



ARL Bimonthly Report 229

August 2003

Opportunities for Research Libraries in the NSF Cyberinfrastructure Program

by David G. Messerschmitt, Roger A. Strauch Professor of Electrical Engineering and Computer Sciences and Acting Dean, School of Information Management and Systems, University of California, Berkeley, and Member of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure

A National Science Foundation (NSF) Blue-Ribbon Advisory Panel recently recommended a new Advanced Cyberinfrastructure Program (ACP).¹ The program offers a chance to reformulate many processes of scientific investigation around the unique opportunities of information technology (IT), and for libraries to contribute to scholarly activity in science and engineering research in new ways. Research libraries house core competencies and expertise highly relevant to ACP and its challenges. However, the natural organization of scientific repositories around disciplinary needs presents challenges to the institutionally based organization predominant in the research library community.

Digital Science

Science and engineering research has long emphasized a give-and-take between theoretical and experimental methodologies, increasingly supported and supplemented by computational modeling (extending theory into new domains), data analysis (extending experimentation), and their combination. The NSF panel observed that research activities based on information technologies have reached a scale and importance that warrants giving them status as a third leg of scientific research methodology, which I term here *digital science*. An increasing number of scientists engage predominantly in digital science, as others engage in theory and experimentation.

The U.S. National Virtual Observatory (NVO) <<http://www.us-vo.org/>> illustrates the growing importance of digital science. Without constructing a new telescope, but simply creating a large repository of observational data and a set of tools for accessing and manipulating this data, astronomers have created what they call "the world's best telescope."² This non-telescope is expected to yield major new discoveries by aggregating and manipulating, for each small patch of sky, data collected at many different times by many different telescopes at many different wavelengths. The NVO also levels the playing field, opening up opportunities for major discoveries from scientists (and amateurs) in all corners of the world.

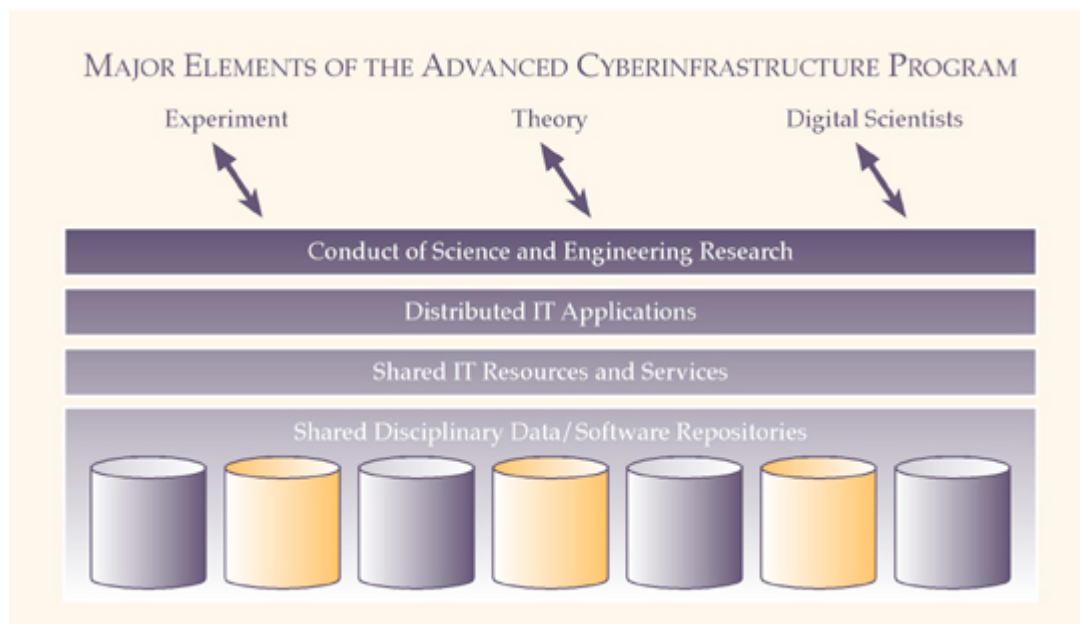
A second example illustrates other possibilities and challenges presented by digital science. The Advanced National Seismic System (ANSS) <<http://www.anss.org/>> will consolidate and interconnect 16 regional seismic monitoring networks in the U.S. into a single (although incomplete) national network.³ In contrast to astronomical observatories, it is unlikely that more than one seismic sensor will monitor any given location. However, such a network will capture a vast number of seismic events, both natural (earthquakes, volcanic eruptions, landslides, etc.) and man-made (including terrorist-originating). Real-time processing of this data may provide sufficient advance warning to shut down critical and vulnerable facilities (e.g., gas mains, transportation systems, and nuclear plants) and will direct emergency services' responses. After the fact, this information is a primary resource for geophysical scientific investigation. Future generations of investigators may make new discoveries based on mining the totality of the collected data by comparing events across geography and time.

Together these examples illustrate the benefits of integrating data acquisition, processing, storage, and access. They also illustrate the critical roles of both data organization and preservation. Digital science includes at least five complementary elements:

- collection of data from the physical world (using distributed sensors and instruments);
- distributed and remote access to organized repositories of such data;
- computation using theoretical models and experimental data;
- presentation of results for scientific visualization and interpretation; and
- support for collaboration among scientists.

The ACP recognizes the importance of these activities with new levels of management attention (led by NSF, but with coordinated activities in other federal agencies and internationally) and funding (estimated at \$1 billion per year by NSF, to be supplemented by other agencies).

The Advanced Cyberinfrastructure Program



Some major elements of the ACP are illustrated in the figure above. Digital science and

engineering research often involves close coordination of theory, experiment, and collaboration among digital scientists, so geographically distributed collaboration and access to geographically distributed sensor networks and instrumentation are crucial. Much of digital science is conducted by authoring (or in many cases executing existing) discipline-specific and generic software that automates data collection and capture, computational models and data analysis, visualization of the results, and collaboration. Software is a primary tool of a digital scientist, just as microscopes and telescopes and pencil and paper are tools of experimental and theoretical scientists.

Cyberinfrastructure encompasses the bottom two layers supporting these activities. First, it provides computational and communication resources, software, and services that are shared by the digital science community. Second, it provides repositories of shared data and software that can be appended, accessed, and utilized in the course of those applications.

"Resources and services" comprise both information technology and human resources. An example of a technology-based shared service is authentication and conditional access to repositories without regard to institution or nationality. One major goal is to capture in shared resources and services much of what is common among applications, and also to provide tools and services that make applications much easier to develop (so that scientists can focus more on their science). But many crucial resources and services are people-based, such as supporting users in accessing data repositories and using and developing software applications.




One issue for the panel was to define what needs to be coordinated or centralized, as opposed to what is best delegated to local groups. The major centralized activities proposed include shared supercomputers with power and capacity beyond the reach of individual institutions, shared data repository centers focused on capturing, organizing, and preserving data and software, and shared development centers for the production, integration, maintenance, and support of software tools and infrastructure.

The ACP would also support research into information technologies, new uses of IT and new organizations for scientific investigation, addressing shortcomings of the technology, and exploring ways that science and engineering research can be revolutionized through IT.

What Needs Preservation?

One of the central themes of ACP is preservation. Today the system for preserving and granting access to scientific data is informal at best. In practice much data is unavailable or eventually lost. One goal of the ACP is to insure the selective long-term preservation of this data and, beyond this, the stewardship and curation of these repositories so they are easily discoverable, identifiable, and accessible. Here "accessibility" refers to software applications and instruments as well as scientists, and access for not only reading but also (conditionally) for additions and changes. By "organized" and "identifiable," we mean consciously and conscientiously structured to make repositories more valuable to scientific investigation, such as by function, location, time, etc., and annotated in ways that make repositories searchable and documented in machine-readable form. By "preserved," we mean available in this manner far into the future (centuries and millennia).

WHAT NEEDS PRESERVATION?

	Information	Logic	Presentation
			
What it is:	Data, Content, Metadata	Logic, Processes, Algorithms	Interaction, Visualization
Targets of Preservation:	Data, Metadata	Software: Models, Simulations	Outcomes, Visualization

Information (represented by data and its descriptive and structural metadata) is the most obvious target of preservation, but shown above are other preservation needs. In the course of scientific investigation, the logic, processes, and algorithms are documented not only by scholarly papers (themselves a target of preservation) but also the software that realizes the models, simulations, and data analysis. This software should be selectively preserved for critical analysis *and* for its future reuse, modification, and execution so that others can reproduce, build upon, and extend outcomes. In addition, the results of major computations (especially where the software that generated them is not preserved) should be selectively preserved for future critical analysis and reuse.

Software has a dual role, as a human- and machine-readable information artifact and as a behavioral artifact resulting from its execution.⁴ The preservation of software as a behavioral artifact requires technical breakthroughs because, absent special measures, today's software will certainly not be executable in future computing environments. This is a major technical and operational challenge, considerably more challenging than data preservation.

In fact, digital science repositories cannot be cleanly separated into passive information artifacts (e.g., "documents" and "data") and behavioral artifacts (e.g., software). This distinction is rapidly blurring, as access to data is increasingly intermediated by various behavioral software-mediated functions. A number of examples illustrate this point:

- Documents, including scholarly publishing and communication, increasingly incorporate "active content," such as audio or video or animations.
- Digital rights management incorporates techniques like watermarking (to identify the origin) and encryption (to enforce conditional access) that require software mediation.
- Modern database management systems do not make "raw" data available, but access is software-mediated through query languages such as SQL.
- Modern object-oriented programming methodologies prohibit direct access to data, requiring that it be intermediated by software procedures.⁵
- Audio or video need not be accompanied by detailed formatting standards if software

is available to translate from whatever representation is chosen to (uncompressed and unencrypted) standards for playback or display.

It is persuasive that raw data in isolation should rarely be the exclusive target of preservation--structural rules, organizational paradigms, and contextual information (all represented as supporting metadata) must be preserved as well. For scientific data, there are thus two distinct forms of metadata:

- **Structural and organizational metadata** exposes (in machine-understandable form) the data structures and semantics necessary to interact dynamically with the data. This could be the subject of direct standardization, but the trend is to avoid the inflexibility of this approach through software intermediation, and to define descriptive languages for discovery and use of these abstracted representations.⁶ This is consistent with the observation that much scientific data will represent active content, as for example animations in the visualization of the results of scientific modeling or historical evolution of observational parameters.
- **Descriptive metadata** captures the relevant context of the scientific data--what experimentation or modeling it was based on, when and where it was captured, and a host of similar information vital to future investigators.

I assert, therefore, that future digital science data and software preservation targets are not separate, but should be assumed from the beginning to be largely inseparable. In the next section, we refer to "preservation artifacts" to mean both data and software and (frequently) their combination.⁷ The software preservation issue assumes far greater significance than might first appear.

Reformulating the Processes of Scientific Investigation

In any domain of application, the first use of IT is invariably the automation of existing processes, but historical experience suggests that major benefits follow when processes are reconsidered and reformulated in light of the unique characteristics and capabilities of the technology. Digital science is a major opportunity for reformulating the processes of scientific investigation, since capture, preservation, and access can in many instances be designed largely from scratch in an IT-rich environment.

Federated and networked repositories render place largely irrelevant. The physical storage of even logically federated repositories need not be centralized--a repository with the appearance of centralization can be composed from geographically separated sub-repositories.⁸ A globally accessible sub-repository or whole repository need be created and managed only once, in one place.

The management of these repositories can be divided into four basic functions:

- **Data stewardship**⁹--This includes provisioning and operations of facilities, acquiring, installing, operating, and maintaining the physical storage/processing and networking infrastructure. It also includes backup, replication, and mirroring operations on data to ensure its integrity and long-term preservation. This is the primary defense against loss due to deterioration of physical media, natural disaster, or sabotage. This function should be transparent to (and strongly separated from) any knowledge of the structure or semantics of the data being managed.
- **Content curation**¹⁰--This includes the logical organization of repositories, as well as the

definition and maintenance of metadata standards (both structural and semantic), to ensure that preservation artifacts can be located, accessed, and interpreted as needed, with tools to support both machine and human search, navigation, and access. As these repositories will become truly gigantic over the millennia, this is crucial to avoiding a debilitating "needle in a haystack" problem.

- **Origination services**--Individual users (individuals or groups and their instruments, sensors, and machines) create, structure, and attach metadata, and store scientific data and software during the course of scientific investigations in accordance with logical organizational standards or guidelines established by content curators. An important user service function aids originators of digital artifacts to condition them for the repository, or does this conditioning on their behalf (a traditional "publisher" function).
- **Access services**--Users (and their instruments and machines) access data without changing it during the conduct of new scientific and engineering investigations. An important user-service function aids consumers of digital artifacts (a traditional "library" or "museum" function), including finding what they need, linking it to their own instruments and machines, and configuring, executing, and using any associated software.

All four functions have an important human-resource element. Staff performing data stewardship must be physically co-resident with the storage facilities, but there is no similar requirement for other functions. Thus, modern networking de-emphasizes the role of place--the physical and logical mapping of data can be quite different, and distributed repositories can appear to be organized in any manner we please through appropriate federation.¹¹ This admits essentially complete freedom to define a human organizational structure and geographical dispersal of each of these roles as appropriate. In terms of efficiency and effectiveness, geographic proximity is hugely significant for human organizations and largely irrelevant for storage and processing facilities.

Preservation and User-Service Roles

Clearly there is an institutional role in ACP that parallels (if not replicates) functions traditionally performed by publishers as well as libraries and museums. Individual research grants don't fund the long-term preservation and access services for digital artifacts gathered or developed at considerable expense, nor are investigators themselves invariably prepared or motivated to carry this out, especially over time frames extending beyond a career or lifetime. As the NVO illustrates, there will be direct scientific returns for centralized and well organized repositories, as opposed to a proliferation of project-based repositories. It is unlikely that commercial publishers will find this an attractive opportunity, given the relatively infrequent (but high-value) uses (many far in the future). It is therefore envisioned that NSF (and other agencies) would fund the organization, preservation, and user service roles of the ACP, centralized and separate from individual investigative grants. They would likely be performed by non-profit or commercial organizations under contract to NSF.

Academic research libraries have focused on serving their respective institutions, notably also cooperating with and sharing resources with other institutions. A few libraries serve a discipline rather than an institution (an example is the U.S. National Library of Medicine <<http://www.nlm.nih.gov/>>). Much intrinsic value of the repositories in the ACP will follow from their discipline-based (as opposed to institutional-based) organization, as illustrated by the NVO example. Both the organization of and access to preservation artifacts should be transparent across both institutional and international boundaries. This is a mismatch with the current emphasis of research libraries on serving predominantly users

within their own institutions across all disciplines, but certainly does not preclude libraries' contribution and participation, especially in light of the organizational freedoms afforded by emerging technologies and the prospects for specific public funding for a broader service role in the ACP.

There are other reasons that the ACP should not be viewed as a traditional institutional responsibility, and digital science repositories should not be thought of as a collection or federation of "institutional repositories," although such repositories can play significant roles in filling gaps, or making up for temporary shortcomings, or providing for non-scientific disciplines.¹² Complete coverage of scientific disciplines would not be assured if this depended on the long-term guaranteed participation and assumed budgetary responsibility of every institution. The fragmentation of institutional repositories would likely be inefficient, since every institution would have to maintain the expertise to cover *all* disciplines. In fact, there are considerable economies of scale and scope--particularly in the domain knowledge required to support originators and consumers of digital artifacts--in the centralization of responsibility for individual disciplines. There would also be a problematic disconnect between responsibility (to users across the world) and budgetary sources (appropriately focused on serving internal users).

Stepping back for a moment, the most natural greenfield organization (starting from scratch) of the repository functions of the ACP would be something like the following:

- A small number of contracted centers could perform data stewardship of digital artifacts. These could be specialized commercial organizations (these already exist for similar purposes) located in low-wage regions. There is no need to organize this function along disciplinary boundaries, or to geographically constrain it. Mirroring of these sites would assure preservation in light of natural disasters and sabotage, and an appropriate networking and caching infrastructure could achieve performance goals.
- One (or at most a few) content curation centers could perform content curation for each discipline (or a few related disciplines). In addition to the development and maintenance of standards (such as for metadata) and related software (such as tools), each such center could accumulate the expertise and focus necessary to support the discipline-specific aspects of origination services and access services, and provisioning of support services for both institutions and individual users.
- In addition to possibly hosting content curation centers, individual institutions should provide origination and access-service functions to local users, emphasizing *discipline-blind* aspects while interfacing with disciplinary content curation centers on behalf of users.

The parallels between these three functions and the printer and bookbindery, publisher, and library for printed materials are evident. Binderies are content-blind, publishers tend to specialize in disciplines, and libraries predominantly serve local users across all disciplines.

While I have emphasized a disciplinary granularity for the content curation activities, this is not totally appropriate for ACP either. There will be considerable commonality across the needs of distinct disciplines, and these should be captured not only for efficiency but also to avoid balkanization that will make interdisciplinary efforts more difficult in the future. The experience in the commercial world has been that IT is often a major barrier to change (such as new products or mergers and acquisitions). In the future, digital science should insure that IT is an enabler of (and not barrier to) interdisciplinary forms of digital science--researchers in one discipline should find the incorporation of the repositories of

other disciplines into their research to be natural and well supported. This will require that ACP specifically look for pan-scientific commonalities (including standards and software), working cooperatively with and among the collection of disciplinary centers.

Aside from the organizational and granularity issues, there are other distinctions that make a difference. I already mentioned the importance of preservation of software as both information and behavioral artifact, including supporting active content and software-mediated data access. This is far more difficult and sophisticated than similar roles in the print world, and is more proximate to the museum world (especially interactive science museums). In addition, these collections will not simply be accessed for reading, but may be dynamically incorporated into distributed computing applications for appending, changing, and reading, adding entirely new responsibilities and support issues. For both reasons, user support will be crucial, and the choices made in partitioning this support structure between the user's institution and centralized groups has major implications to efficiency and effectiveness.

Returning to the greenfield opportunity, one of the supposed "benefits" of networking is purported to be disintermediation--direct interaction between originator and consumer. Actual experience in the commercial world and elsewhere has been somewhat different. While intermediary functions change, sometimes radically, they rarely disappear.¹³ Can the traditional intermediary role of the library and museum be molded to the needs of digital science? Will new institutions arise to meet these needs?

Library Contributions

I have argued that ACP repositories should be organized along disciplinary and pan-scientific rather than institutional lines. Even if you accept this, research libraries (including institutionally based ones) can play important natural roles within ACP. I have already mentioned support and interface functions for local users. Two other possible roles are as the home of disciplinary centers of content curation, and contributing to the research and design activities of the ACP. All these roles build on unique core competencies that libraries have developed and nurtured over centuries.

Research libraries (singly or in consortia) would be a natural home for content curation centers, based on funding from NSF and other agencies rather than their local institutions. These centers would design the organizational paradigms and detailed standards for structural and descriptive metadata, provide content origination and access services to scientists worldwide (or at least nationwide), and provide local institutionally based user support. Libraries obviously already bring competency and experience to these functions, although they would doubtless have to build disciplinary and software development and support expertise far beyond what they possess today, and will require collaboration with both domain and computer science communities.

Many of the issues faced by ACP are not well understood, and for this reason ACP includes a large research function, including research in information technology, in the scientific disciplines themselves, and in the social sciences. I mentioned previously the capture of commonalities (such as metadata) among disciplines, maintaining ready interoperability across disciplines, and preserving software as a behavioral artifact. But there are many, many other issues, many of them familiar to librarians. The issues surrounding metadata and software intermediation of scientific data are poorly understood. Our assertions above about the scale and scope economies of centralization are examples of larger economic

issues requiring study and quantification. Conditional and role-based access requirements for scientific data (such as the differences between pre- and post-publication access) have not been addressed. The ownership and intellectual property issues are extraordinarily challenging. There are many public and institutional policy issues, such as the responsibility of investigators obtaining public funding to add their data to repositories (many of them feel such data is personal property), and possible liability, privacy, free speech, and homeland security issues. Research libraries could themselves become involved in research underlying the design of the ACP and the digital science it supports in collaboration with domain and computer scientists.

There is a larger opportunity here for research libraries. It is clear that IT will (or at least should, unless we choose to ignore it) radically transform research and scholarly discourse, including the traditional communication and archival functions of publication and access. But what does this mean precisely? Libraries are already studying and experimenting with institutional repositories, which would fundamentally alter the relationship of libraries, scholars, and publishers.¹⁴ This is a major step, but only scratches the surface of what may be possible, and what may be beneficial or appropriate. The ACP offers an opportunity to confront these issues boldly and directly with financial support from NSF and other agencies. Although ACP addresses only the realm of scientific and engineering research, the ideas generated and lessons learned should have broader implications. The need to build a largely new platform for scholarly discourse in science and engineering affords a once-in-a-lifetime opportunity to deeply reflect upon and contribute to the future of scholarly discourse more generally.

Acknowledgements

I appreciate valuable comments on early drafts provided by Clifford Lynch and G. Jaia Barrett.

--Copyright © 2003 David G. Messerschmitt

1. Daniel E. Atkins et al., "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure," January 2003, <<http://www.cise.nsf.gov/evnt/reports/toc.htm>>. [back to text](#)
2. Bruce Schechter, "Telescopes of the World, Unite! A Cosmic Database Emerges," *New York Times*, May 20, 2003, late edition, section F, p. 1. [back to text](#)
3. Lisa M. Pinsker, "The Urban Evolution of U.S. Earthquake Monitoring," *Geotimes*, Oct. 2002, <http://www.geotimes.org/oct02/feature_anss.html>. [back to text](#)
4. David G. Messerschmitt and Clemens Szyperski, *Software Ecosystem: Understanding an Indispensable Technology and Industry* (Cambridge, Mass.: MIT Press, 2003). [back to text](#)
5. Specifically all data is encapsulated in "objects," and access to that data must be intermediated by "methods" associated with those objects. [back to text](#)
6. This is illustrated by "Interface Definition Languages" for object repository brokers <<http://java.sun.com/products/jdk/idl/>> and the "Web Services Description Language" <<http://www.w3.org/TR/wsdl>>. [back to text](#)
7. Of course, passive documents characteristic of traditional forms of scholarly communication will continue to be prevalent, both within digital science and especially in other disciplines. [back to text](#)
8. There are technical issues requiring research, but I speak here in terms of possibilities, as opposed to the current state of the technology. [back to text](#)

9. This might also be termed “physical stewardship,” in that it emphasizes the physical representations of data. [back to text](#)
10. This might also be termed “logical curation,” in that it focuses on the logical representation, structure, and organization of repositories and is transparent to their physical representation. [back to text](#)
11. Of course, this does have systems performance implications, which can be manipulated through appropriate caching mechanisms. [back to text](#)
12. Clifford A. Lynch, “Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age,” *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC*, no. 226 (February 2003): 1–7, <<http://www.arl.org/newsltr/226/ir.html>>. [back to text](#)
13. An illustration of this in the commercial world is eBay <<http://www.ebay.com/>>, which allows buyers and sellers to conduct business directly but also creates new intermediaries (eBay itself, as well as payment escrow services). [back to text](#)
14. Clifford. A. Lynch, *ibid.* [back to text](#)

☐ [Back to top](#)

Messerschmitt, David G. "Opportunities for Research Libraries in the NSF Cyberinfrastructure Program." *ARL*, no. 229 (August 2003): 1-7. <<http://www.arl.org/newsltr/229/cyber.html>>.

[Table of Contents for Issue 229](#) | [Other Current Issues Articles](#)
[Other Networked Information Articles](#) | [Other Scholarly Communication Articles](#)



[ARL Bimonthly Report Home](#) [ARL Home](#)

ARL policy is to grant blanket permission to reprint any article in the *Bimonthly Report* for educational use as long as full attribution is made. Exceptions to this policy may be noted for certain articles. This is in addition to the rights provided under sections 107 and 108 of the Copyright Act. For commercial use, a reprint request should be sent to ARL Director of Information Services, [Julia Blixrud](#).

© The Association of Research Libraries
Maintained by: [ARL Web Administrator](#) Site Design Consultant: [Chris Webster](#) Last Modified: August 25, 2003