The Future of Scholarly Communication:

Building the Infrastructure for Cyberscholarship

creates opportunitie
for new forms of research an
scholarship that are qualitativel
traditional ways o
academic publications an
Find...
research data. The widesprea
of content in digita
provides an infrastructur
Find NEW
for novel forms of research. T
support cyberscholarship, suc
content must be captured, managed
preserved in ways that ar
different fro

Report of a workshop held in Phoenix, Arizona
April 17–19, 2007

Sponsored by the National Science Foundation and the Joint Information Systems Committee

William Y. Arms
Ronald L. Larsen

September 12, 2007

The Future of Scholarly Communication:
Building the Infrastructure for Cyberscholarship

TABLE OF CONTENTS

UMC6630-1207

Acknowledgements

This workshop was funded, in part, by National Science Foundation grant number IIS-0645988 and the Joint Information Systems Committee. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

The authors gratefully acknowledge the support of the Coalition for Networked Information, which provided inestimable logistical assistance in planning, organizing, and running the workshop.

Infrastructure for Cyberscholarship

1. Summary

On April 17–19, 2007, the National Science Foundation (NSF) of the United States and the British Joint Information Systems Committee (JISC) held a workshop in Phoenix, Arizona. The objective of the workshop was to build on the findings of recent Cyberinfrastructure reports ([ACLS], [CI]) to identify opportunities and strategies for managing information created and used by researchers and scholars in the sciences, social sciences, and the humanities. Participants included representatives from Europe and the United States with affiliations in government, higher education, industry, and private foundations.

The fundamental conclusions of the workshop were:

- The widespread availability of digital content creates opportunities for new forms of research and scholarship that are qualitatively different from traditional ways of using academic publications and research data. We call this "cyberscholarship."
- The widespread availability of content in digital formats provides an infrastructure for novel forms of research. To support cyberscholarship, such content must be captured, managed, and preserved in ways that are significantly different from conventional methods.

As with other forms of infrastructure, common interests are served best by agreement on general principles that are expressed as a set of standards and approaches that, once adopted, become transparent to the user. Without such agreements, locally optimal decisions may preclude global advancement. Therefore, the workshop concluded that:

- Development of the infrastructure requires coordination at a national and international level. In Britain, JISC can provide this coordination. In the United States, there is no single agency with this mission; we recommend an interagency coordinating committee. The Federal Coordinating Council for Science, Engineering and Technology (FCCSET), which coordinated much of the U.S. government's role in developing high performance computing in the 1990s, provides a good model for the proposed Federal Coordinating Council on Cyberscholarship (FC^{3}S). International coordination also should engage organizations such as the European Strategy Forum on Research Infrastructures (ESFRI), the German research foundation DFG, and the Max Planck Digital Library.
- Development of the content infrastructure requires a blend of interdisciplinary research and development that engages scientists, technologists, and humanities scholars.

The time is right for a focused international effort to experiment, explore, and finally build the infrastructure for cyberscholarship.

- We propose a seven-year timetable for implementation of the infrastructure. The first three years will focus on developing and testing a set of prototypes, followed by implementation of coordinated systems and services.

2. Cyberscholarship

During the past 20 years, scientific and scholarly information has moved from physical media to digital forms. The first generation of digital libraries and online datasets used the new technology to mimic previous practices. We now are seeing the next phase: the emergence of new forms of scientific discovery and scholarly research that previously were impossible. These new approaches might be called "cyberscholarship."

2.1 New forms of scholarship

As an example of cyberscholarship, the National Institutes of Health has been a pioneer in building tools that seek latent patterns in the various databases, indexes, and other materials that it manages [NCBI]. Computer programs analyze vast amounts of information that could never be processed manually. This is sometimes referred to as "data-driven science." Some have described data-driven science as a new paradigm of research. This may be an overstatement, but there is no doubt that digital information is leading to new forms of scholarship.

Machine-readable content in the humanities

If we are serious about data-driven scholarship, then we must respond decisively but wisely to the imperative to collect, to curate, to preserve, and to provide access to the content that we believe, to the best of our abilities, will have enduring value. And we need to do so now.

Abby Smith, historian and cultural resources analyst and consultant

In a completely different field, Gregory Crane, a humanities researcher, recently made the simple but profound statement, "When collections get large, only the computer reads every word." A scholar can read only one document at a time, but a supercomputer can analyze millions, discovering patterns that no human could observe.

Other new forms of scholarship are based on communication. Ubiquitous networks allow people to interact in novel and productive ways. Early examples included arXiv.org, the physics preprint archive, and bio-medical databases such as the protein database. The collective development of such resources speeds up the exchange of ideas and enables scientists from around the world to become active contributors. More recently we have seen serious science being developed using collaborative tools such as wikis and blogs.

2.2 Content as infrastructure

Scholars in all disciplines long have recognized the importance of investing in physical infrastructure: the bricks and mortar of libraries, museums, and archives, as well as scientific instrumentation such as telescopes and accelerators. In the cyber age, collections of digital content and the software to interpret them have become the foundation for discovery; they have entered the realm of infrastructure.

Infrastructure for data-driven science

The National Virtual Observatory describes itself as "a new way of doing astronomy, moving from an era of observations of small, carefully selected samples of objects in one or a few wavelength bands, to the use of multi-wavelength data for millions, if not billions, of objects. Such datasets will allow researchers to discover subtle but significant patterns in statistically rich and unbiased databases, and to understand complex astrophysical systems through the comparison of data to numerical simulations."

From: www.us-vo.org

When content becomes infrastructure, there is value in investment to support it. The preservation and organization of information for new forms of scholarship enable others to discover unexpected and novel associations without having to replicate the primary data. The side bar offers an example of these capabilities at the National Virtual Observatory.

As another example, the Internet Archive has preserved a huge collection of historic Web pages, but the burden of extracting, organizing, and analyzing datasets for individual research projects has impeded its use in social science research. The Cornell Web Lab [CWL] is addressing this issue by providing an integrated laboratory environment for historical research using Web resources.

These examples are exceptions. In most fields, there is no comparable infrastructure for managing digital content. Either the primary research data is discarded after publication of the results or it is published in forms that do not support cyberscholarship.

Cyberscholarship is possible only when most of the content in a field is accessible to computer programs. The reason to focus on content as infrastructure now is that many disciplines are at a tipping point where digital content is the norm, permitting automatic analyses not previously possible. Analog forms have become the exception. Almost all text, data, images, audio, video, and software now are created in digital formats; in addition, a large proportion of the content available only in analog formats is being converted to digital forms. Almost all scientific journals are online, as are reports, working papers, conference proceedings, and datasets. At this point, projects are under way to digitize all the books in several major research libraries. Now is the time to build the infrastructure to ensure that this digital content remains accessible and usable for current and future scholars.

3. The goal

The workshop participants set the following goal:

> Ensure that all publicly funded research products and primary resources will be readily available, accessible, and usable via common infrastructure and tools through space and time, and across disciplines, stages of research, and modes of human expression.

This goal is not tied to a specific view of scholarship. The new infrastructure will enable both cyberscholarship and conventional forms of research. However, the aim is to go far beyond a system that merely replicates the traditional methods used for physical media in the digital domain.

We advocate a target date of 2015 to reach this goal. As elaborated in Section 6, this provides time for a phased development process in which a series of prototypes feeds into a stable infrastructure.

Barriers to data-driven science

For many years, therefore, we have worked to create systems that can automatically read the current chemical literature, aggregate the data, add semantics and metadata, and allow scientific hypotheses to be tested. More ambitiously, it is possible for the system to extract patterns or unusual observations from which new hypotheses might be constructed. This is reflected in our OSCAR and CrystalEye systems. ... Our thesis is that the current scientific literature, were it to be presented in semantically accessible form, contains huge amounts of undiscovered science. However, the apathy of the academic, scientific, and information communities, coupled with the indifference or even active hostility and greed of many publishers, renders literature-data-driven science still inaccessible.

Peter Murray-Rust, University of Cambridge

The remainder of Section 3 expands this goal into several components: capturing content, access for research, value-added services, and preservation for the long term. Section 4 describes challenges that have to be faced, both technical and social. Section 5 discusses organizational questions, particularly those concerned with long-term stability. Finally, Section 6 brings these themes together in an outline plan. The sidebars placed throughout this report are short excerpts from position papers submitted prior to the NSF/JISC workshop in Phoenix.

3.1 Capturing content

The first step in achieving this goal is to ensure that the products of research are captured and made available for others to use. The academic community provides strong incentives to publish the results of research. Equally strong incentives are needed to ensure that, in all disciplines, research is not considered complete until the data used to generate the results are placed in an appropriate form for future researchers.

Some disciplines have admirable traditions of publishing their data in a form that lends itself to reuse. Astrophysicists have the repository of data at the Harvard-Smithsonian Center for Astrophysics [HSCA]. Also, the Inter-University Consortium for Political and Social Research [ICPSR] has been collecting survey datasets for decades. But, as illustrated in the sidebar on this page, other disciplines are not served as well.

Funding agencies such as the NSF can have a considerable impact by establishing guidelines, norms, and incentives for publishing data. While the form of publication varies depending on the context, the general principle should be that when a project uses public funds to generate data, create software, or collect information via surveys, the project has the responsibility to make that information available to future researchers. Universities have a particular responsibility, as the public funds so much of their research.

3.2 Access for research

In a traditional library, the scholar personally browses the collection, searches through the catalog, and takes books off the shelves. With very large digital collections, the equivalent functions are performed by computer programs acting as agents for people. Researchers do not interact with the collections directly. They use computer programs to extract small parts of the collection and rarely view individual items except after automated preliminary screening.

This requires content to be organized for access by computers. Text and data must be in formats that support machine processing. Computer systems must provide application program interfaces (APIs). Display formats that are very easy for humans to read such as PDF are awkward for machine processing; mark-up languages such as XML are much more effective.

The technology of TeraGrid

"Using high-performance network connections, the TeraGrid integrates high-performance computers, data resources and tools, and high-end experimental facilities around the country. Currently, TeraGrid resources include more than 250 teraflops of computing capability and more than 30 petabytes of online and archival data storage, with rapid access and retrieval over high-performance networks. Researchers can also access more than 100 discipline-specific databases. With this combination of resources, the TeraGrid is the world's largest, most comprehensive distributed cyberinfrastructure for open scientific research."

From: www.teragrid.org

Access is more than formats and APIs, however. Cyberscholarship requires software tools, some of which are complex to develop, and it requires access that is unimpeded by complex legal or financial barriers. The limited use of data resources on the TeraGrid is indicative of the challenges and lack of incentives in making research data usable for people who are not actively part of the original research program. Success stories such as TREC for information retrieval research [Voorhees] or the Human Genome Project [HGP] have devoted substantial expertise to creating the necessary infrastructure and managing the datasets with a very clear understanding of how they fit the research practices in their fields.

High performance computing traditionally has focused on supercomputing centers where the emphasis is on computation. Cyberscholarship needs superdata centers, which combine the storage and organization of vast amounts of data with substantial computing power to analyze it. Building such centers requires investment and long-term commitment on the part of an organization or discipline. While equipment can be purchased, expertise takes longer to establish. Superdata centers and the researchers who use them will need several years before they become truly effective. These topics are well described in the recent report, *NSF's Cyberinfrastructure Vision for 21st Century Discovery* [CI].

3.3 Value-added services

Tools for cyberscholarship

Increasingly, value resides in the relationships between researchers, papers, experimental data and the ancillary supporting materials, associated dialogue from comments and reviews, updates to the original work, etc. Typically, when hypertext browsing is used to follow links manually for subject headings, thesauri, textual concepts, and categories, the user can only traverse a small portion of a large knowledge space. To manage and utilize the potentially rich and complex nodes and connections in a large knowledge system such as the distributed Web, system-aided reasoning methods would be useful to suggest relevant knowledge intelligently to the user.

As our systems grow more sophisticated, we will see applications that support not just links between authors and papers but relationships between users, data and information repositories, and communities. What is required is a mechanism to support these relationships that leads to information exchange, adaptation, and recombination – which, in itself, will constitute a new type of data repository. A new generation of information retrieval tools and applications are being designed that will support self-organizing knowledge on distributed networks driven by human interaction. This capability would allow a physicist or biochemist to collaborate with colleagues in the life sciences without having to learn an entirely new vocabulary.

Rick Luce, Emory University

In the past, scientists had to make their own tools. Access to content is of little value if every cyberscholar has to write custom software. Value-added services such as tools for searching and organizing the information extend the intellectual reach of all, from novices through experts.

Value-added services for the humanities

Four basic classes of value-added service emerge:

1) catalogue services identify the discrete objects within a collection (editions of Vergi's Aeneid, books about Vergil);

2) named entity services identify semantically significant objects embedded within collection objects (references to Vergil or the Aeneid within other documents);

3) customization and personalization (given a particular passage of the Aeneid, what would be of interest to an intermediate student of Latin vs. a professional Latinist?);

4) structured user contributions (e.g., users tell the library that a particular word in a passage of Vergil has a particular sense or plays a grammatical role in the sentence). Summarization, visualization, machine translation, and other technologies all play roles within one or more of these value-added services.

Gregory Crane, Tufts University

Many of these tools are specific to a discipline. However, as discussed in the tools for cyberscholarship side bar, there are many categories of software tools that have broad application. We need a process that encourages researchers to develop general purpose packages rather than write individual programs for each experiment.

Open-source software development has been a cost-effective approach for creating such general purpose research tools, enabling contributions from a broad base of developers throughout the research community. But open-source development does not just happen. It needs people, expertise, and a stable organization. Some packages developed in this way have subsequently evolved to become successful commercial products, such as the major statistical packages.

3.4 Curation and preservation

Curation and preservation of the ever-growing volume of digital materials poses particularly daunting challenges. In the United States, the Library of Congress' National Digital Information Infrastructure and Preservation Program [NDIIPP] is leading the efforts to address these challenges. (See the position paper by Laura Campbell of the Library of Congress.) In Britain, the Digital Curation Centre has a similar role [DCC].

The construction and curation of datasets

It may be sufficient for funding agencies and journal and book publishers to mandate that original datasets on which new publications are based be deposited and maintained in publicly accessible repositories. However, there are some fields that are thinking even more innovatively and are trying to build peer-review systems around the data so that they can be judged formally on qualities of coherence, design, consistency, reliability of access, and so on. With JISC support in the UK, scientists and professional associations in the field of meteorology have joined to establish a new kind of electronic publication called a data journal, where practitioners would submit data sets for peer review and dissemination. [Gadian] With Mellon support, in the field of nineteenth century literary studies, Jerry McGann at the University of Virginia has organized scholarly societies into a federation for the purpose of providing peer review for data in the form of online documentary editions of nineteenth-century authors. [Nowviskie] And Bernard Frischer, who is a specialist in online virtual reconstructions of archaeological sites, has received NSF support to plan a journal-like outlet that would provide peer review of virtual reconstructions [SAVE].

Don Waters, Andrew W. Mellon Foundation

These programs emphasize the long term. Less attention has been given to the short-term problem that most research data is discarded very soon after creation. The most common reason is that research is often funded through short-term grants. When the grant expires, the researcher doesn't have the resources, the incentives, and the responsibility to preserve or share the products of the work.

The sidebar above describes some current activities in the areas of data sharing and organization. These examples extend the concept of peer review to the construction of datasets, but in many cases it is sufficient for researchers to simply make their raw data available without a formal publication process. Biomedical databases such as the one for the Human Genome Project are collaborative projects where scientists routinely deposit their raw data for the benefit of others. Disciplines are different and may need to be treated differently, but clear norms backed up by appropriate incentives are needed for primary data sources to become routinely available.

4. Challenges

4.1 Variety

In developing the cyberinfrastructure for content, there are two extremes to be avoided. The first is to assume that a single approach will fit all categories of content. There is too much variety in types of data for this to be possible. Consider, for example, the differences among digitized books, astronomy data, Web pages, the genome database, census data, educational courseware, and weather observations. The forms of the information differ along almost every technical dimension: scale, structure, complexity, and so on. They also have different challenges in terms of privacy, copyright, standardization, and commercial value. Most importantly, they support very different scholarly communities. The uses that are made of them vary greatly, and they need fundamentally different data structures and computational tools.

The other extreme is to assume that every category of content needs a completely different approach and that nothing can be gained by cooperation and shared services. This viewpoint fails to recognize that apparently different categories of content may have many aspects in common. For example, there are only a limited number of system architectures for managing large quantities of information. Encoding schemes and mark-up languages are relatively standardized; the legal and economic frameworks have much in common.

Between these extremes lies a middle ground in which a relatively small set of approaches supports a wide variety of content. As yet, there is only limited understanding of how to select these approaches. Therefore, this report advocates developing a coordinated set of prototype systems, leading to implementation of a flexible operational system that would illuminate, through experience, those approaches with a greater span of influence.

4.2 Resistance to change

Transformative change requires articulation of technologies, policies, and processes in the context of human organizations. The human organizations themselves must adapt, and these adaptations can be even more difficult than developing the technologies or revising the policies and procedures.

Institutional resistance to change

What body other than the university or research center itself a) employs the researchers, b) provides subventions for press runs of fewer than 500, and c) comes up with the budget for the library to buy the subsidized but still expensive books for the researchers they employ? If funding is not the real issue—there is a lot of money in that system—could it be that we are expecting too much in the way of critical path decisions on institutional policy and intellectual property from publishers comfortable with, or desperate to preserve, a traditional model and librarians just sufficiently well funded to be comfortable with their own and publishers'/distributors' traditional roles?

Ken Hamma, J. Paul Getty Trust

Unfortunately, the academic community has a track record of resistance to new forms of scholarly communication. Leading researchers — and the professionals who support them — have mastered established forms of publication and data management. From their perspective, it is easy to see digital content as nothing more than a new representation of familiar forms: of books and journals, accessed by catalogs and indexes, constrained by conventional views of peer review and copyright. The challenge is to raise the new, alternative modes of scholarly communication to the same level of repute as more traditional forms have enjoyed, overcoming resistance to the ideas behind such projects as the preprint archives (e.g., arXiv.org), search engines, automatic information systems (e.g., Google Scholar), and community-generated information (e.g., Wikipedia).

The shortcomings of the current environment for scholarly communication are well known and evident. Journal articles include too little information to replicate an experiment. Restrictions justified by copyright, patents, trade secrets, and security, and the high costs of access, add up to a situation that is far from optimal. Yet this sub-optimal system has vigorous supporters, many of whom benefit from its idiosyncrasies. For example, the high cost of access benefits people who belong to the wealthy organizations that can afford that access. Journal profits subsidize academic societies. Universities use publication patterns as an approximate measure of excellence.

Younger scholars who grew up with the Web are less likely to be restrained by the habits of the past. Often — but not always — they are early adopters of innovations such as Web search engines, Google Scholar, Wikipedia, and blog science. Yet they come under intense pressure early in their careers to conform to the publication norms of the past.

4.3 Scale and complexity

It is essential that the academic community has the technical expertise to control its own destiny. Cyberscholarship is technically difficult. There are deep challenges in managing the scale and complexity of information. While scholarship always has depended on data, we recently have seen an enormous increase in the volume and complexity of data.

With disappointingly few exceptions, the academic community has largely abandoned large-scale digital libraries to other organizations, primarily commercial companies. This problem is particularly severe with the Web. Organizations that run large-scale production services such as Google, Yahoo, and the Internet Archive have developed expertise that enables them to introduce novel technology and new services at low marginal costs. The academic community has been left behind in this development process, which is unfortunate. Universities are isolated from a vibrant area of research and innovation and forced into alliances with commercial companies, which may be regretted in the long term. Where the academic community has been successful in managing huge amounts of data, the leadership recently has come from either supercomputing centers or individual disciplines that are accustomed to very large datasets.

Meanwhile, there is a fertile but largely independent line of research by the digital library community into the management of smaller but still significant collections of highly complex information, often supported by rich metadata. Fedora is the best known example of this work [FEDORA].

Commercial partnerships can be extremely helpful, but more is needed. The NSF's program in Cyberinfrastructure provides an excellent place for the various independent lines of development to meet, share expertise, and develop common processes and technology, where appropriate.

4.4 Access

The persistence of proprietary interests and the unremitting lack of openness

The quest for open access is not a matter of content communism. Without open access, the Web is bound to replicate the insular structure of information in the print world. Lack of open access constitutes one of the main obstacles to the full exploitation of the innovative potential of the Web for research and scholarship.

In the sciences, open access refers to publications as well as their hinterland of data, simulations, software, etc. In the humanities, open access should similarly refer not only to publications but also to testimonies of cultural heritage, to historical works of art, literature, and science, to image, film and sound collections, to statistical data, etc. There is, however, a major difference between the humanities and the sciences: while in science the raw data constituting the hinterland of research are typically produced and kept by the same people who write the publications, authors in the humanities are as a rule not those who collect and preserve cultural heritage or provide access to it.

Research institutions and cultural heritage institutions tend to perceive their interests in different ways. While most research institutions see their mission only half accomplished if they are not employing the optimal tools for granting access to their output, holders of cultural heritage tend to conceive electronic reproductions not as a new way of preserving and sharing the memory of mankind but merely as a new source of revenue they can use in compensation of dwindling public funds to fulfill their traditional function. The duplication of the world of cultural heritage in the electronic medium has actually triggered a gold rush. It motivates museums, libraries, and archives as well as private companies to stake out proprietary claims in this new territory. They tend to speculate on the quick commercial exploitation of resources rather than fostering their integration into a global representation of human knowledge. Unfortunately, they are assisted in this exclusive policy by those humanists who are all too willing to compromise open access in the interest of their own exclusive, academic niches.

The arduous goal of open access in the humanities can only be achieved when public institutions no longer invest in endeavors with proprietary output.

Malcolm Hyman and Jürgen Renn
Max Planck Institute for the History of Science

Developments of cyberscholarship are hampered by the profusion of intellectual property rights and business practices that restrict access to information. Some of these restrictions are necessary, particularly those that protect privacy or trade-secret information, but there is less justification for others. Science and scholarship have a privileged position in society. Governments fund scientific research and national libraries. Universities are supported by taxpayers and enhanced by the generous tax benefits given to not-for-profit organizations. A fundamental goal of the new infrastructure is to make the results of these efforts benefit the society that supports them.

Engineering challenges of stability

When it comes to cost-effective, long-term preservation of large datasets, everywhere we look there are problems crying out for research and innovation across a broad range of disciplines. We don't know how to specify, measure, model, and predict system behavior under a realistic set of threats. We lack data on the threats, and there are significant obstacles to obtaining it. Even if it were possible to characterize system performance and cost at a point in time, the highly dynamic nature of the environment makes taking rational long-term decisions hard. The structure of the funding mechanisms themselves contributes to this difficulty. The relationship between the funders and the services actually preserving the data has significant practical and legal difficulties.

This should not be taken as a counsel of despair. Clearly, these databases exist and are in daily use by their research communities. Some prudent replication, together with good system administration practices, keep them reasonably safe in the short term. And this buys time for the underlying problems to be addressed. The risk is rather one of complacency leading to inaction.

David Rosenthal, Stanford University

Seeking patterns across heterogeneous collections is impossible without access. There is evidence that descriptions of research are more widely read if they are openly accessible [Lawrence]. This benefits the authors, future researchers who build on that work, and the agencies that fund the research. We strongly support the movement in both Europe and the United States to require open access to all papers that describe research supported by grants from the taxpayers. We also support efforts to require that the data products of such research be made available for future research in convenient formats, subject only to the constraints of privacy and the appropriate protection of trade secrets and classified information.

This is another area where the younger scholars who have grown up with the Web have different expectations. They are reluctant to use information that is not online. They are accustomed to new models of content dissemination on the Web, which are characterized by rapid and easy access, sharing, and "swarming" around content. They grow impatient with clumsy and outmoded ways of protecting/locking down content, and regard recent interpretations of copyright law by certain media companies as a perversion of the very principles of respect for intellectual and artistic creations that copyright law is supposed to foster.

4.5 Stability

The content infrastructure requires stability of organizations, technology, and data. The infrastructure is costly to create and to support for the long term. Numerous groups have studied the questions of sustainability, e.g., the Sustainability Standing Committee of the National Science Digital Library [NSDLSSC], without coming to any general conclusions.

The engineering problems of sustaining complex information over long periods are formidable, but they are not overwhelming. They are recognized as serious subjects for research. Several good computing groups are interested in them. We encourage funding agencies such as the NSF to support their efforts.

The economic and organizational problems are much more challenging. As David Rosenthal discusses on his blog [DSHR], the San Diego Supercomputer Center reports that its cost of sustaining one disk plus three tape replicas is $3K/TB/yr [SDSC]. The rapidly decreasing cost of disk media contributes only a small part to this, so the overall cost is not expected to drop rapidly. Clearly, there are data that will not be preserved for the long term.

Although many organizations have a long-term stake in the content infrastructure, none has made a commitment for the long-term yet. In the United States, the research funding model used by the NSF and other agencies avoids long-term commitments. Universities often support the current infrastructure through their library budgets, some of which are substantial. However, they have difficulty redeploying their funds away from services that support their own institution to underwrite shared services that support the academic community as a whole.

5. Organizational questions

5.1 Local optimization, collaboration, competition, and trust

Before describing the path ahead, we should consider whether the goals of cyberscholarship might be met by letting existing organizations develop their individual strategies. Could we assume that new activities will develop spontaneously and solve all the problems?

This seems unlikely. If universities are to take the lead in changing the information infrastructure, they will need to work together; yet universities are fundamentally competitors. They compete for faculty, students, grants, and publicity. While researchers in different universities often collaborate on a one-to-one basis, almost all large-scale collaborations between institutions depend on external funds and rarely continue after the funding expires. Competition among universities, particularly in the humanities, often is reflected in competition among their libraries.

If universities have difficulty generating the necessary change, perhaps the entrepreneurial spirit of the private sector will fill the gaps. Undoubtedly, the private sector has been — and will continue to be — an important source of such innovative services as LexisNexis, Science Citation Index, and Google Scholar. But the demand for ever-increasing profits also can be a barrier. For example, the financial model of journal publishing uses copyright law to restrict access to information. Google's search engine is built on a concept of popularity, based on its own proprietary data collection, which may inhibit the discovery of unusual or unrecognized information. Some areas that are vitally important to science and scholarship appear to have little potential to generate revenue.

The conclusion is that local optimization is not enough. Left to themselves, the various organizations will create a fragmented, suboptimal world and may even create serious diseconomies that could harm the scientific and cultural future.

Interdisciplinary collaboration within institutions faces challenges because of the discipline structure of universities, but here the signs are more hopeful. Several of the position papers describe work in which humanists, computer scientists, and professionals are collaborating across disciplines. The cultural differences among sciences, social sciences, and humanities should not be minimized, but, when the intellectual challenges are genuine, interdisciplinary research thrives.

Finally, there is the question of trust. With conventional media, there are many copies of each work, and each institution looks after its own copy. With the content infrastructure, no institution will store more than a fraction of the information that it uses; researchers will have to trust each other. The position paper by Fran Berman of the San Diego Supercomputing Center describes one approach to this challenge [Berman].

5.2 Coordination

Since local optimization is not enough, some form of coordinated leadership is needed.

In the United States, it is likely that a federal agency or a group of agencies will support key parts of the cyberinfrastructure, but that is far from assured. At present, no agency has the mandate to take the lead in this effort. Several agencies have missions that address some of the goals of cyberscholarship. They include the National Science Foundation, the National Institutes of Health, the Library of Congress and the other national libraries, the Institute of Museum and Library Services, and mission agencies such as NASA and NOAA. Private foundations, notably the Andrew W. Mellon Foundation, have made great efforts to fill some of the gaps. Universities and not-for-profit organizations such as the Internet Archive and the J. Paul Getty Trust are making important contributions. But these organizations, working independently, have been unable to create a coordinated approach.

We recommend that a federal coordinating committee be formed to coordinate the work of the federal agencies in this area. We could use as a model the Federal Coordinating Council for Science, Engineering and Technology (FCCSET) of federal agencies that coordinated much of the U.S. government's role in high performance computing in the 1990s.

Institutional repositories in the Netherlands

"Our mission of disseminating knowledge is only half complete if the information is not made widely and readily available to society. New possibilities of knowledge dissemination not only through the classical form but also and increasingly through the open access paradigm via the Internet have to be supported. We define open access as a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community."

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. October 2003. Signed in the Netherlands by KNAW, NWO, SURF, and the universities of Groningen, Leiden, Amsterdam, Delft, Eindhoven, Wageningen, and Utrecht.

It is of paramount importance in an internationally competitive knowledge economy that the knowledge that is created finds its way to the research community, to society and to private enterprise. Providing broad access is a crucial requirement, and this can only be achieved in a communal approach. [The SURF Strategic Plan] shows the commitment of all universities, universities of applied sciences, the Royal Netherlands Academy of Arts and Sciences (KNAW), the Netherlands Organisation for Scientific Research (NWO), The National Library of the Netherlands (KB), and TNO to a strong effort from SURF in the field of scholarly and knowledge communication.

Bas Cordewener, SURF Foundation

In Europe, there is a tradition of centralized planning for research and central management of universities. In the Netherlands, SURF already has demonstrated what vigorous central leadership can achieve when linked to a strong set of goals. The DARIAH initiative, organized primarily by JISC, DANS (NL), MPDL (Germany), and CNRS (France), aims to provide infrastructure supporting access to all surviving humanities and cultural heritage information for European countries. The Joint Information Systems Committee (JISC) is the natural organization to lead the British effort in cyberscholarship.

Cyberscholarship is inherently international. Strong international collaboration is needed in developing the cyberinfrastructure. This does not imply that a single standardized approach should be followed. Indeed, for reasons discussed in Section 4.1, we advocate a number of differing approaches. But the selection of differing approaches should be based on the categories of content and the research that they support, not on the perceived inconvenience of working with foreign partners.

As a first step, it was noteworthy that the NSF and JISC sponsored this planning workshop jointly, with attendees from several European organizations. The United States and the countries of the European Union have much to gain from working together.

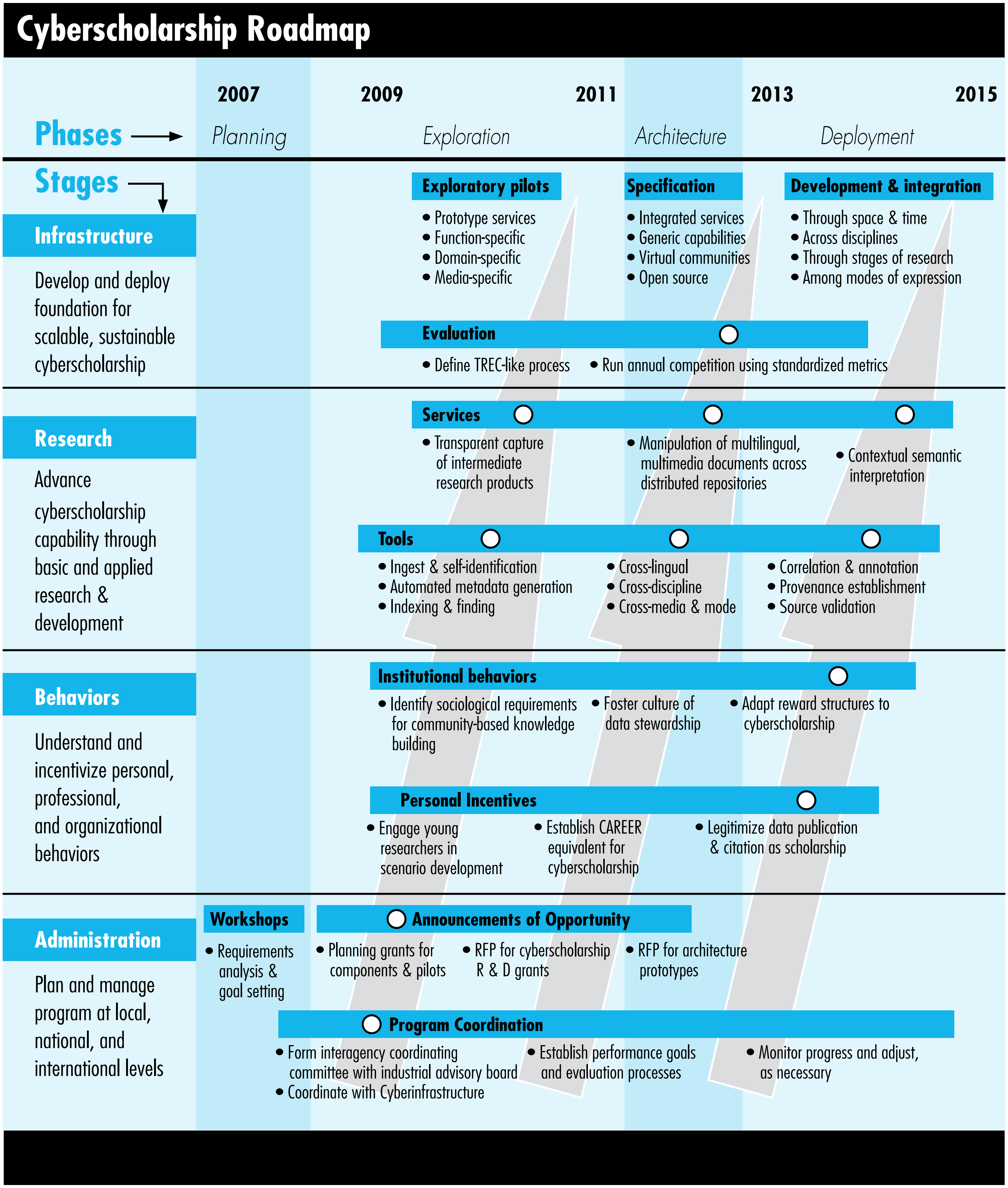
6. A plan

6.1 Research and implementation

We propose a seven-year target for the implementation of the infrastructure for cyberscholarship.

The goal of establishing an infrastructure for cyberscholarship by 2015 is aggressive but achievable when coordinated with other initiatives in the United States, Britain, and elsewhere. A three-phase program is proposed over seven years: a three-year research prototype phase that explores several competing alternatives, a one-year architecture specification phase that integrates the best ideas from the prototypes, followed by a three-year research and implementation phase in which content infrastructure is deployed and research on value-added services continues. Throughout the seven years, an evaluation component will provide the appropriate focus on measurable capability across comparable services. A "roadmap" for the program is suggested in the following figure.

Cyberscholarship Roadmap



2007 2009 2011 2013 2015

Phases → Planning | Exploration | Architecture | Deployment

Stages

Infrastructure

Develop and deploy foundation for scalable, sustainable cyberscholarship

Exploratory pilots

- Prototype services
- Function-specific
- Domain-specific
- Media-specific

Specification

- Integrated services
- Generic capabilities
- Virtual communities
- Open source

Development & integration

- Through space & time
- Across disciplines
- Through stages of research
- Among modes of expression

Evaluation ○

- Define TREC-like process
- Run annual competition using standardized metrics

Research

Advance cyberscholarship capability through basic and applied research & development

Services ○ ○ ○

- Transparent capture of intermediate research products
- Manipulation of multilingual, multimedia documents across distributed repositories
- Contextual semantic interpretation

Tools ○ ○ ○

- Ingest & self-identification
- Automated metadata generation
- Indexing & finding
- Cross-lingual
- Cross-discipline
- Cross-media & mode
- Correlation & annotation
- Provenance establishment
- Source validation

Behaviors

Understand and incentivize personal, professional, and organizational behaviors

Institutional behaviors ○

- Identify sociological requirements for community-based knowledge building
- Foster culture of data stewardship
- Adapt reward structures to cyberscholarship

Personal Incentives ○

- Engage young researchers in scenario development
- Establish CAREER equivalent for cyberscholarship
- Legitimize data publication & citation as scholarship

Administration

Plan and manage program at local, national, and international levels

Workshops

- Requirements analysis & goal setting

○ **Announcements of Opportunity**

- Planning grants for components & pilots
- RFP for cyberscholarship R & D grants
- RFP for architecture prototypes

○ **Program Coordination**

- Form interagency coordinating committee with industrial advisory board
- Coordinate with Cyberinfrastructure
- Establish performance goals and evaluation processes
- Monitor progress and adjust, as necessary

The roadmap envisions four classes of work organized in "stages." The administrative stage is conducted by the responsible funding agencies. In the United States, this would include but not be limited to the NSF, IMLS, NEH, NIH, Library of Congress, and DoD. Administrative agents working in collaboration both nationally and internationally would establish program objectives and budgets and issue announcements of opportunity, broad agency announcements, or requests for proposals in order to solicit and subsequently support cyberscholarship research proposals.

Evidence strongly suggests that development of an effective cyberscholarship infrastructure requires far-reaching organizational adaptation and attention to personal motivations and incentives. This area is of such fundamental importance that it has been identified as the second stage, labeled "behaviors."

The third (research) stage focuses specifically on developing the tools and services that comprise the technological underpinning of cyberscholarship. This includes tools leading to the automatic ingest, identification, indexing, management, and analysis of scholarly communications across language barriers and among disciplines. To be widely adopted, these tools need to be seamlessly integrated into an infrastructure that provides transparent services to the scholarly community.

A series of instrumented, exploratory pilot projects kicks off the fourth (infrastructure) stage. The objective of this stage is to deploy the first generation of cyberscholarship infrastructure derived from advanced research and experimentation in tools, services, institutional behaviors, and personal incentives.

These four stages are to be coordinated over an eight-year period that is organized into the four phases of program planning, research exploration, architecture specification, and infrastructure deployment.

The NSF is the likely leader of a U.S. program coordinated through the proposed FC^{3}S. The foundation is optimized around research support for projects that last only a few years. Longer-term projects typically are assembled through a sequence of short-term projects. This model is suitable for the research prototype phase, and the NSF's Office of Cyberinfrastructure expects to fund several superdata center prototypes in the near future. However, a more comprehensive, sustainable funding model deserves attention. For the subsequent phases, the FC^{3}S will need to consider how long-term, stable support will be provided for the content infrastructure underlying cyberscholarship.

The organizational questions are simpler in Britain, and JISC may be able to use the first phase as the initial stage of an iterative development process, with less emphasis on experimental prototypes.

6.2 The planning phase

Through disparate efforts, much of the planning phase has been completed. The NSF report on Cyberinfrastructure [CI] and the Andrew W. Mellon Foundation-funded report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences [ACLS] provide the foundation. The report of this workshop aspires to provide programmatic shape and direction to a comprehensive U.S. program leading to a new generation of infrastructure for cyberscholarship.

6.3 The exploration phase

Developing the infrastructure for cyberscholarship involves research and implementation. Each of the topics discussed in Section 3 has important research questions, but three topics will receive particular attention during the prototype phase.

- Cyberscholarship depends on the effective management of very large collections of data. This is an active area of research in high-performance computing and a priority for the NSF's current Cyberinfrastructure program. What type and size of computer systems are best suited for this purpose? What services should they provide, and how should they be organized?
- Section 3.4 of this report emphasized the importance of value-added services that will drive research in cyberscholarship. What software-development projects will have the most impact, and what is needed to make them succeed?
- Scholars across many disciplines continue to discover how massive amounts of online content, with novel tools for analysis, can transform conventional disciplines. These domain experts' insights are key to the development of the content infrastructure. Actual research is needed using the prototype systems. As an example, sociologists who wish to study social networks on the Web need to learn what is technically feasible in practice; conversely, computer scientists who are building analysis tools need input from real users.

The experimental prototypes will require the intellectual contributions of many disciplines. The intention of the prototype environment is not to serve all disciplines but to test those ideas thoroughly with some substantial research problems and to think through the type of information services that are needed in other areas. At least three research prototypes should be included in this phase of the program.

A parallel measurement and evaluation effort should be initiated at the same time as the research prototypes in order to prepare for the next phase.

6.4 Architecture specification phase

Three years of developing research prototypes will not result in mature capabilities but will indicate profitable directions for further development. During the architecture-specification phase, the best results from the exploration phase will be harvested and a specification developed for the implementation phase to follow. While there will be a steady progression from experimentation to production, flexibility is important throughout the process and beyond. The architecture needs to provide a framework on which a range of production services can be built. It must not be too prescriptive, and it must not impose a mindset of how researchers in the various domains will conduct research in the future.

6.5 Deployment phase

The third phase of the program will develop widely deployable content infrastructure supporting cyberscholarship. Technically, this is unlikely to be a single monolithic system, and it certainly should be designed for continual enhancement.

A strategy for long-term sustainability of the content infrastructure will need to be in place by the time the implementation phase begins, for it is this infrastructure that will become the repository for data from all of the other short-term research projects. Indeed, with an appropriate content infrastructure in place, scholars need not concern themselves with the sustainability of their data any more than they concern themselves today with the sustainability of their published papers.

7. Summary of conclusions and recommendations

- The widespread availability of digital content is creating opportunities for new forms of research and scholarship that are qualitatively different from the traditional way of using academic publications and research data. We call this "cyberscholarship." (Section 2.1)
- The widespread availability of content in digital formats provides an infrastructure for novel forms of research. To support cyberscholarship, it must be captured, managed, and preserved in ways that are significantly different from conventional methods. (Section 2.2)
- Development of the infrastructure requires coordination at the national and international levels. JISC can provide this coordination in Britain, and SURF fulfills this responsibility in the Netherlands. The European Union's Seventh Framework Programme guides the coordinated development of research infrastructure supporting cyberscholarship. In the United States, because no single agency with this mission exists, we recommend a coordinating committee of the appropriate federal agencies. (Section 5.2)
- Development of the content infrastructure requires a blend of research — both discipline-specific and in the enabling computer science — and implementation. (Sections 4.1, 4.2, and 4.3)
- We propose a seven-year timetable for implementation of the infrastructure. The first three years will emphasize a set of prototypes, followed by implementation of a coordinated group of systems and services. (Sections 6.1, 6.2, 6.3, 6.4, and 6.5)

8. Notes and references

Many of the workshop participants submitted position papers before the workshop. The sidebars in this report are short excerpts from some of these papers, but they represent only a fraction of ideas and experience in them.

The full papers are online at www.sis.pitt.edu/~repwkshop

[ACLS]— for the report of the Commission on Cyberinfrastructure for the Humanities and Social Sciences, see www.acls.org/cyberinfrastructure/cyber.htm

[Berman]— for Fran Berman's position paper, see www.sis.pitt.edu/~repwkshop/papers/berman_schottlaender.html

[CI]—the NSF report on Cyberinfrastructure is available at www.nsf.gov/od/oci/ci_v5.pdf

[CWL]— for a description of the Cornell Web Lab, see www.infosci.cornell.edu/SIN/WebLab

[DCC]— for information regarding the Digital Curation Centre, see www.dcc.ac.uk

[DSHR]— for David Rosenthal's discussion of data storage, see http://blog.dshr.org

[FEDORA]— for information about Fedora, see www.fedora.info

[Gadian]—Dr. Alan Gadian, principal investigator, The Overlay Journal Infrastructure for Meteorological Sciences (OJIMS) Project. Available at: www.see.leeds.ac.uk/research/ias/dynamics/current/ojims.html

[HGP]— for information regarding the Human Genome Project, see www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[HSCA]— for a description of the Harvard-Smithsonian Center for Astrophysics, see http://cfa-www.harvard.edu

[ICPSR]— for a description of the Inter-University Consortium for Political and Social Research, see www.icpsr.umich.edu

[Lawrence]—Steve Lawrence, "Online or Invisible," available at http://citeseer.ist.psu.edu/online-nature01, edited version published in *Nature*, 411(6837) 521, 2001.

[Nowviskie]—Bethany Nowviskie and Jerome McGann, "NINES: A Federated Model for Integrating Digital Scholarship," September 2005, available at www.nines.org/about/9swhitepaper.pdf

[NCBI]— for a summary of tools for data mining offered by the National Center for Biotechnology Information, see www.ncbi.nlm.nih.gov/Tools

[NDIIPP]— for information regarding the National Digital Information Infrastructure and Preservation Program, see www.digitalpreservation.gov

[NSDLSSC]— for information regarding the NSDL Sustainability Standing Committee, see http://sustain.comm.nsdl.org

[SAVE]—The SAVE (Serving and Archiving Virtual Environments) project. See www.iath.virginia.edu/save

[SDSC]—"Disk and Tape Storage Cost Models," Richard L. Moore et al, *Archiving* 2007.

[Voorhees]—Ellen M. Voorhees and Donna K. Harman, eds., *TREC, Experiment and Evaluation in Information Retrieval*, MIT Press, 2005

Appendices

Appendix 1: Workshop Participants

Sheila Anderson
Director of Arts and Humanities Data Service
Kings College, London

William Y. Arms
Professor
Computer Science
Cornell University

Fran Berman
Director
San Diego Supercomputer Center
University of California, San Diego

Laura E. Campbell
Associate Librarian for Strategic Initiatives
Library of Congress

Sayeed Choudhury
Associate Director for Library Digital Programs
Hodson Director of the Digital Knowledge Center
Johns Hopkins University

Dave Cook
Head of Policy and Corporate Services
JISC

Bas Cordewener
Manager
International Collaboration
SURF Foundation

Gregory Crane
Professor of Classics
Tufts University

Ian Dolphin
Head of e-Strategy
University of Hull

Matthew Dovey
Programme Director, e-Research
JISC

Linda Frueh
Regional Director
Washington, D.C.
Internet Archive

Paolo Galluzzi
Director
Instituto e Museo Nazionale di Storia della Scienza
Florence, Italy

Jerry Goldman
Professor
Department of Political Science
Northwestern University

Stephen M. Griffin
Program Director
Division of Information and Intelligent Systems
National Science Foundation

Andrew Gruen (observer)
Northwestern University

Kenneth Hamma
Executive Director
Digital Policy and Initiatives
J. Paul Getty Trust

Mark Kornbluh
Director
MATRIX
Michigan State University

Carl Lagoze
Senior Research Associate
Information Science
Cornell University

Ronald L. Larsen
Dean and Professor
School of Information Sciences
University of Pittsburgh

Mark Liberman
Professor
Linguistics/Computer and Information Sciences
University of Pennsylvania

Joan Lippincott
Associate Executive Director
Coalition for Networked Information

Rick E. Luce
Vice Provost/Director, University Libraries
Emory University

Clifford A. Lynch
Executive Director
Coalition for Networked Information

Janet H. Murray
Professor and Director of Graduate Studies
School of Literature, Communication and Culture
Georgia Institute of Technology

Peter Murray-Rust
Unilever Centre for Molecular Sciences Informatics
University of Cambridge

Michael L. Nelson
Department of Computer Science
Old Dominion University

Joyce Ray
Associate Deputy Director for Library Services
Institute of Museum and Library Services (IMLS)

Malcolm Read
Executive Secretary
JISC

Jürgen Renn
Director
Max Planck Institute for the History of Science

David Rosenthal
LOCKSS Program
Stanford University

Brian E. C. Schottlaender
University Librarian
University of California, San Diego

Abby Smith
Independent Consultant
Historian and Cultural Resources Analyst

Beth Stewart
Technology Coordinator
Digital Humanities Initiative

Alexander Szalay
Professor
Department of Physics and Astronomy
Johns Hopkins University

Eric F. Van de Velde
Director of Library Information Technology
California Institute of Technology

Howard D. Wactlar
Vice Provost for Research Computing
School of Computer Science
Carnegie Mellon University

Donald J. Waters
Program Officer for Scholarly Communication
The Andrew W. Mellon Foundation

Bradley C. Wheeler
CIO and Dean of IT
Indiana University

Norman Wiseman
Head of Services and Outreach
JISC

Appendix 2: Position Papers

Position papers available at www.sis.pitt.edu/~repwkshop

Bill Arms	"Repositories for Large-Scale Digital Libraries"
Fran Berman and Brian E.C. Schottlaender	"The Need for Formalized Trust in Digital Repository Collaborative Infrastructure"
Laura E. Campbell	"How Digital Technologies Have Changed the Library of Congress: Inside and Outside"
Sayeed Choudhury	"The Relationship between Data and Scholarly Communication"
Bas Cordewener	"Institutional Repositories in the Netherlands, a National and International Perspective"
Gregory Crane	"Repositories, Cyberinfrastructure and the Humanities"
Linda Frueh	"Access Tools: Bridging Individuals to Information"
Jerry Goldman and Andrew Gruen	"Complexity and Scale in Audio Archives"
Babak Hamidzadeh	"Scale: A Repository Challenge"
Ken Hamma	"Professionally Indisposed to Change"
Rick E. Luce	"eDatabase Lessons for an eData World"
Janet H. Murray	"Genre Creation as Cognition and Collective Knowledge Making"
Peter Murray-Rust	"Data-Driven Science - a Scientist's View"
Michael L. Nelson	"I Don't Know and I Don't Care"
Joyce Ray	"Discussion Group on Individual Users"
Jürgen Renn	"From Research Challenges of the Humanities to the Epistemic Web (Web 3.0)"
David Rosenthal	"Engineering Issues In Preserving Large Databases"
Abby Smith	"Thoughts on Scale and Complexity"
Beth Stewart	"NEH and Digital Humanities"
Eric F. Van de Velde	"Workshop on Data-Driven Science & Scholarship: Organizations"