The Draft Report of the American Council of Learned Societies' Commission on Cyberinfrastructure for Humanities and Social Sciences (for public comment)

Commission Members:

Paul Courant
Provost & Professor of Economics
University of Michigan

Sarah Fraser
Associate Professor and Chair
Art History, Northwestern University

Mike Goodchild
Director, Center for Spatially Integrated Social Science
Professor, Geography
University of California, Santa Barbara

Margaret Hedstrom
Associate Professor, School of Information
University of Michigan

Charles Henry
Vice President and Chief Information Officer
Rice University

Peter B. Kaufman
President, Intelligent Television

Jerome McGann
John Stewart Bryan Professor
English, University of Virginia

Roy Rosenzweig
Mark and Barbara Fried Professor of History & New Media
Director, Center for History & New Media
George Mason University

John Unsworth (Chair)
Dean and Professor
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign

Bruce Zuckerman
Professor, School of Religion
Director, Archaeological Research Collection
University of Southern California

Editor:

Marlo Welshons
Assistant Dean for Publications and Communications
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign

Domestic Advisors to the Commission:

Dan Atkins
Professor, School of Information
Director, Alliance for Community Technology
University of Michigan

James Herbert
Senior NSF/NEH Advisor
National Science Foundation
Clifford Lynch, Director
Coalition for Networked Information

Deanna Marcum
Associate Librarian for Library Services
Library of Congress

Abby Smith
Director of Programs
Council on Library and Information Resources
Washington, DC

Steve Wheatley, Vice-President
American Council of Learned Societies

International Advisors to the Commission:

Dr. Sigrun Eckelmann
Programmdirektorin
Organisationseinheit
Bereich Wissenschaftliche Informationssysteme
Deutsche Forschungsgemeinschaft

Muriel Foulonneau
French Ministry of Culture, Minerva project, and European Commission
Visiting Assistant Professor
University of Illinois, Urbana-Champaign

Dr. Stefan Gradmann / Stellvertretender Direktor
Regionales Rechenzentrum der Universität Hamburg
Hamburg, Germany

Bjørn Henrichsen
Adm.dir. / Exec. Director
Norsk samfunnsvitenskapelig datatjeneste AS (NSD)
Norwegian Social Science Data Services Ltd.
Bergen, Norway

Dr Michael Jubb
Director of Policy and Programmes
Arts and Humanities Research Board
Bristol, United Kingdom

Jaap Kloosterman
International Institute of Social History
Amsterdam - Netherlands

David Moorman, Senior Policy Advisor / Conseiller principal des politiques
Social Sciences and Humanities Research Council
Conseil de recherches en sciences humaines du Canada

Professor David Robey, Programme Director
ICT in Arts and Humanities Research
Arts and Humanities Research Board
School of Modern Languages
University of Reading, England

Harold Short, Director
Centre for Computing in the Humanities
King's College London

Colin Steele
Emeritus Fellow
University Librarian, Australian National University (1980-2002)
and Director Scholarly Information Strategies (2002-2003)
The Australian National University
Canberra, Australia

Public Information-Gathering Meetings:

April 27th, 2004 Washington DC
May 22nd, 2004 Evanston, IL
June 19th, 2004 New York, NY
August 21st, 2004 Berkeley, CA
September 18th, 2004 Los Angeles, CA
October 26th, 2004 Baltimore, MD

Testimony and Background Materials:

http://www.acls.org/cyberinfrastructure/cyber.htm

Comments to:

cybercomments@listserv.acls.org

by December 31, 2005.

Table of Contents

Introduction: A Grand Challenge for the Humanities and Social Sciences

The cultural record is currently fragmented over more or less arbitrary institutional boundaries—for example, the relevant materials for understanding one artist will be held in a dozen different museums, twenty libraries, and ten archives. The resources required for work in the humanities and the social sciences are comprehensive, diverse, and complex, yet these resources are often destroyed, censored, redacted, restricted, or suppressed. When they survive, they are often to be found far away from the site of their creation and use, carried off as spoils of war, relocated in a museum, or hidden away in private collections. At present, we have the opportunity to reintegrate the cultural record, connecting its disparate parts and making the resulting whole available to one and all, over the network.

This goal constitutes a true grand challenge problem, one that would require intensive collaboration among scholars across all the disciplines of the humanities and the social sciences—cooperating with librarians, curators, and archivists—and it would require the involvement of many others, including experts in the sciences, business, and entertainment, as well as active participation from the general public. Like most grand challenges, this one can be simply stated: make it possible for people to explore the totality of our accumulated global cultural heritage, now scattered throughout libraries, archives, or museums. To do this would require using tools developed to navigate vast catalogs of born-digital, digitized, *and* physical materials—because not everything will be in digital form any time soon. The result of such an effort would be of enormous value not just to humanists and social scientists, but to everyone interested in the human record.

The fact is that that librarians, curators, archivists, and the private sector are already aligning around this objective. Librarians speak increasingly today of building the “global digital library.” Museum curators speak of “heading toward a kind of digital global museum.” Archives for some time now have been talking about and experimenting with virtual finding aids that provide unified online access to records that are physically dispersed. And Google, which has already cataloged more than eight billion web pages and one billion images, has as its stated mission “to organize the world's information and make it universally accessible and useful.” Google has also launched the Google Print project, which has made it seem technically possible to digitize collections of millions of books.

But if we were to truly create this vast collection, even Google wouldn't be sufficient to help make sense of it all, or to explore connections, or find patterns. To do that, we would need advances in both tools and standards. One effort in this area that makes such advances seem possible is the Electronic Cultural Atlas Initiative (ECAI), which is a very practical attempt to make virtual collections of scholarly data from around the globe accessible through a common interface.[1] A little further out on this same path lies the Semantic Web, which aims to be "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."[2]

All of this is by way of saying that this grand challenge is not as far-fetched as it might sound, and is, in principle, something that could be done, and—like landing on the moon or deciphering the human genome—it meets many of the criteria recently proposed by Tony Hoare and colleagues in the United Kingdom[3] for identifying grand challenges:

- It has international scope: participation would increase the research profile of a nation.
- It goes beyond what is initially possible, and requires development of techniques and tools unknown at the start of the project.
- It calls for planned co-operation among identified research teams and schools.
- It encourages and benefits from competition among identified individuals and teams, with clear criteria on who is winning, or has won.
- It necessitates collaboration of several identified research specialties, theoretical and/or practical.
- It decomposes into identified intermediate research goals, whose achievement brings scientific or economic benefit, even if the project as a whole fails.
- It should be rather obvious how far and when the challenge has been met (or not).
- It should lead to radical paradigm shift, breaking free from the dead hand of legacy.
- It is not likely to be met simply from commercially motivated evolutionary advance.

Four other criteria are included in this list, and these four may perhaps be the most important (though Hoare and his colleagues are careful to say that a grand challenge needn't meet all the criteria listed):

[1] Electronic Cultural Atlas Initiative, http://www.ecai.org/

[2] Tim Berners-Lee, James Hendler, Ora Lassila, "The Semantic Web," Scientific American, May 2001. Available at http://www.w3.org/2001/sw/

[3] "Criteria for a grand challenge," as suggested by the UKCRC grand challenges working party. Revised by Tony Hoare, May 30, 2002. http://www.cra.org/Activities/grand.challenges/hoare.pdf

- It arises from scientific curiosity about the foundation, the nature or the limits of the scientific discipline.
- It was formulated long ago, and still stands.
- It is generally comprehensible, and captures the imagination of the general public, as well as the esteem of scientists in other disciplines.
- It has enthusiastic support from (almost) the entire research community, even those who do not participate.

In a real sense, the notion of an "ultimate digital library-museum-archive" does arise from curiosity about the foundation, the nature, and the limits of the humanities and the social sciences, and certainly, a determined pursuit of this challenge would expose all of those things in new, interesting, and useful ways. In fact, the humanities and the social sciences are a long-standing effort to capture, preserve, and carry forward human culture and social life, and we have imagined the goal of comprehensiveness from the library of Alexandria to HG Wells' World Brain, the World-Wide Web.

The purpose of the grand-challenge example is to suggest that we can and should be ambitious in our thinking about what advances in information technology and communications networks have to offer the humanities and social sciences, and vice-versa, and how such advances can ultimately serve the general public. Infrastructure is costly and difficult to build, and therefore one wants it to matter as much, and last as long, as possible. As we think about the information infrastructure that we are building now, and about the fact that we will be using it for years to come, few mistakes we could make would be more expensive than underestimating the activity it will have to support. It would be easy to make exactly that mistake in the humanities and the social sciences, though, because relative to other sectors of the academy, the humanities and the social sciences are underrepresented and underinvested in cyberinfrastructure—and yet over the next generation, the uses we could have for it are potentially more demanding, and the benefits from them more inspiring, than those of many other constituents of that infrastructure. We can see that this is so just by looking at some of the examples already visible today.

Over the past two decades, the emergence of the Internet has transformed the world, and along with it the practice of the humanities and social sciences. Digital resources, networks, and tools have influenced not just the ways that scholars make sense of human cultures and societies but also the ways that these understandings are communicated to students and the general public. We are, moreover, on the verge of even greater transformations in the next decade as scholars, students, and citizens embrace a digitized cultural heritage in new and more sophisticated ways.

Recognizing that this rapid transition to digital knowledge environments is well underway, and understanding that the humanities and the social sciences have important contributions to make in designing, building, and operating this environment, the American Council of Learned Societies (ACLS) appointed a Commission on Cyberinfrastructure for the Humanities & Social Sciences. The Commission was charged with describing and analyzing the current state of humanities and social science cyberinfrastructure; articulating the requirements and the potential contributions of the humanities and the social sciences in developing a cyberinfrastructure for information, teaching, and research; and recommending areas of emphasis and coordination for the various agencies and institutions, public and private, that contribute to the development of this infrastructure.

Commission members were chosen from a representative, rather than exhaustive, list of disciplines and institutions: included are humanities scholars and social scientists, administrators and entrepreneurs, from universities and organizations public and private, large and small. Members were informed by the testimonies of scholars, librarians, museum directors, social scientists, representatives of government and private funding agencies, and many other kinds of people throughout a series of public meetings held in Washington, DC, New York City, Chicago, Los Angeles, Berkeley, and Baltimore during 2004; by information gleaned from national and international reports by other groups on related missions; and by advisors to the Commission, selected for particularly relevant expertise.

Throughout this period of research, hearings, and consultations, it has become clear that as more and more of us live greater and greater portions of our personal, social, and professional lives online, we will want—in fact, we will require—an online environment that cultivates, rather than frustrates or distorts, the richness of human experience, the diversity of human languages and cultures, and the full range of human creativity. Such an environment will not emerge by chance, but only by design, and will be better if the insights and methods of the humanities and the social sciences—clarity of expression, nuanced interpretation that uncovers meaning even in scattered or garbled information, centuries of experience in knowledge organization—are applied throughout its evolution. These methods and contributions are fundamentally important to the digital environment and the information revolution that is upon us. In fact, these capabilities are more important as the volume of digital resources grow, as complexity increases, and as we struggle to preserve and make sense of billions upon billions of sources of information.

But what is also clear is that achieving this potential requires overcoming some daunting barriers—insufficient training, outdated policies, unsatisfactory tools, incomplete resources, inadequate access. The barriers to this possibility are not primarily technological, but economic, legal, and institutional. The effort required to realize this potential is not insignificant, but these limitations are small compared to the potential benefits. This report calls for an investment not just of money but also of leadership—from commerce, education, government, and foundations—in order to realize the promise of cyberinfrastructure for the cultural record.

What is Cyberinfrastructure?

The Oxford English Dictionary defines infrastructure as “a collective term for the subordinate parts of an undertaking; substructure, foundation; specifically the permanent installations forming a basis for military operations, as airfields, naval bases, training establishments.” Although the earliest usages denote “fixed military facilities such as airfields, base installations and transport systems,” by 1971 the term has been stretched to include “a very complex infrastructure of scores of vernacular languages.” In this same year, 1971, the first humanities project to use *cyber*infrastructure had already been launched—Michael Hart’s Project Gutenberg[4], which still thrives today, making available thousands of out-of-copyright literary and historical texts (the first one was the Declaration of Independence).

Coined by the National Science Foundation (NSF) to characterize infrastructure based upon distributed computer, information, and communication technology, the newer term *cyberinfrastructure* was later popularized by the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure[5]. As the Panel suggested, “If *infrastructure* is required for an *industrial* economy, then we could say that *cyberinfrastructure* is required for a *knowledge* economy.”

One characteristic of infrastructure is that it is deeply *embedded* in the way we do our work and it is very *transparent* so that we do not have to think about how to use or interact with it. For example, the act of driving

[4] Project Gutenberg, http://www.gutenberg.org/

[5] *Revolutionizing Science and Engineering Through Cyberinfrastructure:* Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. (January 2003). Available: http://www.nsf.gov/cise/sci/reports/atkins.pdf

a car is embedded in both physical systems of minor and major roads, but also in social systems of licensing drivers, setting speed limits, and knowing when to yield at a four-way stop. Infrastructure is *built on top of an installed base;* it is *built in modular increments*, not all at once or everywhere at once. The development of *standards* and agreements about social conventions enable components of an infrastructure to work together and expand its reach and scope. For example, the infrastructure for a global telecommunications system now exists, but it was built on an installed base that evolved over the course of more than a century, and it only became global in scope during the last few decades.

Cyberinfrastructure is less visible than physical infrastructure for several reasons. First, many of its "built" components, such as high-speed networks and advanced computational devices are distributed and hidden from the end user. Second, many of the components of cyberinfrastructure are intangibles, such as software, design process, and human skill and know-how. Third, cyberinfrastructure has not yet matured to the point that it has fully achieved the characteristics of embeddedness, transparency, or global reach and scope. However, cyberinfrastructure is being built—and adopted—quite rapidly, and so it is important to have broad scholarly participation in its construction. As with other types of infrastructure, once cyberinfrastructure is built, it will be much harder to alter or improve its foundations.

Following the Panel's recommendations (referred to as the "Atkins report"), NSF adopted cyberinfrastructure as a framework for developing programs that coordinate the development of advances in computer science and engineering with the research needs of what NSF calls "domain sciences," that is, scientific tools or strategies that are specific to particular domains of study or disciplines. While the humanities and social sciences lack such a program, consensus is emerging that these areas are ripe for such investment and a concerted leadership, and recently the Final Report of the NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences" (F. Berman and H. Brady, available at www.sdsc.edu/sbe/) made a strong proposal for an NSF agenda in the social sciences.

It is important to be clear in our definition of "the humanities and social sciences" here. It is tempting to think of the humanities and social sciences as separate worlds, defined both by the commonalities within each world, and by the differences between them. But despite the appeal of such a simple model, in reality of course many disciplines are sometimes placed into one division and sometimes into the other (e.g., history), others have a degree of commonality with both, as well as substantial internal variation (e.g., various ethnic study disciplines), and disciplines such as geography or psychology are sometimes placed in the social sciences and sometimes in the natural sciences.

Rather than two distinct worlds, it seems more appropriate, at least in the context of this report, to think of the humanities and social sciences as arrayed along a continuum. At one end are the quantitative and experimental social sciences, with their rigorous adherence to the norms of science and the requisite substantial funding streams from agencies such as NSF. At the other are the humanities and the more humanistic social sciences such as cultural anthropology, history, area studies, political theory, with their very different traditions of scholarship, including little if any external funding for research, with books rather than journals as the primary publication mechanism, with few if any opportunities for postdoctoral students, and with single-authored rather than team-authored publications. The disciplines at the first extreme are sufficiently like the physical and life sciences in their needs and practices as to fit neatly within the vision of cyberinfrastructure outlined by the Atkins report. Those toward the other extreme have very different needs, and it is those needs that are examined in this report.

Cyberinfrastructure for the humanities and social sciences seems less peculiar once we reflect that scholarship already has an infrastructure. The infrastructure of scholarship was built over centuries: its foundation consists of the diverse collections of primary sources in libraries, archives, and museums; the bibliographies, finding aids, citation systems, and concordances that make that information retrievable; standards that are embodied in cataloging and classification systems; the journals and university presses that distribute the information; and the editors, librarians, archivists, and curators who link the operation of this structure to the scholars who use it. All of these structures have both extensions and analogues in the digital realm.

For the humanities and social sciences, information infrastructure is not primarily technical: it promotes learning about, research into, and public access to the products of human culture. People and policies, tools and technologies, human resources, information resources, and capital resources make all of this possible. It is important here to establish the primacy of people in the information infrastructure, for it is from them that all other elements devolve: they set the policies that govern the infrastructure; they develop the tools and technologies for the infrastructure; and they allocate the resources that enable the infrastructure’s development and maintenance.

The Atkins report articulated two layers of cyberinfrastructure: technical and enabling. These can be described as:

- *Technical infrastructure* comprises the middleware, applications, exchange protocols, and so forth, and are (or can be) largely shared among the sciences, the humanities, and social science;
- *Enabling infrastructure* is made up of the institutional policies within the academy and influenced by a broader legal context and social norms—the intellectual property and privacy rights regimes that govern access and use, the development and adoption of standards both within and across domains of inquiry, tools and services, and the education and professional training of those who are building and using the cyberinfrastructure.

Within the technical infrastructure, different research goals and different data types that characterize computational science and computational humanities and social science will require local variations in technical infrastructure, but this is the less daunting of the two layers. Instead, we believe that the enabling, or *human*, infrastructure presents specific challenges to humanists and social scientists, and is of particular importance.[6]

For this reason, our report does not focus on humanities computing and social science computing per se, but computing in the service of the humanities and the social sciences. Nor is computing itself the central topic of the report: we heard time and again of the many factors that come to the fore when talking about access to and use of the human record—policies concerning access to copyrighted material, privacy in social science data, cultures in the academy that do not support the use of information technology for research, and the chronic underfunding of the key organizations that support humanities and social science—museums, libraries, archives, university presses, and schools. The Commission heard time and again from those who wanted more advanced software applications, greater bandwidth, and more access to information technology and expertise in information technology. But these same people also urged that technology not be considered apart from the host of other social and economic factors it takes to realize the aspirations they articulated.

In fact, the relative availability of digital data in different disciplines indicates an important distinction between the cyberinfrastructure that supports science and engineering and that which is optimized for the humanities and social sciences. Data about objects of attention (and therefore, analytical methods) in the sciences are digital—not entirely,

[6] Similar conclusions with respect to the sciences can be found in Paul A. David's 2004 Research Report for the Oxford Internet Institute, "Toward a cyberinfrastructure for scientific collaboration: providing its 'soft' foundations may be the hardest part" http://www.oii.ox.ac.uk/resources/publications/RR4.pdf

but predominantly. By contrast, the preponderance of the human record is still in analog form, and the most highly prized methods of analyzing this record depend on insight and creativity—the very things we value in the record itself. For the humanities and the social sciences, then, an effective cyberinfrastructure will have to support the computer-assisted use of both analog and digital resources, and it will have to inspire creativity and provoke insight. In order to do this, it will have to be an infrastructure that enables communication and collaboration, while still permitting contemplation; it will have to embody an understanding of the continuity between digital and analog, rather than promoting the rhetoric of discontinuity; and it will have to connect the expert and the amateur, the teacher and the student, in ways that take full advantage of the ubiquity of networks, the public's enthusiasm for cultural materials accessible online, and the dedication of those who have committed their careers to the humanities and the social sciences.

This report therefore talks not only about the kind of information technologies needed for the humanities and social sciences, but about how information technology can enhance teaching, facilitate research collaboration, and increase public access to the rich legacy of human cultures across time and space. In that discussion, it addresses the particular needs and contributions of those directly engaged in teaching, research, and cultural work, but it also places those needs and contributions in a larger context, namely the public good that these activities, collectively, produce.

So far, information technology in the academic context has seen its greatest investments in the study of the natural world. Science and engineering have made great strides in using information technology to understand and shape the world around us, from the cosmic to the sub-atomic. This report addresses the possibilities inherent in the application of these same technologies to the vastly more messy and idiosyncratic realm of human experience. Scholars and teachers have been using information technology to explore and interpret the human record for more than fifty years, but it is time now to move beyond experimentation, pilot projects, and proofs of concept, to a full integration of these technologies into the core business of teaching, learning, and research in the humanities and social sciences, and into the public and private realms where knowledge communities are shaping and being shaped by these technologies.

This report also addresses the disruptive nature of new information technologies as they affect fields of inquiry into human culture that are based largely on recorded information. China and the rest of East Asia used woodblock printing as early as the seventh century, but it took until the tenth before it had transformed public, religious, and imperial life.

Johannes Gutenberg developed the printing press in the 1400s, but it took several decades for the information revolution that he and his colleagues fomented to get underway (careful as they were to make their printed books look indistinguishable from manuscripts). In the same way, it will take more than one generation or two to feel comfortable with the ways and means of living online, of digital culture, and of networked communication and publishing.

This subject is disruptive in other ways. Thinking about cyberinfrastructure requires thinking about structures: of knowledge, of the academy, and of the wider society. Scholars in the humanities have naturally been alarmed as two of the most essential elements of the infrastructure of humanities scholarship—the research library and the university press—have been threatened with failure by recent economic, organizational, and technological changes. Indeed, many have first thought about the systematic impact of digital technologies in connection with the "crises" of our cultural memory apparatus: libraries, archives, university presses, and museums.

A well-designed approach to building a cyberinfrastructure for the humanities and social sciences presents an opportunity to act with these forces and not be acted upon by them. This can also be the means to address what many have seen as, to quote the report on *Reinvigorating the Humanities* of the American Association of Universities, as the "decidedly mixed" state of the humanities, which is undergoing "a shaking out of old and entrenched attitudes and expectations. . . [a] shaking out of superannuated structures."[7]

The case for why and how to seize this opportunity is presented in the following three sections: the first chapter articulates a vision for the future of the humanities and social sciences, the second highlights some of the fundamental constraints in these areas, and the third articulates a framework for effecting the changes that are necessary to overcome those constraints and realize that future.

[7] American Association of Universities, *Reinvigorating the Humanities: Enhancing Research and Education on Campus and Beyond* (Washington: AAU, 2004), 4. Available at: http://www.aau.edu/issues/HumRpt.pdf

Chapter 1: Culture Online

Libraries, archives, and museums are cultural infrastructure. So are schools, for that matter. So are university presses. In the humanities, textual editing (which cuts across disciplines and communities of practice) contributes to the growth of cultural infrastructure, since it creates critical editions for scholarship in a number of disciplines, and these critical editions in digital form are another form of cultural infrastructure. Libraries provide shared texts in reliable editions; archives, libraries, and museums provide the documentary record on the basis of which those editions are constructed; and schools (still, sometimes) teach the theoretical and practical skills needed to produce editing. Each of these institutions now needs to fulfill its infrastructural role in a world where information is digital, communication is networked, and our ways of knowing are embodied in software.

Scholarship has always relied on technology, and the technologies of scholarship have most often been used to transmit information across space and time. Plato may hold pride of place in the long line of those who lament the transition from one medium to another—in his case (in the *Phaedrus*) from speech to writing—but he has been followed by many who note what is lost as well as what is gained with each innovation. From carved stone to scroll, from bamboo slips, palm leaves, and vellum to paper, from scribal hand to movable type—these were all giant leaps in cultural history. Scholarship has depended on these advances, and it has studied them, from the spread of Buddhism via woodblock prints, and the birth of print culture in fifteenth-century Europe, to the influence of print on political events in France during the Enlightenment to the influence of posters and movies on the illiterate peasantry of Bolshevized Russia in the 1920s. And today, sociologists, ethicists, anthropologists, and political scientists all around the world are doing pioneering work on Internet culture.

Networked access to relevant information sources in the humanities and social sciences has increased dramatically in recent years. Project MUSE offers over 250 online, full-text contemporary journals in the humanities, arts, and social sciences. The journals can be searched by key words, and the reader can follow links to relevant footnotes and other related journal articles. JSTOR, for Journal Storage, is a large archive of older publications, some extending back a hundred years. Currently, JSTOR contains 400 journals from 230 publishers, with over 14 million pages. A new project, entitled ARTStor, is based on the premise of JSTOR and focuses on art images drawn from many time periods and cultures.

ARTStor holds hundreds of thousands of images contributed by museums, archeological teams, and photo archives, as well as tools and indexes that facilitate productive use of this vast collection. InteLex Past Masters is a large dataset of full texts, usually in the form of complete works of major thinkers in the social sciences. Economics, political thought and theory, and sociology figure prominently. Social scientists and students often turn to this Web site for trusted editions of Charles Darwin, Herbert Spencer, and Adam Smith. For authors who wrote in foreign languages, an English translation is provided. *Research Papers in Economics* (RePEc) is another example of a fairly new and widely accessed database. It contains over 250,000 working papers, journal articles, and book and chapter listings, all pertaining to research in economics. *Cogprints* is often the first place scholars go for information pertinent to the study of cognition. Psychology, anthropology, and other social sciences that include elements of cognitive study are represented by a wealth of digitized research.

The Shoah Visual History Archive

In 1993, Steven Spielberg was filming *Schindler's List* in Jackson, Mississippi. During the shoot, he met with many Holocaust survivors and asked what he could do for them. Most responded that they wished that their memories and recollections of the war years could be recorded and preserved, and not lost to time. In due course, Mr. Spielberg created the Survivors of the Shoah Visual History Archive, an immense database of 52,000 video interviews with survivors of the concentration camps and others who were persecuted during the Second World War. Their testimonies were taken in 56 countries and recorded in 32 languages. The archive, which would require thirteen and a half years to view non-stop, is now housed and administered in Los Angeles. Citizens of Jackson researched this database, and discovered that 19 Holocaust survivors still lived in the city. After some negotiations, copies of the videos of those survivors and other relevant tapes were sent to Jackson, where they are available through the town's public library as an openly accessible resource that combines facets of local history, sweeping world events, and deeply personal stories. Through these videos and those of the tens of thousands of survivors, we may all become witnesses to a key event in 20th century history that might otherwise have been lost in the passage of time and the passing of that generation.

In the mid 1980s, the American Council of Learned Societies surveyed almost 4,000 scholars in the humanities and social sciences to capture a comprehensive portrait of what they "think about a wide range of issues of greatest concern to their careers, their disciplines, and higher education in general." In the book-length report of the survey, the very first finding highlighted is "the rapid increase in computer use." "In 1980," the report notes, only "about 2 percent of all respondents either owned a computer or had one on loan for their exclusive use." But by 1985, it observes with some obvious excitement, "the number was 45 percent, most of whom used it not only for routine word processing but for other purposes as well." Those "other purposes" were, however, clearly minority pursuits. Only about one in five scholars reported using online library catalogs or databases; only one in ten used email; just 7 percent (most of them in classics or linguistics) said that they had used a computer for "theme, text, semantic or language analysis." Three years later, the Research Libraries Group (RLG) published a detailed assessment of information needs in the humanities and social sciences. The humanist scholars interviewed were consistent across disciplines, articulating a pervasive need to create more machine readable catalogs, indexes, and other finding aids. There was little interest in making the contents of those repositories available in digital form, in part because the technology was still nascent and untested, coupled with a prevailing acceptance of the informal, book-based, and often serendipitous browsing methods of scholarship that had played so fundamental a role in humanities research for centuries. Image databases for two- and three-dimensional objects were largely beyond the capacities of the technology, and the economics, of the time.

The RLG report on social sciences showed those disciplines to be more technologically dependent than the humanities; almost every social science discipline in 1988 had a trusted machine-readable index associated with scholarship and research in the relevant academic fields. The social sciences were more interested in the availability of electronic databases and datasets for research support, with examples such as the census and ICPSR materials already well established in multiple disciplines. Scholars in these disciplines also expressed interest in using technology to improve access to conference papers, unpublished research and technical reports.

The Mellon International Dunhuang Archive (MIDA)

The Mellon Dunhuang Archive is the product of a major and ongoing multi-institutional, multi-national effort to recreate high-quality digital reconstructions of the murals, manuscripts, and sculpture of several hundred Buddhist cave shrines in Dunhuang, China, a uniquely important cultural crossroads on the ancient Silk Route in the Gobi Desert. Using digital cameras, a team from Northwestern University, in collaboration with the Dunhuang Research Academy (the Chinese administrative unit that seeks to preserve and document the caves), has photographed the wall paintings and sculpture in forty-two grottoes. The team captured and rendered high-resolution representations of the caves in two-dimensional image files and three-dimensional visual representations that can be viewed using "virtual reality" technology, making available material that otherwise would be inaccessible even in person due to the height, darkness, and location of the paintings. Ultimately, the Mellon International Dunhuang Archive seeks to reunite "virtually" and present to scholars a rich body of primary source materials that remain in China and those carried to off-site collections around the world. The contents of the Archive are being placed on ARTstor, an independent non-profit publisher of art history images (http://www.artstor.org/info/).

The Archive of the Indigenous Languages of Latin America

At the University of Texas at Austin, Joel Sherzer has for decades researched and gathered data on the indigenous languages of Latin America, from Tierra del Fuego to the highlands of Mexico. These are languages that are destined to be changed and in some cases obliterated by the increasingly dominant languages of Portuguese, Spanish, and English across the hemisphere. Professor Sherzer and many colleagues have braved rainforests, mountaintops, and urban jungles to record the sounds of these vanishing cultures, and these recordings have been digitized and transferred to servers at the University of Texas. Without the recordings made by these researchers, our only evidence of the existence of these languages might be secondhand, in books and journals, and would be significantly less extensive. One language this team has recorded is already extinct; others will surely follow. Future scholarship and, more broadly, the basic human need to know and understand our origins, demand the preservation of this precious data.

In 1997, the American Council of Learned Societies issued a study focusing on information technology in the humanities. Published less than ten years after the RLG reports, it clearly indicated a change in the degree of acceptance of technology in the humanities, the level of technical knowledge, and sense that information technology could enrich and influence research. Chief recommendations included a call for a national strategy for digitizing texts, images, sound, and other media pertinent to the cultural heritage, and the cited need for coordinated large-scale projects to effect this digitization; more pervasive technical standards; greater attention to the challenges of preservation of digital information over time, including ongoing accessibility even as operating systems and hardware configurations changed; and a need to promote within the universities a more hospitable environment for computer-supported arts and humanities.

That the findings and recommendations of the 1988 report would seem almost quaint to those scholars interviewed less than a decade later underscores the revolutionary advances in information technology that now permeate the world of humanists and social scientists. Almost every scholar regards a computer as basic equipment; colleagues view those who write their books and articles without the assistance of word processing software as objects of curiosity. Email and instant messaging has broadened circles of communication and increased the amount and, arguably, the quality of debate among dispersed scholarly communities.

Today, both scholars and the general public are highly dependent on computer networks and digital resources to do their jobs and to learn about the world. Whereas in 1985 those using email were a tiny avant-garde, in 2005 a two-hour shut down of an email server paralyzes the entire population it supports. When the American Historical Association’s email connection went down for ten days in 2004, the outage caused so much consternation that the Association called its lawyers to take action against the Internet provider and decided that it would need two suppliers in the future. Computer redundancy—once a concept confined to space missions—now turns out to be required for historical missions. As recently as a decade ago, a scholar employed at a small liberal arts college in the Midwest who needed to survey the scholarly literature for a new project might have had to drive a couple of hundred miles to a major university library to find and copy the relevant journal articles. Now, his college probably subscribes to digital databases where many, if not all, of the articles can be downloaded in an hour or less. As recently as a decade ago, a scholar working at a community college in the South had only very sporadic contact with other scholars in her specialty. Once in a while, a scholarly meeting might be scheduled for a big city within driving distance or she might be lucky enough to persuade her dean to send her to a more distant meeting. Today, listserv discussions arrive in her mailbox every few hours, and she regularly emails drafts of her latest articles to colleagues around the world for comment and discussion.

These changes are leading to new forms of organized scholarship. New communities are forming around technologies that enhance the research and teaching of their disciplines. One example is NINES, a scholarly collective founded to develop a publishing environment for integrated, peer-reviewed online scholarship centered in nineteenth-century studies, British and American[8]. Traditional disciplines are changing profoundly. Classics, linguistics, anthropology, and many of the social sciences cannot be taught easily, or in some cases well, without recourse to digital technology and online resources. New disciplines are emerging. New fields of study, such as archaeometrics, archaeogenetics, music informatics, new facets of bioethics, and the anthropology of the Internet augur new methods of research and new discoveries that may in turn lead to other new fields of inquiry. Every day we see more cross-disciplinary work involving humanities and social sciences with engineering and the sciences. Many of the above-mentioned fields rely on networked resources, information technology, and digital tools.

Cultural Infrastructure and the Public

The digital revolution in the humanities and social sciences has had profoundly democratic consequences both for scholars and the general public. It is astonishing to think of the information now accessible to anyone with a computer and a network connection, especially when compared to the state of our digital resources only fifteen years ago. In 1990, there was no World Wide Web; today, the total number of Web

[8] NINES, http://www.nines.org/

pages defies estimation. Millions of Americans now go online for everything from purchasing consumer goods—Amazon.com is one of the fastest growing companies in the world; Ebay hosts hundreds of thousands of auctions each year—to researching an incalculable number of topics. As evidence that a radical reorientation of information-seeking behavior is far more widespread than the near-universal reliance on search engines among faculty and students: it would be difficult for many inside or outside the academy to imagine a day without Google.

The Internet and the World Wide Web have also allowed unprecedented access to medical information. More Americans now turn first to the Web for knowledge about medical conditions, medications and their side affects, and medical research before consulting their physicians. A search for Web sites pertaining to heart disease yields 7.5 million Web addresses; a similar search for diabetes yields over 13 million Web locations. Nearly 70% of adult Americans access the Internet for many of these services, as well as to read newspapers, schedule appointments, send email to friends and family, and pay their bills.[9] Our access to the Internet and its resources can be fairly characterized as indispensable. Just as the loss of Internet connectivity occasioned a crisis for professional historians in the American Historical Association, a similar digital outage would occasion a crisis for many, including students who rely on Internet sources to complete their homework, either in their homes, schools, or libraries.

For the general public, putting the documentary record of the past online means that record is open to people who rarely had access before. If digitized properly, many online materials, both text and images, can become accessible through screen readers and other assistive technologies to those with visual impairments or other disabilities. The analog Library of Congress currently does not welcome high school students—its reading rooms, no less its special collections, routinely turn them away. But now the Library's American Memory Web site allows high school students to enter the virtual archive on the same terms of access as the most senior historian or member of Congress.

Not surprisingly, another feature of this remarkable connectivity and access to resources is that it has brought scholars and scholarship more closely in communication with non-scholarly audiences. Humanists and social scientists now routinely hear from students and members of the general public who have found their email addresses and have questions to ask. Scholars who have created Web sites out of their scholarly work

[9] See http://www.pewinternet.org/trends/User_Demo_08.09.05.htm, and more generally, the Pew Internet &American Life project and its reports, such as "Internet: The Mainstreaming of Online Life" (http://www.pewinternet.org/pdfs/Internet_Status_2005.pdf).

are invariably startled to find that they have tapped into entirely new audiences. Non-academic users of the University of North Carolina's archival Web site, Documenting the American South, speak eloquently of how they "felt privileged to have access to these primary sources as if they had entered an inner sanctum where they did not fully belong," reports university librarian Joe Hewitt.

The Field Museum

The Field Museum in Chicago has over 600,000 objects in anthropology, 275,000 volumes in its research library, 500,000 photographs, and 2,500 linear feet of documents in its institutional archives. To take but one example, the Field Museum maintains an important collection of documents and images from the 1893 World's Columbian Exposition, an event that not only gave birth to the Museum itself, but also gave rise to twentieth-century research on and public awareness of anthropology and conservation. Collections such as those in the Field Museum are just a small piece of the extensive and actively used resources found in museums of art, natural history, maritime history, and other topics. Although these collections could be made available online, almost none of them are now. The artifacts in these collections are increasingly a primary foundation of scholarly research, as well as being used in documentaries, television programs, feature films, and other kinds of entertainment. Unlocking this content by digitizing it, and aggregating it by sharing it over the network, would profoundly change the way we undertake research and education worldwide.

But even the vast amount of material currently available online is nothing compared to what could still be made available. The material in analog form in the Library of Congress's reading room dwarfs that available through the American Memory collection, and with a mission to provide "texts, images, and audio files related to Southern history, literature, and culture from the colonial period through the first decades of the 20th century," much unfinished work remains for Joe Hewitt and his colleagues documenting the American South. To those who previously had no easy access, online collections do open doors, but many more doors remain closed.

In earlier ages, the problem was that there was not enough information: literacy was scarce, reading material was expensive and rare. In the 21st century, though, we are deluged with data. In "How Much Information," Peter Lyman and Hal Varian have tracked the steadily increasing amounts of information produced each year, in all media: in 2003, they estimated production of 300 terabytes (TB) of print, 25TB of movies, 375,000TB of digital photography, 987TB of radio, 8,000TB of television, 58TB of audio CDs—and that doesn't count software (like video games) or materials originally produced for the Web, or more ephemeral forms of digital information like phone calls or instant messaging[10].

[10] Lyman, Peter and Hal R. Varian, "How Much Information," 2003. Available at http://www.sims.berkeley.edu/how-much-info-2003

Without question, we have at our fingertips more information than at any earlier time in history. Academic libraries are exemplary of this digital growth. About 2% of a library's acquisition budget was spent on electronic subscriptions and resources in 1996; today the average is 30%, with some institutions spending over half of their collections budget on digital materials, most of which are electronic versions of journals that they may or may not continue to receive in print form.

How can we make sense of this surfeit? The practice of scholarship must adapt, and cyberinfrastructure not only offers a chance to do that, it requires a response from the humanities and the social sciences. We have remarkable opportunities to advance our understanding of human cultures and societies past, present, and future, but only if scholars can rethink outward forms and settled practices and in the process discover a new analytic and interpretive power for the humanities and the social sciences in this age of change.

The landscape of digital humanities and social sciences is populated by many examples of networked computing providing unprecedented access to a variety of cultural artifacts. Thanks to high-end digital imaging, we can examine and compare ancient cuneiform inscriptions with new precision and clarity,[11] and we can see the much-damaged manuscript of Beowulf in a way that renders the text more legible than the original, and we can "peel back" successive conservation treatments to see how the varying states of the artifact over time have influenced interpretation.[12] Other ambitious and comprehensive editing projects reproduce the complex genealogy of a medieval text[13] or recreate the many sources and states of the work produced across an entire lifetime by a prominent author in the age of print.[14] Three-dimensional modeling makes it possible to recreate Roman forums,[15] medieval cathedrals,[16] and Victorian exhibitions,[17] and these models provide more than just a sense of place: the process of building them can give us a deeper understanding of how the original structures themselves were built. Digital video reformats fragile film and gives us access to rare footage of dance performances from the early decades of the last century. Mapping technology allows us to understand the rapid spread of religious hysteria in the Massachusetts Bay Colony during the seventeenth century,[18] or to

[11] Cuneiform Digital Library Initiative. (2005). UCLA and Max Planck Institute. http://cdli.ucla.edu/

[12] The Electronic Beowulf. (2003). British Library. http://www.uky.edu/~kiernan/eBeowulf/guide.htm

[13] The Piers Plowman Electronic Archive. (2005). http://jefferson.village.virginia.edu/seenet/piers/

[14] The Rossetti Archive. (2005). http://www.rossettiarchive.org/

[15] Cultural Virtual Reality Lab. (2005). UCLA. http://www.cvrlab.org/

[16] Salisbury Project, Cathedral Model (2005) UVa. http://www3.iath.virginia.edu/salisbury/model/index.html

[17] The Crystal Palace. (2005). UVa. http://jefferson.village.virginia.edu/london/model/

[18] The Salem Witch Trials. (2005). UVa. http://etext.virginia.edu/salem/witchcraft/home.html

observe the evolution of the built and natural environment around Boston's Back Bay over two centuries.[19] The Valley of the Shadow project contains extensive records in the form of digitized diaries and letters, newspapers and statistical records, photographs and other images of the period leading up to and following the Civil War; it also has animated maps of battles that visually reconstruct troop movements, points of battle engagement, and other data drawn from army and navy records of the time.[20] Academic and public libraries, museums, and historical archive programs contribute digital data to the project, which has a mission to make information on American culture, architecture, performing arts, presidents, as well as the histories of women, Native Americans, and African Americans, accessible to the general public.

These and other digital projects highlight the capacity of digital technology to make the past more present, the distant closer, and the strange more familiar. Cyberinfrastructure offers us new ways of seeing art and sculpture, new ways of bearing witness to history, new ways of hearing and remembering human languages, new ways of reading texts—both ancient and modern. That same infrastructure can allow us to work in collaboration with distant partners who share our interests, who provide complementary expertise, and whom we may only rarely meet face-to-face. All of that is, in some sense, about access—either access to colleagues; or access via digital representations to distant, damaged, or disappeared physical artifacts; or intellectual access to the meaning or significance of these artifacts.

These projects are also distinguished for the resources they make available, but as the Field Museum, Library of Congress, and Documenting the American South examples earlier make clear, much more remains to be digitized. In some important sense, we have yet to realize the promise of our digital collections, beyond convenience of access and the raw power of aggregation. The existence of these collections, then, poses a challenge: how do we use them to ask new questions and answer old ones? Meeting that challenge will—increasingly, over the next generation—involve not only access to the materials of the humanities and social sciences in digital form, but also the use of tools that enable collaboration and turn that access into insight. Scholars in the humanities and social sciences cannot depend on colleagues in computer science or engineering to build these tools: they have their own research to do, and only in rare cases does that research involve the application of well understood technologies in new domains. By the same token, commercial interests cannot be depended on to develop all of the software tools that are needed, though they may

[19] Evolutionary Infrastructure. (2005). UVa. http://www3.iath.virginia.edu/backbay/
[20] The Valley of the Shadow. (2005). UVa. http://valley.vcdh.virginia.edu/

have an important contribution to make in accelerating the pace of digitization, as in Google Print's recent agreements to do wholesale digitization of research library collections. If the promise of the digital library is to be realized, then humanists and social scientists need to contribute to the design and development of tools for digital humanities and social science, and support systems for that development effort will need to be built—research centers that are national repositories of expertise, postdoctoral programs that emphasize digital scholarship, and graduate programs that train the rising generation in computational methods.

If this emphasis on the importance of the computational seems unwarranted, consider how seemingly mundane tools have already changed the way we work: email, Web pages, search engines, full-text search capabilities, and digital media were either non-existent or scarcely known a generation ago, but now none are exotic or even new. They are used by school children and grandparents, by hobbyists and community organizers, as well as by scientists and scholars, and they allow individuals of all ages from all over the world to interact with one another and with the cultural record more deeply, more democratically, more effectively, and more conveniently. Furthermore, they all support the creation and interpretation of culture, and they make available to all the rewards of learning and of remembering. It seems self-evident that the enormous increases in computational power, storage capacity, network ubiquity, and computer literacy that we've seen since the invention of email should make it possible to imagine, design, and deploy tools with proportionately greater impact on teaching, research, and the society at large.

What will those tools do? A general answer to that question was offered to the Commission in its first public hearing, in Washington, DC, by Michael Jensen, electronic publisher for the National Academies Press: "human interpretation is the heart of the humanities. . . . devising computer-assisted ways for humans to interpret more effectively vast arrays of the human enterprise is the major challenge." In practice, this means that tools for use with digital libraries will need to enable the user to find patterns of significance in very large collections of information, across many different types of data. Today, we refer to this kind of work as "data-mining" (or sometimes, when it is confined to text, text-mining), and we see it used mostly in corporate settings, for purposes of competitive intelligence, and sometimes in scientific applications, as a strategy for finding meaning in very large datasets.

Data-mining is only one investigative method, or class of methods, that might become more useful in the humanities and the social sciences, as we harness greater computing power and bring it to bear on larger and larger collections, often with outcomes in areas other than that for which the data was originally collected. It is raised here in order to suggest that we can (and should) expect more from digital libraries than searching and browsing. Beyond this, we can imagine many other ways of animating and exploring the reintegrated cultural record, through simulations that reverse-engineer historical events to understand what caused them and how things might have turned out differently, through game-play that allows us to tinker with the creation and reception of works of art, through role-playing in social situations with autonomous agents—the possibilities will only expand, as computers become better and better at dealing with human language and cognition, and at representing existing and conceivable worlds with ever-greater realism.

However, as the Commission heard in its Berkeley meeting, "the social sciences and humanities are different from the physical and biological sciences in the variety, complexity, incomprehensibility, and intractability of the entities that are studied. Consequently, the physical and biological science models in the National Science Foundation's report *Revolutionizing Science and Engineering through Cyberinfrastructure* do not directly apply to the social sciences, which have different kinds of problems. These problems make it difficult to understand social reality in the first instance, and they pose special problems for creating cyberinfrastructure for the social sciences. But they also provide interesting challenges for computer scientists, digital librarians, and social scientists themselves. Perhaps most importantly, overcoming these problems provides the opportunity to revolutionize the social sciences" (Henry Brady).

By the same token, overcoming these problems could also revolutionize cyberinfrastructure, and that provides some incentive to address what Brady calls the "variety, complexity, incomprehensibility, and intractability" that characterizes the human record. If we can design the software tools, the computer networks, the digital libraries, archives, and museums that we need in order to assemble, preserve, and examine that record, we will have done something much more difficult than supercomputing alone, and the impact of the accomplishment will be felt far beyond the disciplines that will be revolutionized in the process. Yet many barriers stand between where we are now and a future in which we might realize something like the unification of the cultural record. Some of these challenges are technical, but by far the most formidable are human and social—whether legal, organizational, disciplinary, political, or economic. Humanists and social scientists are experts in such human and social problems, so perhaps we can address them, but doing so will require a serious engagement. The next chapter describes what the Commission has come to believe are the greatest of these challenges.

Chapter 2: The Digital Migration

A well-designed cyberinfrastructure, as we have argued, will promote democratic access to vast digital resources, contribute to the discovery of new knowledge, transform teaching, and foster the public good. Indeed, it has already begun to do some of those things. The adoption and diffusion of digital technology might be compared to some of the great human migrations in ages past. Throughout history, waves of people have moved from one physical location to another, establishing new outposts of civilization, spreading agriculture, industry, and technology, and transforming society in the process. In the digital migration, the paths we traverse are high-speed networks; the knowledge we acquire is largely through distributed repositories of information; and the new worlds we discover come to us, rather than our coming to them.

What is required to complete this migration? As we compare ourselves to Australia, Canada, England, and Europe, we see that there has been proportionally much greater support for the development of a broadly accessible, broadly useful cyberinfrastructure in these countries than in the United States. To begin the work of correcting this, the Commission has identified seven key barriers to building robust cyberinfrastructure. These include the money needed to fund the work, the nature of digital work in the humanities and social sciences, laws and policies that govern access to and distribution data, and the organizational cultures that many perceive as inhibiting rather than supporting innovation.

1. Funding

By any standard, available funding for building an American cyberinfrastructure is meager, as is American research funding in general: according to Vint Cerf in the Wall Street Journal (July 27, 2005), "our total national spending on R&D is 2.7% of our GDP, and now ranks sixth in the world, in relative terms, behind Israel (4.4%), Sweden (3.8%), Finland (3.4%), Japan (3.0%) and Iceland (2.9%). The federal government's share of total national R&D spending has fallen from 66% in 1964 to 25%" in 2005. In 2003, the Atkins report recommended annual expenditures of $1 billion to create a cyberinfrastructure for science and engineering: in 2005, funding specifically designated to shared cyberinfrastructure at NSF was about $123 million. Even that is a great deal more than what has been spent to date on cyberinfrastructure needs peculiar to the humanities and social sciences. Even if the humanities and social sciences realize some substantial benefits from NSF investment in cyberinfrastructure, it seems clear that additional investment from other sources will be necessary.

Federal funding patterns reveal the generally limited public funding support for humanities and social sciences. Health research accounts for more than half of federal spending on basic (non-defense) research: the National Institute of Health's budget in 2004 was around $28 billion. The National Science Foundation, which provides some funding for the social sciences and almost none for the humanities, was $5.5 billion. Within that about ten percent, or $584 million went to the Directorate for Computer and Information Science and Engineering (CISE), which until recently had the primary responsibility for cyberinfrastructure—of course, CISE's budget also funds NSF's portfolio of basic research in the computer and information sciences and related areas. The NSF now has an Office of Cyberinfrastructure, which has initiated a comprehensive strategic planning process to guide the agency's investments in cyberinfrastructure for science and engineering. Federal funding for humanities-related projects is paltry in comparison. The 2004 budgets of the most important agencies—the National Endowment for the Arts ($139 million), the National Endowment for the Humanities ($162 million), and the Institute for Museum and Library Services ($262 million) taken together do not even equal the budget for CISE, which is itself only one-tenth of the NSF budget and one-fiftieth of the NIH budget. Additionally, the ability of the NEA, NEH, and IMLS to fund cyberinfrastructure directly is diminished because much of the money in these agency budgets goes directly to states through block grants that the agencies have little control over.

Private foundations are important sources of support in the humanities and the social sciences, but they cannot make up all the difference. Indeed, no single private foundation in the United States—with the exception of the Gates Foundation, which primarily funds health issues—has an annual giving amount that equals the budget of CISE—and among the large private foundations, few are focused on humanities and social science[21]. Nevertheless, philanthropic sources have so far played a disproportionately large role in funding the experimentation in digital projects in the humanities. Key foundations—notably the Andrew W. Mellon Foundation and others such as the Getty Trust, the Carnegie Corporation, the William and Flora Hewlett, David and Lucile Packard, and Alfred P. Sloan foundations—have made strategic investments in building resources or seeding them. But when such online enterprises reach maturity and try to move from project to program, the problem of sustainability can become insurmountable.

[21] The Foundation Center. "Foundation Growth and Giving Estimates," 2005. Available at http://fdncenter.org/research/trends_analysis/pdf/fgge05.pdf

The private sector has also provided remarkable instances of individual philanthropy, with the emergence of a new cadre of digital collectors such as Brewster Kahle (the Internet Archive), Rick Prelinger (Archive Films), and David Rumsey (the David Rumsey Map Collection). They not only collect high-value resources for humanities and social sciences but also are committed to providing free access to them on the Web and developing cutting-edge services to enable their use. But these efforts—however laudable—cannot by themselves fill the gap.

2. The characteristics of data in the humanities and social sciences

Digitizing the products of human culture and society poses intrinsic problems of complexity and scale. The complexity of data concerning human cultures—data that are multilingual, historically specific, geographically dispersed, and highly ambiguous in meaning—makes digitization complex and expensive. Like science data, humanities and social sciences data are also massive in scale, perhaps more so if one tries to imagine what it would be like to gather all the content of all the museums, libraries, and archives into one space at one time, or even to digitally record the daily life of a single human being.

Moreover, a critical mass of information is often necessary for understanding both the context and the specifics of an artifact or event, so that often a very large dataset of multimedia content—image, text, sound, moving image, audio—is required. Humanities scholars are often concerned with how meaning is created, communicated, manipulated, and perceived through discourse. Recent trends in scholarship have broadened the definition of what falls into the category of discourse, and many scholars formerly comfortable working only with texts now turn regularly to architecture and urban planning blueprints, art, music, video games, film and television, fashion illustrations, billboards, dance videos, gesture, graffiti, food, rituals, as well as blogs.

While difficult to achieve, the value of critical mass or functional completeness is easily demonstrated. The Shoah archive described earlier has a signal authenticity in large part because it is so comprehensive. The tale of what happened to one or two families, in one or two villages, in one or two countries, during the Holocaust is worth recording and disseminating—but how much greater is the knowledge we gain from the completeness of the record? It is not mere quantity that matters. In history, art history, classics, or any scholarly enterprise that benefits from a comprehensive comparative approach, quantity can become quality.

The problems are multiplied by the multiple audiences for humanities and social science data, where there can be many subject specialists who want access to the same sources for different reasons. The beautiful *Roman de la Rose* project, a digital collection of the major illuminated manuscripts of one of the most popular medieval literary works, is used by literary scholars, art historians, linguists, social historians, and preservation specialists, each of whom has a different disciplinary perspective and vocabulary.[22] Even more important, cultural documents often have student and popular audiences, and since those audiences require further contextualization, the data or evidence itself needs to be more self-describing and self-contextualizing.

3. Barriers created by intellectual property restrictions

The framers of the U. S. Constitution sought to balance the rights of the creators of intellectual property and the claims of the larger community. Article 1, Section 8, grants Congress the power to give "authors and inventors the exclusive right to their respective writings and discoveries," but it also specifies that such rights be granted only "for limited terms" and with the purpose of promoting "the progress of science and the useful arts." Today, many people (including most of those from whom the Commission heard) fear that the balance has been upset and the property claims of rights holders are interfering with the promotion of intellectual and educational progress.

Indeed, the most notable recent U.S. Supreme Court decision on copyright—*Eldred v. Ashcroft*—involved someone who was seeking to disseminate works in the humanities to a broad public. Eric Eldred was the organizer of the Eldritch Press Web site, dedicated to providing free books by such authors as Nathaniel Hawthorne. He sued to overturn the Sonny Bono Copyright Term Extension Act of 1998 (CTEA) on the grounds that its twenty-year extension subverted the constitutional provision of "limited" copyright terms and did nothing to promote new creativity. Eldred had wanted to add to his Web site Robert Frost's poetry collection *New Hampshire*, which was slated to pass into the public domain in 1998.[23] But the CTEA halted his plans. Eldred's loss in the Supreme Court effectively ended the plans of thousands of other digitizers to add historical and cultural works from the 1920s to the public Web for another twenty years. Many scholars, librarians, and grassroots digitizers like Eldred believe that when that time comes, Congress will again extend copyright, benefiting the owners of profitable commercial works at the expense of the public domain. The promised,

[22] Roman de la Rose, http://rose.mse.jhu.edu/

[23] See http://www.legalaffairs.org/issues/March-April-2004/story_lessig_marapr04.msp

democratic digital access to our cultural heritage currently ends in 1923. All of Hawthorne is up on the Web, but most of F. Scott Fitzgerald is not.

Equally frustrating is that many lesser-known works of creativity and culture—not just books, but also photos, drawings, films, and other materials—from the 1920s and later years cannot be made available online simply because the rights holders are difficult or impossible to find. Because recent copyright law has eliminated the requirement that rights-holders formally apply for renewal, the copyrights of these so-called "orphan works" are automatically extended. Although such works often lack commercial value, and the non-textual materials especially may prove to be the most critical in terms of scholarship and preserving cultural memory, the expense and difficulty of locating the rights-holders blocks their digitization. Moreover, even more complex issues arise in providing access to unpublished works (manuscripts and letters, for example), a category of particular importance to the humanities. The vast majority of primary sources that are protected by copyright are protected because of sweeping legislation that traps things under copyright, in cases where authors had no intention to publish, nor any economic interest. Many sound recordings, too, are effectively "protected" from being reproduced in the practice of scholarship until the latter half of the 21st century, when any scholar now practicing is likely to be dead.[24]

Current copyright laws not only keep older works from becoming available in digital form, they also threaten the preservation of born-digital works. The copyright code as it exists now has several important provisions that foster access and preservation—qualities that are jeopardized by the transition to digital distribution. One of the exceptions in the current copyright code, Section 108, allows for libraries and archives to copy works (in quantities specified by case law) to preserve the intellectual content. This has included copying works from one medium to another, such as brittle paper to microfilm or nitrate film to safety stock. Copying to digital form is allowed for preservation purposes (*not* for access), but it is not clear that all the forms of copying and normalization needed for secure digital archiving are, in fact, allowable under the law. Another exception, the doctrine of First Sale (Section 107), allows the purchaser of an item such as a book subsequent unregulated use of that book. This is the reason libraries can lend books to readers and that book owners can sell used books to other readers. Both sections of the law have lost their rationale in the age of digital replication. The result is that libraries are now offered e-books and e-journals not for purchase, but for limited, licensed access. Libraries cannot copy content

[24] Cliff Lynch: sound recordings are protected till someting like 2060 before they pass into the public domain because of the interaction of federal and state laws, at least in New York State.

of the databases for preservation; and they are limited in the number of individuals to whom they can give access, both onsite and online.

This has seriously jeopardized the preservation of published electronic materials, a problem that is bound to escalate as more and more content is distributed in database form. It has also eroded the ability of public libraries, indeed any library that is not exceptionally well funded, to serve its patrons, and it has placed innovative efforts to preserve the Web, such as the Internet Archive, in an ambiguous legal position.

These are classic examples of the unintended consequences of technology innovation. In this case the technology solves one access problem—and it is a significant one—only to create a new one. It affords us the opportunity of greatly increased accessibility and collaboration and, given the current state of intellectual property law, it also presents the risk that we will be come unable to study our own culture and cultural development. In other words, we could become much worse off than we have been historically, not least because the law thwarts a reliable and cost-effective means to preserve cultural content as a service to the public.[25]

Finally, although current copyright laws pose the most significant restrictions, a host of other legal doctrines can present problems to humanists and social scientists, including trademark, patent, rights to likeness, and rights to privacy laws. The particular problem of privacy laws is addressed in the next section, but it is important to note that while the legal tradition of copyright is a textual tradition, as evidence and scholarly communication are extended beyond the textual, scholars encounter other various legal requirements involving non-textual materials. The need for signed releases from people appearing in photos is just one example. And because the networked environment in which we work is intrinsically international, international laws can create incompatibilities and disagreements about permissible speech of all kinds.

4. The public record: barriers to contemporary social science

Emerging constraints increasingly impede contemporary social science research. One is the growing societal concern for privacy. In some countries this has reached the point where a periodic census has become impossible, and is being replaced with a collection of sample surveys. Technology worsens the situation in several ways. It exacerbates the sense that data-collection agencies may be operating secretly or invisibly;

[25] For a concrete example of this, see Jeff Ubois, “New Approaches to Television Archiving” (10:3, March 2005), http://firstmonday.org/issues/issue10_3/ubois/index.html.

it is susceptible to misuse and exploitation; it allows widespread and continuous monitoring; it allows the linking of previously independent records to reveal identity information. Contemporary history and social science are also increasingly constrained by the requirements of Institutional Review Boards (IRBs), which were created to assure the rights of research subjects in the wake of scandals like the Tuskegee experiment in which disenfranchised African Americans were denied treatment for syphilis. Although it is important to conduct social science research in ways that protect and respect the rights of individuals, some worry that “mission creep” on the part of IRBs has pushed them into areas beyond their initial charge and have eroded the First Amendment protections of scholars.[26]

Another recent trend in developed countries is a major shift toward outsourcing data collection, as central governments attempt to downsize. Statistical agencies are being privatized, and traditional services are being discontinued. At the same time the private sector is investing significantly in data gathering, both for its own purposes and to develop business opportunities. But commercial data sources are priced to recover the cost of production, whereas researchers are more used to public sector sources priced at the cost of reproduction, if at all. Private-sector sources are often less subject to the norms of science—replicability, rigorous definitions—and are collected for purposes different from those of social scientists, a fact that often has implications with respect to sampling.

Finally, following September 11th, some federal agencies removed from public access datasets that they regarded as compromising national security. Researchers who had traditionally had unlimited access to public sources, particularly sources of geographic information—locations of cultural and religious institutions, basic topographic data—suddenly found their work halted.

5. Barriers created by the loss and fragility of data

The study of human cultures and creativity is founded on access to the records of the past. The accountability of the government to its people is premised on a clear, accessible, and unbroken audit trail of past actions. Increasingly, scientific studies of the biosphere, the geosphere, and the

[26] See Philip Hamburger, “The New Censorship: Institutional Review Boards,” 2005. Paper ID: U of Chicago, Public Law Working Paper No. 95. Available at http://papers.ssrn.com/paper.taf?abstract_id=721363; C. K. Gunsalus, forthcoming white paper from conference on "An Examination of the Interaction between Human Subject Protection Regulations and Research beyond the Biomedical Sphere," convened by the Center for Advanced Study, College of Liberal Arts and Sciences, College of Law, and the Office of the Vice Chancellor for Research at Univ. Of Illinois. See http://www.cas.uiuc.edu/ethical.html

cosmos are in need of authentic and reliable historical data. Families and communities rely on the bonds of shared memories to nurture the ties that bind them in an increasingly mobile world. Access to the artifacts and records of the past is one of the most valued functions that libraries, archives, and museums have served in the past. Today, we have only begun to consider ways to preserve the political, economic, social, and cultural record of our increasingly digital civilization.[27]

In light of this, the importance of preservation cannot be overstated. Digital data are notoriously fragile, short-lived, and easy to manipulate without leaving evidence of fraud. Preservation requires the scrupulous management of data, from ingest into a repository, through the steps of validation, normalization, storage, migration, and delivery to parties that have been authenticated and authorized to receive that data. These are complex technical procedures dependent on standards and protocols that work quickly and reliably. Preservation was once an obscure backroom operation, of interest chiefly to conservators and archivists. It is now widely recognized as one of the most important elements of a functional cyberinfrastructure.

6. Barriers created by current models of scholarly communication

Scholarly communication is a system that includes scholars, readers, publishers, and libraries. The economies involved in this system include a prestige economy, primary for scholars, important but secondary for the other players; a cash economy, primary for publishers, not very important to content producers in most cases, and important but not actually primary for libraries; and a subsidy economy, primary for libraries, who are subsidized by universities as a public good, and more important to scholars than they generally know. It should not come as a surprise that a system with three different economies at work inside it is difficult to operate successfully, but when it does work, it has a certain elegance: each party contributes from its own sense of mission, and each gets paid in its own currency. At present, though, there seems to be general agreement that the system of scholarly communication is not working—that parts of it are broken, or breaking.

The most fundamental reason for this may be that scholarly publishing—whether practiced by university presses or scholarly societies—has lost sight of its mission, and now operates primarily, and often

[27] For an overview of some of the preservation issues and literature, see Daniel J. Cohen and Roy Rosenzweig, "Preserving Digital History," in *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web.* Philadelphia: University of Pennsylvania Press, 2005. Available at http://chnm.gmu.edu/digitalhistory/preserving/.

unsuccessfully, as a financial rather than an intellectual enterprise. Market economics are not necessarily the best measure of the value of scholarship, and it may not make sense, in the larger context of the information life-cycle in universities, to conceive of scholarly communication as a market commodity rather than as a public good.

In order to operate primarily as an intellectual enterprise, scholarly communication—like research libraries—may need to operate on subsidy. It might earn that subsidy by lowering the cost and increasing the effectiveness of scholarly communication for the university as a whole. Here are some ways in which university presses might collaborate with authors and libraries to do that:

- Administering an online authoring and peer-review environment that encourages authors to produce content in forms that lower library costs for collection and preservation;
- Normalizing content produced outside that environment, to lower the cost of collection and preservation;
- Working with authors, rights-holders, and lawmakers to address intellectual property issues that make it difficult or dangerous for libraries to collect and preserve certain digital content;
- Working with the commercial sector as an advocate for scholarship, to negotiate a common understanding of the fair use of contemporary cultural materials (for example, film, television, music, etc.) in scholarly and educational contexts;
- Providing print on demand for users of free electronic resources in library collections, and managing the income from that activity;
- Marketing online scholarship to maximize its impact and its audience;
- Determining when the size of the audience merits more expensive editorial and production work, and when that work should be handled by the scholar or scholarly project.

All of these are things worth doing, in at least some circumstances, and many of them would contribute directly to the support of authoring or to lowering the cost of collecting and preserving digital content. As such, both should qualify for subsidy, in any self-interested institution—though even without that, these activities could well produce sufficient value for libraries to be paid for in the cash economy in which publishers now largely operate, if publishers were properly capitalized to retool so they could provide such services.

For obvious reasons, institutional subsidies are easiest to justify when the public good they create is consumed locally, within the institution. This has been the case with libraries, and it has not been the case with presses. University presses don't publish local authors exclusively, or even in the main, and the good they produce by publishing is produced for a global, not a local, market. It may be time for institutions to think more broadly about the system of scholarly communication as something cooperatively and consortially subsidized across localities. But even if presses are subsidized to a greater extent, and even if they cooperate with libraries and with authors, and even if they act in various ways to lower the cost of collecting digital content in libraries, developing those digital collections will be expensive, and it will be even more expensive to maintain them over time.

Even with these costs in view, locally owned and locally maintained digital collections may well be a long-term cost-cutting measure, and the key that unlocks the problem of scholarly communication. If universities *don't* own the content they produce—if they don't actually collect it, hold it, and preserve it—then they will be at the mercy of those who do. If universities *do* collect, preserve, and provide open access to the content they produce, then the entire balance of power shifts away from commercial publishing and toward university presses and university libraries.

Without retracting the argument that scholarly publishing should and could be subsidized as a public good, it is worth adding that the simplest analysis of the "crisis in scholarly publishing" is that, with average university press print runs descending into the low hundreds, it is most obviously a problem of audience: one can't afford to physically manufacture anything—books, televisions, or widgets—in lots of 500 or 1000. Perhaps, then, university presses could expand the audience for humanities scholarship by making it more readily available, online. If they did that, scholars might find an audience first, and publication in print second, instead of the other way around. And perhaps then the risk to publishers would be less, because demand sufficient for print would be demonstrated in advance. At present though, many of the most important scholarly publishers in the humanities and social sciences perceive threat rather than opportunity in the digital library. Meanwhile, by contrast, commercial publishers, having signed up most science, technical, and medical journals, are now looking at humanities and social science journals as territory worth colonizing, and they are making a handsome business of renting content to research libraries—who very rarely collect and preserve what they rent.[28]

Experiments in electronic publishing in the humanities and social sciences, and experiments in building and maintaining digital collections

[28] For example, in the field of economics in 1960, almost all 30 journals for that discipline were published by non-profits; by 2000, of 300 journals, two-thirds are published by commercial entities. Willinsky, J. [need rest of citation]

in libraries and institutional respositories, need to be supported as they move toward sustainability. According to Kate Wittenberg, the director of Electronic Publishing in Columbia (EPIC), such enterprises must "find a way in which the technical infrastructure and some aspects of workflow systems might be created centrally and then shared by a variety of projects in the humanities and social sciences. For EPIC and similar organizations, finding an answer to this challenge would be extremely valuable: making use of existing infrastructure to create efficiencies in organizations with minimal staffing." One model outside the United States is Érudit, an initiative of Les Presses de l'Université de Montréal, which offers a range of services tailored to the different types of academic publications and "is intended to serve as an innovative means of promoting and disseminating the results of university research."[29]

Significant new leadership will be required if we are to break out of the current stalemate, though—from provosts, directors of university presses, and scholarly societies as well as from libraries and individual scholars.

7. The culture of scholarship

The culture of the academy is conservative; it does not reward risk-taking in ways that other sectors do. Robert Darnton, a prominent French historian who has been active in pushing the boundaries of his domain onto the Web, remarked at the Commission hearings that the structural elements of the academy have not changed, even as the world has changed around it.[30]

A recent study of the current state of American literary scholarship online identified several cultural features among humanists that seem to militate against change.[31] Despite all evidence that "the future is digital," we have relatively few digital communities, and relatively few platforms for online collaboration. In addition, individuals continue to dominate in a new medium that invites and enables collaboration. Lone scholars, the report remarked, are working in relative isolation building their own content and tools, struggling with their own intellectual property issues, creating their own archiving nightmares.

Many have contrasted this pattern to that found among technology-intensive sciences and engineering. In those domains we see that scientists are working in "large, multidisciplinary teams of researchers in experimental development of large-scale, engineered systems. The

[29] Érudit, http://www.erudit.org/en/index.html
[30] [Darton]
[31] [Brogan]

problems they address cannot be done on a small scale, for it is scale and heterogeneity that makes them both useful and interesting."[32] In contrast to this collaborative model in the sciences, Associate Provost Stephen Brier of CUNY told the Commission, "humanists tend to be more focused on individual theorizing and communicating of ideas and information about their disciplines. Technology is not seen as a necessary, let alone a required, tool for collaboration in the humanities the way it is in the sciences."

It is interesting to compare Brier's assessment of the humanities with this passage from the Atkins report on Cyberinfrastructure in science and engineering:

> The conduct of science and engineering is a social activity, pursued by individuals, collaborations, and formal organizations. Any enlightened application of information technology must take into account not only the mission of science and engineering research but also the organizations and processes adopted in seeking these missions. A major opportunity in the [Advanced Cyberinfrastructure Program] is to rethink and redesign these organizations and processes to make best use of information technology. In fact, this is more than an opportunity; it is a requisite for success. Experience has shown that simply automating existing methodologies and practices is not the most effective use of technology; it is necessary to fundamentally rethink how research is conducted in light of new technological capabilities. Advanced cyberinfrastructure offers the potential to conduct new types of research in new ways. Doing this effectively requires holistic attention to mission, organization, processes, and technology.[33]

Most of those interviewed by the Commission expressed hope that an investment in cyberinfrastructure would defend and extend the values and contributions of the humanities and social sciences: some also expressed the hope that it might allow us to "conduct new types of research in new ways." The Web is at once too porous and too invasive to keep it out of the academy or to keep the academy out of it. Digital information technology facilitates horizontal relationships and communication, and tends to erode vertical hierarchies. Within the academic culture, methods of training and hiring and credentialing are essentially hierarchical. To take advantage of the technology, one must

[32] (Chatham, 11)

[33] 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): 14-15. http://www.communitytechnology.org/nsf_ci_report/report.pdf

engage directly with it, and one must allow traditions of practice to be more flexibly influenced by the technology. Although the ethos of the humanist is the "individual genius" working alone, collaborative humanities teams have shown us that the successful humanist can be highly collaborative.

Yet when humanists and social scientists do want to collaborate using technology, they may find their institutional leadership lacking. Many university administrators don't understand that work in these areas will have start-up costs, and so they don't budget for those costs. As UIC art history chair Bob Brugemann pointed out, "the major problem is not technological or even budgetary. It is a problem in coordination, fragmented authority, and leadership."[34]

[34] need citation

Chapter 3: Crossing the Divide: A Framework for Action

Digital technologies have already changed the nature of teaching, research, and scholarly communication in the humanities and social sciences as well as the ways that scholars talk with non-academic audiences. The Commission believes we are also on the edge of more profound breakthroughs in scholarly and public understandings of society and culture. But these further changes are not inevitable, or inevitably good. We need to act in the present to ensure this future. If we look back to other constitutive changes in scholarship and its relation to the public, we see that all such changes occurred through broadly distributed efforts within an agreed conceptual framework.

For example, in the years following World War II, the GI Bill helped transform American higher education from a system of elite access to one of mass access. Even though not originally intended as an educational measure (it was aimed at mitigating unemployment), it worked easily and effectively within the existing system by giving students the choice of where to enroll and allowing the receiving institutions the ability to determine how many of the “new” students they would accept. The GI Bill itself created no institutions nor mandated institutional behavior, but this simple means of distributing opportunity and resources dramatically expanded the number of people who considered “college” a possibility and prompted colleges and universities to see themselves as national and not local or regional institutions. Established institutions that were responsive to the new opportunities, such as the University of California, flourished.

When the federal government began the direct support of advanced research, the National Science Foundation, the National Institutes of Health, and, later, the National Endowments for the Humanities and the Arts adopted the extramural grant mechanisms pioneered by philanthropic foundations and combined them with the peer-review practices developed within universities to distribute research support on the basis of competitive applications. The competitive “market” for research support reinforced standards of scholarly excellence and relied on the research ambitions of individual scholars to motivate the institutional response of universities in developing their local research infrastructures.

The response of American higher education to the GI Bill, and the process developed by the federal government to fund advanced research, demonstrate that frameworks for action can challenge institutions to build upon existing capacities. This report suggests that cyberinfrastructure is another such framework for guiding decisions, allocating resources, and setting directions. Thinking about structures naturally requires thinking also about functions and their schematic relationship. That the National Science Foundation already has adopted cyberinfrastructure as such a framework underlines the necessity of strategic thinking. The cyberinfrastructure of the humanities and social sciences does not and will not exist independently of the larger academic infrastructure, where the sciences have set priorities. Similarly, academic stakeholders must take account of the even larger social and commercial cyberinfrastructure that is increasingly the platform on which human creativity and social interaction—the subjects of the humanities and social sciences—is expressed and takes place.

The following is a framework for action. First, we present five *underlying principles* that a robust cyberinfrastructure in the humanities and social sciences must exemplify. Second, we indicate seven *fundamental needs* that must be fulfilled to make that infrastructure possible. Each of these culminates in a specific recommendation.

Underlying Principles

An effective infrastructure for the humanities and social sciences will be built according the following five underlying principles:

1. It will facilitate collaboration

Digital technology favors openness and collaboration, from wikis to international teams. Certainly the definition and construction of the cyberinfrastructure should be a collaborative, shared undertaking involving the humanities and social sciences community in the broadest sense. But just as important is that the cyberinfrastructure needs to be designed to foster and support collaboration and sharing, perhaps most significantly when that collaboration crosses disciplinary boundaries, encourages participation by experts outside the academy, and generally brings new perspectives and new methods to bear on the exploration of the cultural record.

Although collaborative work is present in many areas of the humanities and social sciences, the prevailing practice is still for scholars to work alone. Priority should be given to teamwork and to cooperation on large-scale infrastructure-building projects that have demonstrable benefits and that scale out beyond those directly involved. Digital work has changed the “ecology of interaction” within the academy. Stakeholders should seek to foster deeper interaction among all those involved in creating and disseminating knowledge in the humanities and social sciences—scholars, librarians, archivists, teachers, technologists, and publishers.

Collaboration across institutional boundaries will also be essential. Collective action has already been effective in building shared capacities; it will be even more so as institutions of higher education confront the preservation and archiving of digital materials. Those collaborations need to extend beyond the academy, to leverage existing talent, resources, and commitment found in the academy, in the commercial sector, in government, and among the general public. It is not enough to let those outside the academy view the results of this work: each sector has already made significant contributions to the cyberinfrastructure, each has a leadership role to play, each needs to be involved in contributing further to the collection and curation of our cultural heritage.

Such a collaborative strategy will also allow us to adapt the tools and technologies already developed and deployed in other scholarly endeavors and in the marketplace. Science and engineering have already begun to scale up the parts of the cyberinfrastructure they need for their work; they have secured the interest and commitment of key funders in the federal sector. The humanities and social sciences have the enormous advantage of being able to benefit from the work already done by scientists, and to adopt and domesticate the technologies, tools, and practices that we see evolving there.

“My bottom line observation from the content perspective is that integration is key. And for that scholars need support and tools very early in the information lifecycle. In aggregate, the academy needs for them to have information description practices and standards; content, preservation, and exchange formats; information exchange and harvesting protocols; and linking mechanisms—infrastructure that is hard to build. Frankly libraries can't build that infrastructure alone, at least not if operating from or considered primarily in their historical role, which is quite late in the information lifecycle, long after the object, collection, Web site, or what have you has come into existence. For this reason the infrastructure for integration must include not only the technological, but also organizational and collaborative elements. This soft infrastructure—infrastructure that organizes and encourages collaboration between scholars, technologists, and information stewards—should be considered within the ‘enabling infrastructure’ that the commission identified early on.” John Ober, Director of Policy, Planning and Outreach, Office of Scholarly Communication, California Digital Library and Office of the President, UC-Berkeley.

2. It will support experimentation

Although the cyberinfrastructure itself should be stable and reliable, it will need to support ongoing experimentation and evolution. The call for experimentation reinforces the widely expressed need for a change within many of the fields of the humanities and social sciences. We must nurture the culture of risk-taking by creating frameworks within which junior scholars and students are rewarded for dreaming large and devising innovative approaches; laboratories where these ideas are worked out and critically assessed; and rewards for integrating the knowledge gleaned from success and failure into the next iteration. It is important to have explicit mechanisms and traditions of capturing and sharing the lessons learned, which is critical to iterative design. We need to remember that true experimentation always carries with it the possibility of failure. But informative failures are essential to moving forward into the unknown.[35]

3. It will be sustainable.

Sustainability is often thought of as primarily a financial issue: how will this project persist after start-up funding is spent? Indeed, the digital transformation has revealed questions of the financing of research and scholarly communication and preservation that previously were hidden from view in some ways by the practices of libraries and university presses. Many scholars in the humanities may have first encountered the concept of sustainability in discussions with potential funders of digital projects. We need, as Diane Zorich notes, to avoid the tendency of digital initiatives to be "treated as 'special projects' rather than as long-term programs."[36] We also need to consider that charging the full cost of a resource to its users is only one economic model for sustainability: long-term institutional, governmental, and corporate subsidy and sponsorship are also options. Perhaps more to the point, sustainability goes beyond simply paying the bills: *intellectual* sustainability requires human capital. Digital projects need to draw on a pool of trained and engaged personnel, and therefore universities need to develop the programs and the opportunities that produce people with this kind of expertise.

[35] John Unsworth, "The Importance of Failure." *The Journal of Electronic Publishing*, volume 3, number 2 (December 1997). Available at http://www.press.umich.edu/jep/03-02/unsworth.html

[36] Diane Zorich, A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns (Washington: Council on Library and Information Resources, 2003) http://www.clir.org/pubs/reports/pub118/part2.html

Projects and services will be sustained by the demand of users. But deciding when to collaborate to build economies of scale, and when to compete in order to meet and improve specific services, is not easy. Kevin Guthrie, the first director of JSTOR, remarked to the Commission that "technology creates the illusion that anyone can self-publish on the Web or do so at close to no cost at all … The marginal cost of delivering something on the Web is, indeed, next to nothing, after all. But the individual experience is not scalable: it assumes an infrastructure that is already in place and that is inexpensive. Neither is really true."

4. It will facilitate interoperability.

Access to data should be seamless across repositories, which will require standards-based tools and metadata to ensure not only interoperability but also reusability. Currently, many content providers put significant resources online, but often in data silos that have one point of entry and exit and must be searched vertically. Horizontal searching between data silos is not feasible. This disincentive for researchers merely replicates the barriers we find in the analog world with physically dispersed information that researchers can't readily assemble.

Instead, scholars, students, and the general public need to be able to read across and deeply within varied collections of texts, images, sounds, and data. Informed observers foresee the exponential expansion of our capacity to "read" across the human cultural record as "machines are programmed to index, manipulate, mine, aggregate, decompose, and build up scholarly and other forms of content by algorithm."[37] Indeed, the rise of "social software" (like del.icio.us's social bookmarking or Flickr's photo "folksonomies") makes it clear that the *use* of digital information itself produces useful digital information, perhaps most useful for enabling collaborative intelligence. We need tools and standards to achieve this vision, and new forms of collaboration among scholars, librarians, and technologists will be both a method and an outcome of building those tools and establishing those standards.

As NSF Director Bement observes, "with today's electrical grid. . . [m]y neighbor and I can use different appliances to meet our individual needs, as long as the appliances conform to certain electrical standards, they will work reliably." A sufficiently advanced cyberinfrastructure will work similarly: researchers will have "easy access to the computing,

[37] Donald J. Waters, "An Overview of Strategic Issues in Scholarly Communications: A Perspective from the Mellon Foundation"

communication, and information resources they need, while pursuing different avenues of interest using different tools."[38]

Social scientists have been especially urgent on the need for linking and sharing. As Henry Brady pointedly observed, collecting data digitally is one challenge; "an even bigger challenge is to find ways to link these data together to increase their power and utility. . . . Linkage makes it possible to make comparisons, which are fundamental to social science research. But linkage is underdeveloped in the social sciences and we need to think boldly about what we can do to improve it. . . . Social science is about comparing, measuring, and searching for patterns."

Because data generated for one purpose are used in many ways, all human artifacts are of potential value for research and learning. That means that in the digital environment, they should be created, described, and preserved in ways that facilitate reuse. The use of standards is crucial to the sharing of data, as well as to its persistence. Standards help to bring coherence into the chaos of the Web and allow one to find "actionable information" more easily. And yet, as one expert lamented, the culture of standards is weak among the humanities and social sciences. The Commission has noted the sharp tension between the recognition on the one hand that standards for, say, metadata, must be open and generalizable for purposes of interoperation and ease of creation, and the desire on the other hand for maximum expressiveness. We need robust and efficient means to generate metadata automatically as materials are created or scanned. Automating generation of metadata is an area where reducing the enormous overhead on human effort should be a priority for research and development. We also need standardized metadata in fields where nomenclature in English is less broadly used—e.g., Asian Studies, African Studies, and South American Studies—so they can be fully integrated into the unified cultural record.

The cyberinfrastructure must be constructed to be open, modular, easily updated to adapt to new technologies, and built to foster and support knowledge communities. It must serve geneticists and genealogists, historians of Buddhism and collectors of delta blues, filmmakers and dancers, those in the academy, those working in business and industry, and those home-schooling their children.

[38] Ardent L. Bement, Jr., "From Concept to Confluence: Framing our Cyberinfrastructure," Remarks, SBE/CISE CI Workshop, March 16, 2005

5. It will be accessible as a public good.

We have argued that digital information has an inherently democratizing power. But that power can only be unleashed if access to the cultural record is as open as possible—in both intellectual and economic terms, to students as well as life-long learners, across barriers of language and disability. Unfortunately, the record so far on access has been a mixed one. On the one hand, the Web has made available a welter of human knowledge for free to all—the eight million items in the Library of Congress's magnificent American Memory program is but one example. But while "information may want to be free," much digital information is not. Scholars themselves have not always been good at making the fruits of their own work available for free online. Indeed, the "open access" to scholarship movement has gained much more traction in the world of science than in the world of social science and humanities.[39] In addition, commercial entities have taken an increasingly prominent role both in digitizing public-domain cultural heritage and in digitizing cultural heritage materials still under copyright; these collections are often only available to organizations, such as major research libraries, who are able to pay substantial subscription or license fees.

Traditionally, in libraries, the largest costs have been fixed; acquiring and maintaining the space, buying the material, cataloging it, and preserving it. But the actual cost of *using* the material in the library, provided you are in the neighborhood, has always been low. The digital age exaggerates this cost structure, and makes it global: everyone is in the neighborhood. Once produced and put on a server, digital materials become public goods, and the cost of use, almost anywhere in the world, is essentially zero.

The key technical property of a pure public good is that one can add more consumers without diminishing the quantity of the good available to others. Markets are very bad at producing resources that have the characteristics of public goods, because they operate generally by charging for use. So, while there are many successful "business models" for academic libraries, there are none that look remotely like a business, because the services provided by academic libraries cannot be efficiently provided on a charge-for-use basis. One can restrict use, of course—put up a bar, charge for a look, limit use to those in the local domain—but it is very wasteful if everyone does this, because if I charge for something that costs nothing to produce at the margin, I am passing up possible

[39] See John Willinsky, *The Access Principle* (MIT Press, 2006).

value: I could make you better off while doing no harm. Efficiency is what economists love about market economies with respect to private goods—but when there are public goods, charging invariably reduces efficiency, because it reduces social welfare relative to what is possible.

Unfortunately, although public goods can be extended to more users at zero cost, they can cost quite a bit to produce in the first place. The case of digitally produced scholarship is of course an excellent example. Economic theory tells us what we ought to charge in these cases (nothing, at the margin), but it doesn't tell us how to pay for production, nor how much to produce. Economic theory does, however, tells us is that markets will underproduce, and that as a general matter, the solution of public-goods problems requires collective action. The promise of cyberinfrastructure is to provide information and communication as a public good: the Web has already demonstrated that doing so can engender new economies that operate in surprising ways and deliver benefits even economists did not predict.

Needs and Recommendations

These underlying principles will ensure that we move ahead in the right directions and the right ways. But good intentions and even good foundations are not enough. We see seven key requirements that must be met if we are to move ahead.

1. We need to nurture and validate digital scholarship and digitally literate scholars.

The conservatism of which Robert Darton spoke to the Commission has effects that are keenly felt in some academic departments in the humanities and social sciences. He was only one of many who warned that if academic leaders do not step forward to shape the digital domain, the world will shape it without them.

Such warnings reflect a widely shared perception that academic departments in the humanities and social sciences do not adequately reward innovative work in digital form. A handful of examples, recently, challenge this view, but in the most elite universities it is still traditional scholarly work that is most highly valued. As a result, those institutions (and they are often the ones who set the aspirations of a new generation of scholars) do not introduce graduate students to digital means or methods.

So, how will younger scholars in the humanities and social sciences engage these means and methods? The few will find a way on their own, but the many will need more formal venues and opportunities. We recommend the creation of brief (one- to three-week) summer workshops for younger scholars—perhaps located at some of the emergent centers in the digital humanities and social sciences—focusing on how to do research, how to present the products of scholarship, and how to teach in the digital era. One model could be the Canadian Social Sciences and Humanities Research Council's Image, Sound, Text and Technology Institute Program, which provides grants for these sort of workshops. [40] But such workshops should not neglect mid-career scholars who wish to learn about new tools, resources, and approaches. One recent workshop on digital scholarship offered only to younger scholars in one very specific domain—the history of science and technology—found itself vastly oversubscribed.[41]

Over the last 130 years, American higher education has evolved resilient mechanisms for supporting, certifying, and validating scholarship and for training new scholars. In the humanities and related social sciences, these structures are increasingly focused on the individual scholar producing books—often monographs—or articles. Means of support such as research leaves, fellowships, residential research centers, and grants-in-aid perhaps unintentionally reinforce this focus. Peer-review for promotion and tenure appropriately builds upon peer-review for publication, but by doing so may further narrow the ambit of scholarly creativity. While the scholarly community has honored several of the early pioneers of digital scholarship, basic structures of research support and evaluation change more slowly. Digital scholarship makes possible and even requires collaborative work, and produces results much more diverse than the book or article. And it requires new forms of advanced training. These new forms of work hold both practical and intellectual promise that would reward the effort to adjust the foci of current practices of research support.[42]

The same barriers that the Web breaks down between the academy and the public will also erode walls within the academy. Those who become fluent in the technologies that amplify human effort and afford new views into disciplinary subjects, will emerge as leading players, and those who do not may feel threatened, or devalued. We might naturally expect

[40] See http://www.sshrc.ca/web/apply/program_descriptions/itst/workshops_e.asp

[41] The workshop, offered by The Center for History and New Media at George Mason University with funding from the Sloan Foundation, had 75 applicants for 15 slots.

[42] For an example of an alternate form of advanced training and research support, consider the Networked Infrastructure for Nineteenth-century Electronic Scholarship, or NINES, http://www.nines.org/, described in Jerome McGann's "Culture and Technology: The Way We Live Now, What Is To Be Done?" *New Literary History* (36:1, Winter 2005), and online at http://jefferson.village.virginia.edu/%7Ejjm2f/nlh04web.htm

younger colleagues to have greater fluency and ease with new technologies, but they will also be more risk-averse, having tenure at stake. Senior colleagues have both the opportunity and the responsibility to take certain risks in this moment, and they will also be called upon to condone risk-taking in others. Times such as these, when much is in flux and the outcomes are uncertain, provide a rare freedom to make explicit and to examine our core assumptions and aspirations. The relationships between teaching and research, experiment and analysis, theory and practice, can be fruitfully explored and recalibrated. In the next decade, the successful integration of these new technologies into the humanities and social sciences will depend on those who see this time of ferment as one of opportunity, viewing the new methods and new resources as tools to achieve the deepest ambitions of scholarship and learning.

☛***Recommendation:*** *We recommend that the National Endowment for the Humanities, the National Endowment for the Arts, the Institute for Museum and Library Services, the National Academies, the major private foundations, the major scholarly societies, and the individual leaders of the humanities and related social sciences adapt current practices of research training, support, evaluation, and validation to accommodate and foster digital research, teaching, and publishing. We recommend specifically that there be:*

- *Fellowship and research leave for digital scholarship and for collaborative research projects, laboratories;*
- *Policies for tenure and promotion that recognize and reward digital scholarship and scholarly communication; recognition should not only be for scholarship that utilizes the humanities and social sciences cyberinfrastructure but also that which contributes to its design, construction, and growth.*
- *Workshops that bring together scholars and technologists around a set of goals and forge working partnerships with the computer scientists and engineers;*
- *Workshops for younger scholars that introduce them to the methods and possibilities of digital scholarship.*
- *Programs at the national level to develop shared content for teaching, and to share successful practices for teaching, both online and in person, with digital resources.*

2. We need public and institutional policies that foster openness and access.

Because humanists and social scientists study society and culture, their use of the cyberinfrastructure inevitably has social, economic, and political implications and limitations. Laws, policies, and conventions surrounding copyright and privacy are, thus, also an implicit part of the cyberinfrastructure in the social sciences and humanities. We must align current intellectual property laws and privacy policies with the new realities of digital knowledge environments. Policies and the laws that support them must take account of the characteristics of digital content and the practices that make that content productive. The recent effort of the Copyright Office to address the problem of "orphan works"—works whose copyright status is uncertain and, hence, cannot be used by scholars and others—is a welcome sign of a key agency in this debate taking an appropriate leadership role.[43] So, too, is the Library of Congress's current study of Section 108 of the copyright code and its implications for preservation.

The Commission can offer no simple solutions to complex issues of intellectual property—scholars, after all, create as well as use intellectual property and, hence, are on both sides of these contentious debates. But scholars have traditionally embraced openness and sharing, and that should continue in the digital environment. They should not be intimidated by the efforts of rights holders to restrict valid educational uses of materials. They should, for example, make full use of the "fair use" provisions of the copyright laws, which specifically cite educational use—as compared to commercial use—as a significant factor when considering if a use is "fair," or allowable without seeking permission from the copyright holder. Even the generally cautious *Chicago Manual of Style* warns against seeking permission unnecessarily: "the right of fair use is valuable to scholarship, and it should not be allowed to decay because scholars fail to employ it boldly."[44] We think it particularly important to explore more nuanced notions of intellectual property rights, supported by more sophisticated tools, so that the increasing capacity of digital technologies to mine, process, and analyze massive collections of texts not be nullified by laws intended to restrict republication. We support the work of groups like "Creative Commons" that are exploring innovative and nuanced ways to ensure the widest dissemination of works of the human imagination.

[43] For overview, see Scott Carlson, "Whose Work Is It, Anyway?" *Chronicle*, July 29, 2005, at http://chronicle.com/free/v51/i47/47a03301.htm.

[44] *Chicago Manual of Style*, 15th ed., 137.

And while scholars advocate public and legal policies of openness and access, they must similarly urge these policies within their own communities: universities need to consider the impact of their technology transfer and intellectual property policies; university presses and scholarly societies need to envision dissemination models that reflect academic values and lobby for the resources they need to live up to those values; museums need to make their digitized surrogates freely available. All parties should work energetically to ensure that the fruits of scholarly research and analysis are accessible to all those who might use them—from a student preparing a high school project to a parent trying to understand the issues in a school board debate to a tourist about to visit Rome and wanting to understand its art and architecture.

But ensuring public accessibility means more than making available materials for free on the Web to students and citizens. As everyone knows, younger students eagerly seize upon digital materials; the Web is friendly and familiar territory for them, not the *terra incognita* that it sometimes seems to their parents. But while younger students have no trouble getting on the Web, they often don't know what to do when they get there: a 1930s photograph of sharecroppers, with the imprimatur of the Library of Congress's American Memory site, may seem to be a transparent reflection of social and historical "reality" rather than a created and composed artifact with a larger political message. Students (and often their teachers, for that matter) need help in "making sense of evidence." We recommend that resources be devoted to making students (and citizens) into sophisticated and critical consumers of the vast cultural heritage that has been placed at their fingertips. Some of this can be done electronically, but workshops for K–12 teachers who use the Web in their classrooms are badly needed as well.

Social scientists have a different set of accessibility concerns that center on privacy policies. As Myron Gutman, Director of the Inter-university Consortium for Political and Social Research of the University of Michigan, Ann Arbor, told the Commission:

> Social science data have generally been collected with an assurance to participants that their identities will be kept confidential. The more complex the integration of the data, the more individual the information (especially images, geographical locations, or potentially genetic identification), the greater the risk of disclosure. . . . We need to learn how to manage these forms of integrated content so that they can be used in the future without doing harm to the individuals who were generous enough to share their experiences or their behavior with researchers.

☛***Recommendation:*** *A robust cyberinfrastructure for the humanities and social sciences will require a more flexible framework for protecting intellectual property and for making it accessible. It also will require a better legal framework for protecting data and making it accessible. Scholars, academic leaders, librarians should work with policy-makers toward those goals and they should work within their own communities to ensure the widest possible access to scholarship, research, and creativity.*

3. We need open standards and the tools to use them

For hundreds of years, the most important tools of humanists and social scientists were pen or brush and paper. Now, more and more scholars rely on a range of digital tools for research, teaching, and writing:

- curation tools, that allow specialists to extract or supply suitable metadata for catalog descriptions;
- knowledge-organizing tools that allow students and researchers to easily gather and organize their research in a digital environment and support the construction of advanced techniques for knowledge management: gazetteers, thesauri, and other controlled vocabularies;
- analytic and data-mining tools, to process vast amounts of text and data in a search for interesting patterns and anomalies—these tend to be better developed for numeric data than for qualitative data and text at this time;
- adaptive search engines, to support particular types of activity in particular disciplines;
- robust finding and filtering tools that draw upon computational linguistics and statistics, as well as upon discipline-based concepts;
- tools for processing natural language and text;
- image-processing tools that are capable of identifying specific content, interpolating missing data, and displaying and manipulating images;
- geographic information systems and related tools for handling information concerning space and time;
- tools for online peer-review and publishing;
- multilingual tools—lexical resources, morphological libraries, and normalizing and parsing tools—since one of the specific wealths of humanist scholarship is the linguistic diversity within and among documents.

Interoperability in software and in data is never perfect, but in both cases it has a better chance of emerging when information about those resources is open, easy to find, and readily re-usable. Interoperability across the humanities and social science infrastructure therefore requires the continued development and promotion of vendor-independent, open standards for document modeling and data documentation, and open-source methods for software development.

It is not clear who will build new tools for the humanities and social sciences and promote these standards, if not humanists and social scientists themselves, and their organizations. However, the academic sector is too often fragmented and constrained by the academic reward system—in spite of long-standing statements by the Modern Language Association[45], for example, it is still a difficult proposition to get tenured or promoted in an English department for work with digital media. Funding agencies are poorly resourced in the social sciences and humanities. The Commission noted a proliferation of tool-building on a local scale that appears to represent a fair degree of unnecessary redundancy among domains and between commercial and non-commercial spheres.[46] These efforts need to be coordinated; the recent summit on Digital Tools for the Humanities, which was supported by NSF and held at the University of Virginia in September of 2005, is a promising first step in this direction.

☛***Recommendation:*** *Scholars in the humanities and related social sciences should work with librarians and technologist to develop tools for producing, searching, analyzing, vetting, and representing knowledge, as well as standards for the documentation of data of all kinds. Funders, including NSF and NEH, and academic leaders should support the development and maintenance of digital tools and standards for humanities and social science research and instruction. Such support should include the development of spaces for collaboration among the toolmakers and standards-bearers, as well as the scholarly validation of these activities. NEH should coordinate this activity, but to so, it will need to recruit new expertise and new federal funding.   This effort should itself be coordinated with parallel tool- and resource-building activities in organizations like the Digital Library Federation (for example, in their Aquifer project).*

[45] http://www.mla.org/guidelines_evaluation_digital

[46] For examples, see http://echo.gmu.edu/toolcenter-wiki/index.php?title=Main_Page

4. We need centers for innovation, research, and archiving.

Humanists and social scientists have much to gain through collaboration with technologists, possibly by creating interdisciplinary labs and research groups that include both technical and subject expertise. “Once humanities faculty began using the laboratory in their research,” Computer scientist Marc Levoy of University of California told the Commission, “they would also find creative ways to fold its technology into their teaching, for example through project-based assignments in upper-level courses. This would bring humanities students into the lab, some of whom have dual backgrounds, and so could help run the lab.”

These humanities computing labs would be not just teaching spaces but also, as Provost James O’Donnell of Georgetown University explained to the Commission, “zones of experimentation and innovation for humanists, within and without traditional institutions.” O’Donnell adds that those zones should be “part and parcel of the formal academic structure. Ghettos are not the answer. We need instead the creation of privileged but open communities, where the very best young people are challenged to invent, experiment, break things, and succeed.” Some exemplary models of such centers include the American Social History Project/Center for Media and Learning, the Center for History and New Media, the Institute for Advanced Technology in the Humanities. The National Center for Supercomputing Applications at the University of Illinois has recently shown interest in arts, humanities, and social science, and their involvement in this effort would be most welcome.[47]

In addition to university-based centers, we must create regional, and in some cases, national centers for collections of various types of data. Regional data centers could become especially adept at working with states or localities to archive relevant data. They would also be charged with the task of preserving important social science data in machine-readable form. These centers require ongoing funding because they require skilled professionals who can monitor the use of data and undertake “disclosure analysis” to ensure that confidentiality is protected.[48] Funding for these centers will make it possible for researchers to use a broader range of data. The Interuniversity Consortium for Political and Social Research (ICPSR) is one such

[47] The American Social History Project/Center for Media and Learning, http://www.ashp.cuny.edu/; Center for History and New Media, http://chnm.gmu.edu/; Institute for Advanced Technology in the Humanities, http://jefferson.village.virginia.edu/; National Center for Supercomputing Applications, http://www.ncsa.uiuc.edu/

[48] See http://www.fcsm.gov/working-papers/wp22.html

national center, in the social sciences; the Library of Congress national partnerships around the NDIIPP (National Digital Information Infrastructure Preservation Program) are exploring service to other communities, using other business models.

Finally, colleges and universities should consider collective (shared, federated, distributed) archives for the safe deposit and future access of the increasing number of online journals and other materials acquired by subscription. At present, these institutions are taking a substantial risk that the "perpetual access" promised by providers will continue to be available should those providers disappear. Were each individual institution to undertake the archiving and preservation of these materials, the cost-savings from online subscription would be eradicated. However, responsibility for archiving could be shared and distributed, or centralized regionally or in some other way. The Association of Research Libraries is the organization in the best position to lead the discussion of these options, but historical societies, public libraries, including state libraries, should not be left out of that discussion.

☛***Recommendation:*** *A robust cyberinfrastructure should include a varied set of institutions and centers addressing different needs. When human, institutional, or technical resources become too expensive to replicate at every institution, it makes sense to fund a limited number of national centers for fostering cyberinfrastructure: this is what NSF has done in the sciences, it is what should be done in the humanities and social sciences as well. It is reasonable to re-evaluate, on a regular basis, the resources that merit this kind of centralization, but public funds should support national centers of excellence in digital humanities and social science, as crucial seedbeds of further innovation. Universities should continue to develop local centers for digital research and teaching in the humanities and social sciences as well. The humanities and social science cyberinfrastructure should include a network of such centers distributed around the country, including some devoted exclusively to confidential social science data.*

5. We need extensive and reusable digital collections.

The general public, students, teachers, and scholars want to have online access to the full range of primary source materials currently housed in repositories such as museums, historical societies, local libraries and research libraries, special collections, archives, and privately held collections. This would include books and journals, newspapers and magazines, government documents, manuscripts, maps, photographs, satellite images, census data, recorded sound, film, and broadcast television. Information technology also offers ways to reunite dispersed collections, as in the Dunhuang project; to compare exemplars—the Shakespeare quartos, and the many variants of the Roman de la Rose; to assemble the works of single creators, like the photographs of Mathew Brady, or to aggregate disparate examples pertaining to a single theme, such as seventeenth-century illustrations of cultivated exotic fruit in France. Despite concurrent access to more than 30 billion Web pages and annual digital output totaling many times what is held by the Library of Congress, we have really only begun to realize the potential of networked cultural heritage information.

For those materials that might never be digitized because of fragility and other reasons, there should be rich representations in online catalogs and inventories, so that users who go the Web will be able to see the full extent of resources about a given topic, whether those resources are online or not, free or not, accessible or not.

We also need to pay special attention to ensuring that those who are unaffiliated—be they collectors of digital data, photographers, artists, or performers—can add content to the Web, and retrieve content from it, easily.

The extensive digitization of cultural heritage materials is one of the most exciting developments in the humanities and social sciences in the past century. It needs both public and private support and encouragement, and we support all such efforts, whether from public entities like the Library of Congress or private corporations like Google.[49] But we believe that scholars need to be able to influence the development of private and non-profit digital archives alike, and therefore these efforts need to proceed in dialogue with scholarly, library, and non-profit communities to ensure that their fruits are accessible to the widest possible audience and that scholars can make use of them in the most effective ways. The Open Content Alliance represents an interesting emerging example of this kind of dialogue.[50]

Ambitious projects such as the ones undertaken by Google should not lead us to forget about the continued need for investment from the public and non-profit sector: one recent and carefully reasoned estimate suggests that Google Print represents only about a third of the books held in research libraries—and there are many forms other than books in which the cultural record is purveyed, and many books not held by

[49] An illustration of how commercial contribution to the cyberinfrastructure can be used is Andy Powell's discussion paper "The JISC Information Environment and Google." See http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/ie-google/ie-google.pdf

[50] http://www.opencontentalliance.org/index.html

research libraries.[51] In public and non-profit digitization efforts, priority must be placed on those collections that commerce will likely not fund, and those will likely be collections held by small and chronically underfunded museums, archives, libraries, and historical societies that are content-rich and technology-poor, such as Historically Black Colleges and Universities, which are custodians of vast and important collections documenting the lives and heritage of African-Americans.

☛***Recommendation:*** *We support efforts such as the Million Book Project, Project Gutenberg, the Open Content Alliance and other non-commercial digitization projects. Although we see a good deal of book- and journal-digitization going on, there is a considerable need for digitization going unaddressed in museums and archives, including the archives of public broadcasting (PBS & WGBH in the United States, the BBC and ITN in the United Kingdom), and we note that only the Internet Archive is attempting to preserve past states of the Web itself. We endorse efforts such as the Digital Promise Project (www.digitalpromise.org) as an imaginative and effective means of providing public support for the digitization of collections unlikely to attract commercial investment. We also encourage continued investment in this area by the National Endowment for the Humanities, the Institute for Museum and Library Services, the Andrew W. Mellon Foundation, and other funding agencies, both public and private. In addition, we recommend that scholars should cooperate with commercial digitization efforts with the goal of ensuring that they are as well-designed and as widely accessible as possible.*

6. We need to restructure the funding model for the humanities and social sciences

In nearly every venue the Commission took its investigation, the urgent need for significantly more funding and new models of financial sustainability were highlighted. We do not currently have the funding needed to build the exciting online environments and the democratic access we propose. Sustained funding is needed for content development and conversion, for preservation and curation services, for technical development, for expanding the core capacities of the individuals and organizations who make the cyberinfrastructure possible, for tools and methods to interpret and discover new meaning in this vast, new digital world, and for teaching scholars, students, and the public how to operate in it. The academy and affiliated institutions should be asked to fund certain core areas, as they have in the past: preservation and curation of cultural materials, both analog and digital; support for innovative research in the humanities and social sciences; and development of tools

[51] Brian Lavoie et al., "Anatomy of Aggregate Collections: The Example of Google Print for Libraries," *D-Lib Magazine* (11:9, September 2005), http://www.dlib.org/dlib/september05/lavoie/09lavoie.html

and resources for classroom use. Many stressed the urgency for a concerted effort by higher education to ensure that all who pass through its doors emerge as citizens of the digital world, with proven digital literacy.

Many also articulated a need for continued and increasing federal funds, with new streams of funding directed at conversion of rare and unique analog materials, audiovisual media, and other types of content that such commercial entities as Google and Newsbank will not digitize; increased funding for basic research; and increased opportunities for cross-disciplinary projects in the humanities and social sciences, and science and engineering. Several conversations focused on the need for special attention to providing substantial and sustained funding to ensure that key federal agencies such as the national libraries of medicine and agriculture, the Library of Congress, and NARA get up to capacity in the next five years for ambitious programs of creating digital content.

Not all of that funding need be government, philanthropic, or university allocations. Funding priorities include creation and conversion of content; research and development; services; curation, including the development of standards; and preservation. A mix of funding from several different sectors best serves these areas. The development of online services for education and research, from publishing to searching and presentation applications, should continue to rely on mixed models of funding, public and private, non-profit and for-profit. The mix will vary, depending on the service and the audience, but this is one area where the humanities and social sciences can leverage the work already being done in the private sector.

The very visible agreements between research university libraries and Google are but one example of this potential harmony of interests and missions. The private sector also contributes a great deal to innovation and entrepreneurial engagement with such ongoing activities as collecting and preserving, and commercial investment has often benefited scholarship and the dissemination of cultural heritage content in North America. The American Antiquarian Society, for example, the leading repository of pre-1800 printed Americana, has enjoyed a business partnership with ReadEx-Newsbank for fifty years, also involving the investment of millions of dollars. At campus-based technology and media laboratories like the Entertainment Technology Center at Carnegie Mellon, the School of Literature, Communication, and Culture at Georgia Tech, the Massachusetts Institute of Technology Media Lab, the Entertainment Technology Center at the University of Southern California, and the Institute for Advanced Technology in the Humanities at the University of Virginia, corporate supporters and partners have

Documenting Endangered Languages is a joint program between the National Science Foundation and the National Endowment for the Humanities. A single solicitation or set of guidelines has been fashioned that meets the requirements of both agencies. Applications are submitted electronically through NSF's FastLane system and then NEH and NSF staff members collaborate to name external specialist reviewers and members of a sitting panel who meet the standards of the two agencies. An NSF staff member and an NEH staff member chair the panel, meeting jointly, and afterwards direct the most highly rated proposals to normal channels in one or the other of their home agencies. Proposals funded by the NSF or the NEH are subsequently administered by that agency. Both agencies have pledged a minimum amount to support each year's competition and frequently it is possible to find additional funding as well. In FY 2005, the NSF-NEH partnership funded $4.4 million worth of projects.

played an important, often foundational role. While difficult issues will arise in shaping such public-private agreements, we find it impossible to imagine a robust cyberinfrastructure that does not include the vigorous participation of information industries.

That said, scholarship, teaching, and research are conducted as a public good and with public subsidy, either directly from the government or from tax-exempt private philanthropy. It is commonplace to observe that government funding for the humanities and related social sciences is not of the same order of magnitude as support for the sciences and engineering. Even so, some government funders, most notably NEH, while supporting digital projects have eschewed support for digital tools and other elements of the cyberinfrastructure. We hope that this report provokes a reassessment of that practice. Similarly, we hope that NSF might consider the priority attached to problems in the computer sciences that impinge directly on humanities and social science, and vice versa, and NSF should develop and sustain funding streams to support this. The recent call sponsored by the Library of Congress (LC) and NSF for proposals to do research in digital preservation called forth a volume of very sound proposals. Other areas of digital library development, metadata, etc., should be co-sponsored with federal agencies such as the Library of Congress, the Smithsonian, the National Archives and Records Administration (NARA), NSF, the National Institute of Standards and Technology (NIST), and others. We recommend that support for research on preservation and curation, more effective ways to scan, enhanced imaging, capture of audio, metadata generation, and an array of labor-saving techniques that will make what is online more usable.

Significantly, in other parts of the world—Europe, Canada, and Australia, among others—cyberinfrastructure in the humanities and social sciences is more generously funded (relative to the size of the population) than in the United States, and more unified research frameworks integrate the support of humanities and social science with the support of science and engineering. For example, the Australian e-Research framework, a significant initiative being undertaken by two government departments in Australia (the Department of Communication, Information Technology and the Arts, and the Department of Education, Science and Training) emphasizes the need for adequate infrastructure, for accessibility of data and research outputs, and for cultural change among researchers and in all research communities in order to maximize the adoption of e-Research, including the development and retention of appropriate skill bases and the engagement of industry as important stake holders.[52]

We believe that greater access to funds from NSF for work in the digital humanities will benefit both the humanities and computer science: the recent joint initiative of NEH, NSF, and the Smithsonian to fund endangered languages, described in the sidebar above, is one example that shows this is possible. We maintain, more generally, that humanities and social science research should be integrated into the emerging Grid environments, and we believe they have substantial contributions to make in developing the 'Semantic Grid' aspect of such architectures, which now tend to be semiologically simplistic. A promising example in this regard is the work done by Martin Doerr and others for building the CIDOC Conceptual Reference Model as a referential framework for describing cultural artifacts.[53]

The Andrew W. Mellon Foundation is a both a leader and a leading funder of the application of digital technologies to the humanities and social sciences and of the development of systemic means for the expansion, extension, and exploration of the relationship between scholarship in those fields and information technology. The William and Flora Hewlett Foundation, the Packard Institute for the Humanities, the Rockefeller Foundation, and others have also provided support to critical initiatives. While many other private funders have supported digital projects, few have supported the development of the cyberinfrastructure as described in this report. We hope that will change.

[52] See http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/e_research_coord_committee.htm as well as the related terms of reference summary, at http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/tor.htm, and the discussion paper, at http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/discussion_paper.htm

[53] See http://cidoc.ics.forth.gr/

Preservation is an area that demands special attention to ensure that, as a public good, it is guaranteed the fiscal and policy support that it needs. While it is not yet clear what technical mechanisms need to be in place to ensure the integrity and authenticity of digital data over long periods of time, it is clear that the models we presently use for funding archiving will not scale for digital repositories. Given the fact that much of the data and records that will concern future scholars are created in the for-profit sector, there will need to be tax incentives and other financial considerations that make it easy for the cost-effective deposit of data into long-term preservation repositories. Equally important is to develop financial incentives and rewards for those institutions that take on the mission of preserving information on behalf of society.

☛***Recommendation:*** *The development of cyberinfrastructure should be regarded as a strategic priority. Federal, state, and private funders (including the commercial sector) who support development of the broader academic cyberinfrastructure should also support projects that help develop capacity for the humanities and social sciences, in recognition of the fact that large-scale research problems always have social, ethical, aesthetic, historical, and hermeneutic dimensions.*

7. We need leadership.

The basic principles and the underlying needs are clear, but where will the leadership for this effort come from? The humanities and social science communities themselves have the primary responsibility to make the case for this effort, provide visible and sustained leadership, and offer examples for others to emulate. Senior faculty, academic officers, and boards of governors must signal that humanities and social sciences are crucial to national development of a socially useful cyberinfrastructure. They must engage technologists, non-profit and for-profit entities, funders, and policy makers. We need to work with them to forge agendas that will influence the policies made for intellectual property and privacy.

Much of the recent leadership within the academy on issues of digitization in the humanities and social sciences has not come directly from scholars, but from librarians. As the library constitutes the historic infrastructure of scholarship, it is entirely appropriate that librarians have sought to re-ignite scholarly engagement with infrastructural issues. We are now at the point, however, where others need to take up their share of the burden, and library administrators need to seek out the views of humanities scholars and social scientists when convening committees to address these issues. As the task force of the American Association of Universities indicated in its 2004 report, *Reinvigorating the Humanities*, "[u]niversity presidents, provosts and humanities deans" must "support the development and use of digital information and technology in the humanities."[54]

Leadership requires structure. Again, the example of the library community is instructive. The Association of Research Libraries (ARL), the Council on Library and Information Resources (CLIR), and the Research Libraries Group (RLG) have made technological transformation central to their mission and programming and have in turn created vehicles—the Coalition for Networked Information, the Digital Library Federation, the Scholarly Publishing and Academic Resources Coalition—dedicated entirely to providing leadership on these issues. A few cognate efforts in the humanities and related social sciences exist—the Alliance of Digital Humanities Organizations, H-Net, and the Humanities, Arts, Science, and Technology Advanced Collaboratory are three examples—but these have not had the kind of financial support from the humanities and social science communities that ARL, CLIR, or RLG have had from the research library community.

☛***Recommendation:*** *The leadership of the humanities and social sciences—not only distinguished scholars and university administrators, but also learned societies and research centers—should regard the development of a cyberinfrastructure as a strategic priority for the advancement of these fields, not as an optional "special project." The ACLS, the National Academy of Arts and Sciences, the National Endowment for the Humanities, the National Endowment for the Arts, the Institute for Museum and Library Services, as well as the National Academy of Engineering, the National Science Foundation, and key private foundations should convene a series of high-level meetings to develop mechanisms for cultivating and leadership, supporting innovation, and monitoring the development of cyberinfrastructure for the humanities and social sciences.*

[54] American Association of Universities, *Reinvigorating the Humanities: Enhancing Research and Education on Campus and Beyond* (Washington: AAU, 2004), pp. IV, 59-69. Available at http://www.aau.edu/issues/HumRpt.pdf

Conclusion

We can see the possibilities that arise from placing much of the world's cultural heritage—its historical documentation, its literary and artistic achievements, its languages, beliefs, and practices—within the reach of nearly every citizen. We can also see the value of building an infrastructure that gives every citizen not just access to this cultural heritage but the opportunity to participate in its creation and curation. We believe that a major, concerted, and structured investment in the capacities of humanities and social science scholarship to operate in the digital world will help transform these fields of knowledge and the digital world itself.

In brief, we need extensive but coherent digital collections, centers to support their users, and the tools to make use of them; we need governmental and institutional policies to support digital scholarship and new forms of scholarly communication; and we need the funding and the leadership to make it all happen, from our universities and national academies, our scholarly societies, businesses in the culture industry, and visionary individuals in all walks of life.

Appendix I: The Charge to the Commission

As scholars in the humanities and social sciences use digital tools and technologies with increasing sophistication and innovation, they are transforming their practices of collaboration and communication. New forms of scholarship, criticism, and creativity proliferate in arts and letters and in the social sciences, resulting in significant new works accessible and meaningful only in digital form. Many technology-driven projects in these areas have become enormously complex and at the same time indispensable for teaching and research.

For their part, scientists and engineers no longer see digital technologies merely as tools enhancing established research methodologies, but as a force creating environments that enable the creation of new knowledge. The recent National Science Foundation report, "Revolutionizing Science and Engineering through Cyberinfrastructure," argues for large-scale investments across all disciplines to develop the shared technology infrastructure that will support ever-greater capacities. Those capacities would include the development and deployment of new tools; the rapid adoption of best practices; interoperability; the ability to invoke services over the network; secure sharing of facilities; long-term storage of and access to important data; and ready availability of expertise and assistance.

The needs of humanists and scientists converge in this emerging cyberinfrastructure. As the importance of technology-enabled innovation grows across all fields, scholars are increasingly dependent on sophisticated systems for the creation, curation, and preservation of information. They are also dependent on a policy, economic, and legal environment that encourages appropriate and unimpeded access to both digital information and digital tools. It is crucial for the humanities and the social sciences to join scientists and engineers in defining and building this infrastructure so that it meets the needs and incorporates the contributions of humanists and social scientists.

ACLS is sponsoring a national commission to investigate and report on these issues. The Commission will operate throughout 2004, and is charged to:

- Describe and analyze the current state of humanities and social science cyberinfrastructure
- Articulate the requirements and the potential contributions of the humanities and the social sciences in developing a cyberinfrastructure for information, teaching, and research

- Recommend areas of emphasis and coordination for the various agencies and institutions, public and private, that contribute to the development of this cyberinfrastructure

Among the questions to be explored in pursuing these three goals are:

Describe and analyze the current state of humanities and social science cyberinfrastructure.

1. What can be generalized from the already significant digital projects in the humanities and social sciences? Which humanities and social science communities are most active and why? Of those that are not, which might soon, easily and/or profitably, engage more deeply with digital technology? How have those scholars developed computing applications to accomplish their scholarly and expressive goals? Where have they failed to do so, and what can be learned from those failures?
2. What new intellectual strategies, critical methods, and creative practices are emerging in response to technical applications in the humanities? To what extent are disciplines in the humanities transforming themselves through the use of computing and networking technologies? What are the implications of that transformation?
3. What organizations and structures have empowered or impeded the digital humanities? What are examples of successful and durable collaboration between technologists and humanities scholars? Where and how are people being trained to support and engage in such collaborations? What has been the role of libraries, archives, and publishers in these projects?

Articulate the requirements and the potential contributions of the humanities and the social sciences in developing a national cyberinfrastructure for information, teaching, and research.

1. What are the "grand challenge" problems for the humanities and social sciences in the coming decade? Are they tractable to computation? Do they require cyberinfrastructure in some other way?
2. What technological developments can we predict that will have special impact in the humanities and social sciences in the near future?
3. Which are the most important functionalities necessary for new research and development in cyberinfrastructure generally? What kinds of humanities or social science problems are theoretically difficult or expressively complex, or challenge our ability to formulate a computable problem in some other way? What kinds of

humanities or social science problems are computationally intensive, require especially high bandwidth, or present resource challenges in other ways?

4. What are the barriers that confront humanities and social science users who wish to take advantage of state-of-the-art computational, storage, networking, and visualization resources in their research? What can be done to remove these barriers?
5. What impact will the availability of high-performance infrastructure have on enabling cross-disciplinary research? What will high-performance infrastructure mean for the broader social impact of humanities and social sciences?
6. What can be done to improve education and outreach activities in the computer-science and engineering community to broaden access to high-end computing? How can computing expertise in the humanities and social sciences themselves be increased?

Recommend areas of emphasis and coordination for the various agencies and institutions, public and private, that contribute to the development of humanities cyberinfrastructure.

1. What investments in cyberinfrastructure are likely to have the greatest impact on scholarship in the humanities and social sciences?
2. What research infrastructure should be coupled with cyberinfrastructure?
3. How can private and public funding agencies coordinate their efforts and cooperate with universities, research libraries, disciplinary organizations, and others to maximize the benefits of cyberinfrastructure for the humanities and social sciences?
4. How should new investments in infrastructure and technologies be administered so as to include the humanities?