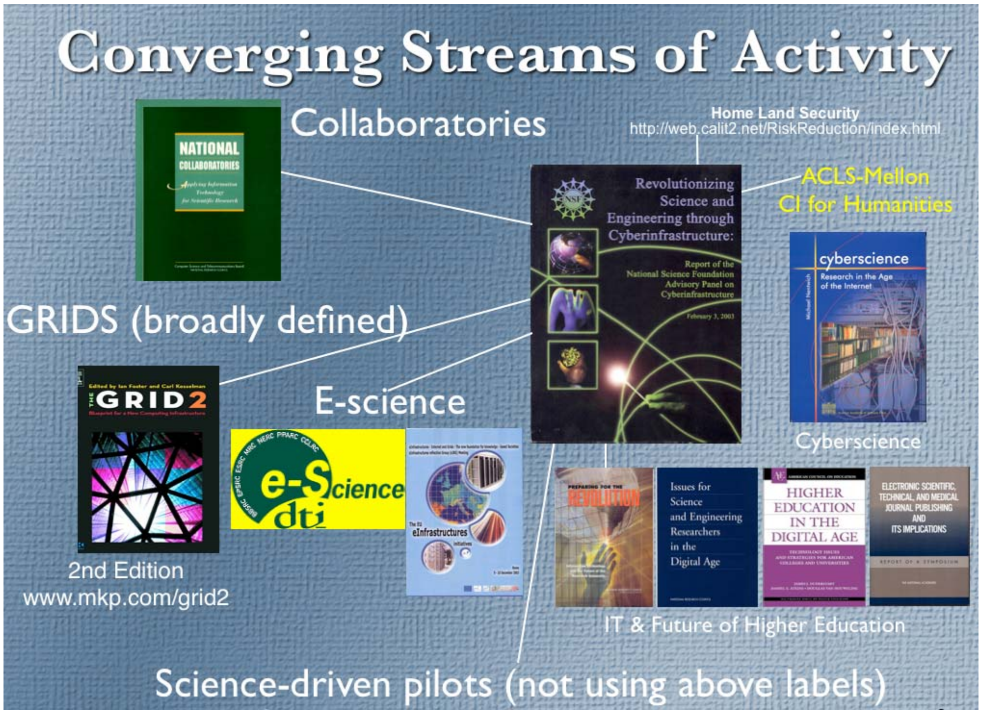
3. Shift to work flow models in scientific research that produce and share more intermediate products on the path to refereed, archival publications. The pre-print server movement is one example, and examples of preprint or e-print servers are easily found with a Google search on these terms.

4. Establishment of "institutional repositories" that more reliably capture, organize, and preserve the digital information products of a university or other knowledge-based institution. The D-Space Project, now becoming a federation of respositories, is a seminal example at www.dspace.org.

A report for a 2004 symposium at the U.S. National Academies on *Electronic Scientific, Technical, and Medical Journal Publishing and Its Implications,* discusses many of the developments, challenges and opportunities for scholarly communication in the digital age. It is available at The National Academies Press at www.nap.edu.

The Cyberinfrastructure Movement and Implications for Research Libraries

In February 2003 the U.S. National Science Foundation (NSF) issued a report from a Blue-Ribbon Advisory Panel on Cyber Infrastructure entitled *Revolutionizing Science and Engineering Through Cyberinfrastructure*. The report is available at http://www.cise.nsf.gov/sci/reports/toc.cfm. It develops a vision of comprehensive, advanced infrastructure based on information and communication technology (*cyberinfrastructure*) that serves as a platform for new organizations and methods for conducting scientific and engineering research and allied education. Names for these new environments include, for example, *collaboratory* [2-4], *grid community* [5], or *e-science community* [6]. Figure 1 illustrates some of the trends feeding into the cyberinfrastructure movement as well as some of the surrounding broader impacts on higher education. The report surveys a growing number of research fields that are creating and using cyberinfrastructure not simply to automate what they have always done, but rather to link human expertise, data, information, computational models, sensor arrays, and other specialized facilities in ways that open fundamentally new paths for research.



Converging Streams of Activity

Collaboratories

Home Land Security
http://web.calit2.net/RiskReduction/index.html

ACLS-Mellon
CI for Humanities

GRIDS (broadly defined)

E-science

2nd Edition
www.mkp.com/grid2

Cyberscience

IT & Future of Higher Education

Science-driven pilots (not using above labels)

NATIONAL COLLABORATORIES
Applying Information Technology for Scientific Research
Revolutionizing Science and Engineering through Cyberinfrastructure:
Report of the National Science Foundation Advisory Panel on Cyberinfrastructure
February 3, 2003
cyberscience
Research in the Age of the Internet
Michael Nentwich
THE GRID 2
Blueprint for a New Computing Infrastructure
e-Science
dti
eInfrastructures
PREPARING FOR THE REVOLUTION
Issues for Science and Engineering Researchers in the Digital Age
HIGHER EDUCATION IN THE DIGITAL AGE
ELECTRONIC SCIENTIFIC, TECHNICAL, AND MEDICAL JOURNAL PUBLISHING AND ITS IMPLICATIONS

Figure 1 – Converging trends for the cyberinfrastructure movement.

The NSF Advisory Panel, based on extensive testimony from a cross section of the NSF research community, concluded that a nascent international revolution is underway. The Advisory Panel recommends bold leadership and major investment to empower it. The principal finding from the report is:

> *…that a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information and communication technology; and pulled by the expanding complexity, scope, and scale of today's problems. The capacity of this technology has crossed thresholds that now make possible a comprehensive "cyberinfrastructure" on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy. The cost of not doing this is high, both in opportunities lost and through increasing fragmentation and balkanization of the research communities.*
>
> *Such environments and organizations, enabled by cyberinfrastructure, are increasingly required to address national and global priorities such as global climate change, protecting our natural environment, applying genomics-proteomics to human health, maintaining national security, mastering the world of nanotechnology, and predicting and protecting against natural and human disasters; as well as to address some of our most fundamental intellectual questions such as the formation of the universe and the fundamental character of matter.*

The "cyberinfrastructure movement" in the U.S." is complemented by similar activities in other countries under names such as e-science, e-infrastructure, and cyberscience. A growing number of visionaries now see this nascent revolution in research in the science and engineering as a harbinger for other fields and ultimately the entire academic enterprise. The American Council of Learned Societies (ACLS) has, for example, initiated a committee to explore the implications of cyberinfrastructure for the humanities and social sciences. The committee, chaired by Professor John Unsworth (unsworth@uiuc.edu) at the University of Illinois is expected to report out in January 2005.

The NSF cyberinfrastructure panel identified the functional stack shown in Figure 2. Cyberinfrastructure is based on networking, operating systems, and middleware to provide the generic capabilities for management, transport, and federation of systems and services (tools) described in the five columns. A community-specific, customized knowledge environment can ideally be created efficiently and effectively using facilities, tools, and toolkits provided at the cyberinfrastructure layer to federate the appropriate resources.

This model assume significant effort to capture and benefit from commonalities across science and engineering disciplines and appropriate levels of coordination and sharing of facilities and expertise to minimize duplication of effort, inefficiency, and excess cost. To achieve advanced capability the model also assumes real collaboration between domain scientists and engineers, computer scientists and engineers, social scientists to help understand the human factors and social-cultural issues.

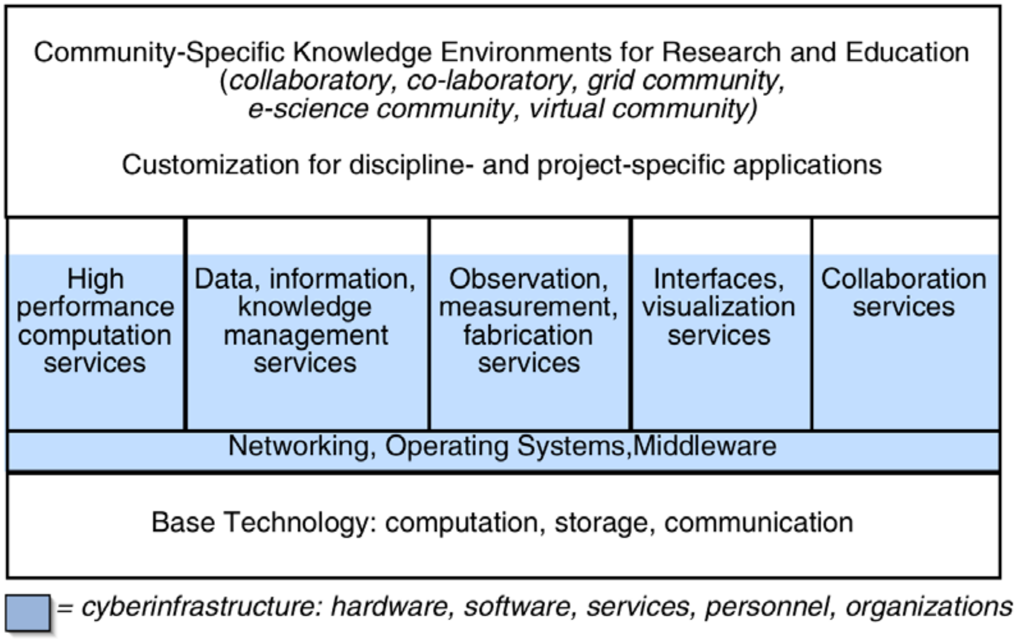

Community-Specific Knowledge Environments for Research and Education
(collaboratory, co-laboratory, grid community,
e-science community, virtual community)
Customization for discipline- and project-specific applications
High performance computation services
Data, information, knowledge management services
Observation, measurement, fabrication services
Interfaces, visualization services
Collaboration services
Networking, Operating Systems,Middleware
Base Technology: computation, storage, communication
= cyberinfrastructure: hardware, software, services, personnel, organizations

Figure 2 - Integrated cyberinfrastructure services to enable new knowledge environments for research and education.

Big ideas from the report relevant to this meeting include the following:

1. Global cyberinfrastructure can become a platform for routine, effective distance-independent activities of knowledge communities. (Goal is not to eliminate same time and place collaboration, but rather to augment it.)

2. World-scale collaborative teams can be common place.

3. Cyberinfrastructure offers new options for what is done, how it is done, and who participates.

4. The Digital Library R&D community has made large contributions to creating this opportunity and now has the opportunity and responsibility to make it real.

The research library community has the opportunity to play a major role in the creation and provision of a greatly expanded set of *data, information, and knowledge management services* (next to leftmost column of Figure 2.) By "expanded" I mean expanded in services, in diversity of the types of material (new digital genres, multimedia forms), expansion in the size of the collections, and expansion in the expectation for long-term access to digital objects. Increasingly research communities will expect these digital services and will look for other sources for them if the research library community does not provide them.

Examples of services included in this block are as follows:

1. online access to complete coverage of the credentialed, archival literature of the relevant to various research communities;

2. stewardship and curation services for enormous collections of scientific data;

3. digital repositories to provide stewardship and access to instructional material and the intermediate products (pre-publication products) of research activities;

4. leadership in digitizing, organizing, and curating unique special collections belonging to the library and providing access to them for the greater academic community (if not general public);

5. providing a host of customization/personalization services to communities and individuals including current awareness information services and community-specific information portals ("virtual special libraries") for individuals, projects and/or research organizations.

Research libraries will in general need to provide these services as part of cooperative, international federations with other libraries and information service providers. Libraries will be judged less and less by what they own and control, and more and more by what they provide (directly or indirectly), often in federation with others. They need to explore new models of cooperation to create common federated collections and functional cooperation with other organizations in their day-to-day activities. Occasional meetings to talk about common problems or even to create buying leverage with publishers will not be enough. The research communities want a comprehensive collection of information and data together with a wide range of services that will often transcend a specific library.

To serve the emerging cyber-infrastructure-enabled, often virtual, research communities spanning multiple universities, research libraries will need to build and operate their services on a compatible "middleware" (software) layer that enables distributed content (data and metadata) and services at various institutions to be integrated together seamlessly from the end users perspective. Middleware also provides directory services as well the "trust layer" that handles authentication and authorization. The NSF Middleware Initiative (www.nsf-middleware.org/) is coordinating R&D activities in this area including support for Globus, Shibboleth, and their convergence into a common, open source middleware layer.

In addition, the Andrew M. Mellon Foundation in the U.S. is sponsoring and coordinating open source middleware activities aimed and providing cost-effective, scalable, and interoperable course and project management systems that work within and across academic institutions. A prime example is the Sakai Project www.sakaiproject.org, a $6.8M *community source* software development project founded by The University of Michigan, Indiana University, MIT, Stanford, the uPortal Consortium, and the Open Knowledge Initiative (OKI). The project is producing open source Collaboration and Learning Environment (CLE) software with the first release in July 2004. The Sakai Educational Partners' Program (SEPP) extends this community source project to other academic institutions around the world.

The NSF Cyberinfrastructure Panel heard much testimony about the need for more systemic and sustained preservation and curation of scientific data. There are several factors driving this need.

1. The enormous and growing power of supercomputers in pipelines, clusters, and grids, is enabling much more comprehensive and accurate simulation of natural phenomena (the atmosphere, the cosmos, the oceans, entire ecologies, etc.). This requires the combined expertise of multidisciplinary specialists, their computational models, *and* their data. Data interchange, interoperability, and re-use by others than the originator is a growing need.

2. The development and refinement of data mining techniques is enabling analysis of large amounts of data in order to extract new kinds of useful information, in problem areas other than the original for which the data was gathered. Datasets that have "done their job" may still later reveal new useful relationships.

3. Developments in smart sensors and arrays are leading to unprecedented ability to observe and measure the physical world. The volume and complexity of data sets is increasing.

4. As more science becomes data and computationally intensive there is the need for more robust organizations to create and refine the associated metadata, to provide mechanisms for review and replication of results.

The United Kingdom e-science activities are in a leadership position in establishing requirements and approaches for scientific data curation and provision. See, for example, the *e-Science Curation Report* at www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf and a *Report of a Digital Data Curation Taskforce* at http://www.jisc.ac.uk/uploaded_documents/ CurationTaskForceFinal1.pdf.

All of this brings new research challenges to the technologists and opportunities for service to research libraries as federated stewards of information. It also implies the need for informatics professionals who are fluent it both information management principles and the substance of discipline they are serving.

More on the topic of "Keeping Academic Libraries at the Center of the University" in the digital age is available at http://www.hti.umich.edu/cgi/t/text/text-idx?type=simple;c=spobooks;cc=spobooks;sid=75bb88ce6fcf8c816736a4857858ba0c;rgn=div1;q1=atkins;view=text;subview=short;sort=occur;idno=bbv9812.0001.001;node=bbv9812.0001.001%3A7.

The panel goes on to recommend that the NSF seek significant new funding rising in a few years to a US$ 1 billion per year and assume the leadership of an Advanced Cyberinfrastructure Program (ACP) with close coordination with other U.S. and international R&D agencies. A central goal is to define and build cyberinfrastructure that facilitates the development of new applications, allows applications to interoperate across institutions and disciplines, ensures that data and software acquired at great expense are preserved and easily available, and empowers enhanced collaboration over distance, time, and disciplines. The individual disciplines must take the lead in defining specialized software and hardware environments for their fields based on common cyberinfrastructure, but in a way that encourages them to give back results for the general good of the research enterprise. Achieving this vision will challenge fundamental understanding of computer and information science and engineering as well as parts of social science, and it will motivate and drive basic research in these areas.

For the past year the NSF has been working to digest and act upon the visions and recommendation from the cyberinfrastructure panel report. One framework under consideration includes the following complementary activities:

1. science and engineering research frontier projects using advanced cyberinfrastructure;

2. development of integrating architectures that support discipline-specific applications using a common, reconfigurable set of open source tools, technologies and services;

3. deployment of foundational cyberinfrastructure including backbone networks, widely shared compute and storage facilities; education and workforce development activities; a portfolio of activities aimed at yielding new knowledge on the science of cyberinfrastructure, including its human and social dimensions and rigorous evaluation and assessment activities;

4. enabling research investments that will create new information technology tools and resources to enrich cyberinfrastructure for the foreseeable future.

Of particular interest to the digital library communities are potential core activities in the area of federated data archives and digital libraries that include 1) comprehensive NSF-wide planning to explore the efficacy of creating a national databank of federated data archives residing in different locations and belonging to multiple domains, individuals and organizations; 2) assessment of interests by specific research communities in establishing and supporting digital libraries; and 3) identifying organizational and economic models that support centralized and/or distributed data archives and digital library investments.

Although complex, challenging, and unfinished, the digital library movement has now become a part of an even larger integrated vision for the impact of information technology. It is now an enabler of the grander concept of *cyberinfrastructure-enabled knowledge communities*. The international community of research libraries and digital library researchers are critical to achieving this next wave of opportunity for global scale knowledge communities devoted to knowledge creation, dissemination, and application for human good.

End