Head in the Clouds and Boots on the Ground: Science, Cyberinfrastructure and CLIR

Amy Friedlander
Council on Library and Information Resources

May 1, 2007

Thank you for inviting me to participate in this program and to discuss some of the recent developments in science and the cyberinfrastructure, primarily in the United States, as well as CLIR's contributions to these discussions. We will spend a few moments reflecting on the trajectory of scientific research and the role of the computationally intensive systems collectively known as the "cyberinfrastructure," before we go on to talk about CLIR's agenda at the convergence of these two topics. I should also preface my comments with a few caveats and qualifications. First, trends in the nature of research and their implications for libraries and our sibling institutions, archives and museums are taking place in the context of a broad re-structuring that consists of at least three related elements:

- a re-thinking and re-organization of the system of higher education in the U.S. over the next 30 years, including support for advanced research, systems of scholarly communication and relationships with the private sector;
- a re-thinking of notions of literacy, how it is measured and what it means to be literate; and
- the internationalization of the conduct of scientific research. Scientists have shared their results and interpretations for millennia. In the last 30 years, collaborations and cooperation has begun to occur earlier in the research process. This phenomenon is particularly obvious in large scale scientific experiments and in the construction of shared databases, such as the Protein Data Bank.

Second, within these concurrent, seismic processes lies a set of interactions and feedbacks with advances in information technology such that the information technology, in its many expressions, is both cause and effect. In this context, it is useful to think of the information technologies as creating a computationally rich environment with a set of properties and affordances rather than as a specific set of technologies (for example, networking or word processing) or toolkits (for example, geospatial, statistical or text analysis packages). Third, I should also observe that making generalizations about a subject as large and heterodox as scientific research is always a perilous business. Much of what I will say about scientific research might be said of research more generally. Moreover, my comments are U.S. centered – because my professional experience has been almost exclusively in or about the U.S. There are almost certainly many counter-examples and nuances that could be introduced to qualify some rather broad assertions. But let us go forward nonetheless, recognizing that these remarks are analogous to a view of earth from the moon.

Cyberinfrastructure: e-Science, Big Science and More Science

The term "cyberinfrastructure" originated in a report by the U.S. National Science Foundation (NSF) where it is defined as the comprehensive infrastructure required to capitalize on advances in information technology, which "integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools."[1] The American Council

[1] National Science Foundation, Cyberinfrastructure Council, Cyberinfrastructure Vision for 21st Century Discovery (March 2007), p. 6.

of Learned Societies subsequently adopted the term in its report on a cyberinfrastructure for the humanities,[2] and the word has crept into routine discourse in higher education and advanced research. Historically, roots of this idea extend back to the development of computer networking in the 1960s and advances in high performance computing in the 1980s that enabled both distributed research teams and access to data and other resources as well as computationally intensive analysis in a range of fields from computer assisted design in engineering and manufacturing to spectroscopy and the construction of massive databases of highly granular information in the life sciences, social sciences and physical sciences. In parallel, the term "E-Science" originated to name a broad research initiative in the United Kingdom, which was formally organized in 2001,[3] and has come to mean computationally intensive research that is executed in distributed network environments or involves large quantities of digital data and is frequently conducted in research teams.[4] Thus, "e-Science" and "cyberinfrastructure," while closely allied, actually represent slightly different emphases: The former connotes a research program; the latter recognizes the role that the engineered and institutional infrastructure plays in instantiating and fostering that research and devotes resources to cultivating that set of systems. This is an important distinction, and one that helps us understand the roles of libraries, archives and museums in supporting and nurturing the research enterprise.

The progress in the sciences toward a shared infrastructure is fairly well understood, at least in its broad form. There are several drivers: One is simply the cost of instrumentation and the limitations to broad distribution of these expensive facilities. Supercolliders, telescopes, sensor arrays and so on are expensive to build, limiting the number that can be reasonably supported. In addition, there are constraints on the number of suitable sites for such facilities. For example, Vassar College built an astronomical observatory in the 19th century for pioneering woman astronomer Maria Mitchell on its campus in Poughkeepsie, New York, 75 miles north of New York City in the Hudson River valley. Prime observing sites for modern observatories, as Sayeed Choudhury can tell us, have fairly stringent requirements, and there are a relative handful of them, Hawaii, Chile, northern Japan, and so on. Satellite-based earth observing systems have similar constraints, where downlinks can be located and how monitoring and observations can be coordinated on a 24/7 global basis, thus driving international collaborations across several facilities.

Cost and geography are two fundamental constraints. Speed of light is a third. Complex manipulation and rendering of very large datasets require the capabilities of a supercomputer or cluster. Moreover, while the network does enable computational tasks to be distributed, some problems result in co-location of resources, and some kinds of problems cannot be parsed into sub-tasks that can be cleanly distributed. Increasingly, there is discussion about different configurations of high capacity computational resources and facilities: grid computing, cloud computing *and* supercomputer centers with high speed lines affording access to a distributed but typically limited community of investigators. These four fundamentals – cost, geographicand environmental requirements, speed of light, and character of the research problem -- together with advances in networking set up a tension between centralization and distribution evident in large scale system architectures and in the social organization of the conduct of the research. Thus, the e-Science (or the research program) acts and is acted upon by the cyberinfrastructure, and the cyberinfrastructure has

[2] American Council on Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, Our Cultural Commonwealth (2007).

[3] About the U.K. e-Science Programme, http://www.rcuk.ac.uk/escience/default.htm

[4] Association of Research Libraries, Agenda for Development E-Science in Research Libraries, November 2007, p. 6; also e-Science, Wikipedia, last modified 2 February 2008; http://en.wikipedia.org/wiki/E-Science.

been evolved and will continue to be evolved to enable the e-Science. And the systems will be both centralized and distributed.

On the one hand, the rise of so-called "big science" is obvious both in the scale of the instrumentation, such as the Large Hadron Collider (LHC) at CERN or the Stanford Linear Accelerator (SLAC) outside of Palo Alto, California, and in the organization of large research teams. On the other hand, networked access to resources, in particular to standardized data in ditial form, has allowed for distributed research. In 1997, the State of Sao Paulo Research Foundation in Brazil (FAPESP) established a network of 30 research laboratories to form a virtual institute devoted to genomic projects, beginning with examination of pathogens in plants of local economic significance, notably citrus fruits and sugar cane. The research agenda subsequently expanded to include livestock and human cancer genomic research, and the ONSA network now comprises more than 60 participating institutions, including international partnerships.[5]

These two examples – CERN and FAPESP – put faces on the phrases, "globalization" or "internationalization" of the conduct of science." As of the end of 2006, CERN counted over 11,000 (11,046) paid and unpaid employees, researchers, and users drawn from more than 20 counties, organized into 13 on-site departments or units.[6] The six experiments at CERN engage international teams that range in size from 1700 scientists from 159 institutes in 37 countries on the ATLAS experiment to 22 scientists from 10 institutes in 4 countries on the LHCf (Large Hadron Collider forward) experiment.[7] The researchers may not be physically present at CERN all of the time, but they do depend upon the experiments that take place at the facility. FAPESP, on the other hand, is a distributed social and organizational system of bench science. Large scale coordination is enabled by the physical network; by consensus within the professional community on the structure of the data, namely the proteins, genomes, and so on; and by access to core datasets around which the conduct of the science has become organized and where the science itself is advanced through contributions to a unified resource of highly granular information. Thus, the infrastructure to support research is both centralized and distributed, and both physical – lines, nodes, equipment, and so on – and informational, encompassing both the logical layer that integrates the disparate hardware and the conceptual structure of the biological data itself that allows simultaneous storage, access, distribution, and analysis by diverse researchers who share a single set of databases.[8]

Earlier engineered infrastructures tended to be conceptualized primarily in terms of their physical representations as roads or telephone lines with their associated organizational and informational control systems. Moreover, there is a relatively clean distinction between operation of the infrastructure and services that were built because the infrastructure existed. An obvious example in the U.S. context was the collocation of stockyards, grain elevators and flour mills with railroad junctions and canal facilities. In contrast, the cyberinfrastructure includes a data layer as an integral component.[9] Now, what we mean by "data" and the ways

[5] The State of São Paulo Research Foundation, http://www.fapesp.br/english/materia.php?data%5Bid_materia%5D=297

[6] European Organization for Nuclear Research, Human Resources Department, CERN Personnel Statistics 2006 (March 2007), https://hr-info.web.cern.ch/hr-info/stats/persstats/PersonnelStats2006.pdf

[7] CERN – The LHC Experiments, http://public.web.cern.ch/Public/en/LHC/LHCExperiments-en.html; CERN – The LHC Experiments: ATLAS, http://public.web.cern.ch/Public/en/LHC/ATLAS-en.html; CERN - The LHC Experiments: LHCf, http://public.web.cern.ch/Public/en/LHC/LHCf-en.html

[8] Note that this single set of database is unitary from the perspective of content; the databases themselves may be replicated and mirrored in several locations to afford access and to enhance security of the content.

[9] NSF Cybinfrastructure Council, March 2007, Chapter 3: Data, Data Analysis, and Visualization (2006-2010). Christine Borgman elaborates on this point in her recent monograph, Scholarship in the Digital

in which data will be managed remain an important discussion. But when data that is independent of the control systems required to manage the engineered network become part of the infrastructure, then the roles of information managing entities, such as libraries, archives, museums, corporate data centers, and so on, assume a larger and more explicit role operation in the infrastructure itself. Rather than being services and resources collocated with the infrastructure, libraries, archives, data centers and similar entities become integral to the operation of the infrastructure as users experience it and expect to use it. Perhaps not all of these agencies' traditional functions will migrate to the infrastructure, but some subset of them will, thus begging more questions: Which services are infrastructure services? Which ones are not? And are all libraries and collecting institutions equally vital to the success of the research enterprise?

Implications for Libraries

Arguably, libraries, archives museums, professional societies and so on have always been part of the organizational infrastructure that has supported research and education, articulating practices, standards and codes of conduct that knit together various components. Interlibrary loan, MARC records, catalogs and abstracts of manuscripts, professional certifications, finding aids, indexes, and ISO standards all speak to codified practices that are intrinsic to modern research. More generally, librarians, archivists and scholars have reached broad consensus on the functions associated with such institutions: They collect, preserve and manage information made available to patrons under conditions that range from completely open to highly restricted access based on the nature of the material, community expectations, and local mores.

Conceptually, infrastructure systems are both hierarchical and scalable so that they meet local conditions but possess overall coherence and interoperate to obtain ubiquity, shareability and broad access. Libraries and their various counterparts embody precisely this set of characteristics: They exist in many languages, meet a range of local or specialized needs, and manage a welter of information artifacts, yet they have formal and informal mechanisms for interacting and exchanging information and training across institutional, geographic and political boundaries. In short, you can probably go into almost any library in the world and recognize where you are even though you also quickly understand that there are differences. And as my examples of CERN and FAPESP imply for research generally, libraries that support education and research are having to learn when to consolidate and centralize, when to distribute tasks, and how to maintain large scale coherence at multiple scales whether within the home college or university or as part of much larger systems. We can see evidence of this tension already in the intense discussions taking place around journal archiving, retention of physical copies, and access to licensed material. In the future, such traditional measures as gate count and physical collection size may become less useful as ways to gauge impact than metrics that capture intangible usage.

The introduction of digital technologies has greatly amplified this tension between the local and the system wide. The cumulative effect of three developments: more power at the desktop; access to high bandwidth networks linking researchers to each other and to other resources; and the existence of key resources such as the scientific datasets and the electronic journals, has been to push functions hitherto associated with the library toward the end user and to increase demands on the library's stewardship and long term data management responsibilities. Students and faculty arrive on campus with fairly high expectations about the extent to which devices and programs will be supported, and campus administrators face

Age (Cambridge: The MIT Press, 2007), see especially, Chapter 6, Data: Input and Output of Scholarship.

the hard choices about what is supported as part of the campus infrastructure.[10] Although the library is much loved and greatly respected, it has receded from view. Surveys of U.S. faculty and librarians commissioned by Ithaka in 2006 found that "in the future, faculty expect to be less dependent on the library and increasingly dependent on electronic materials."[11] Just as technology has de-coupled content from its artifactual form, so too has information from the perspective of these users become separated from the institutions that make it available – the publishers, aggregators, libraries, and so on – even while the apparently "free" access to a journal may be financed by a site license that the library or the university has negotiated.

My point is not to argue open versus restricted access but rather to show that the logic of the technology has been to push functionality to the desktop and to curtain off from that desktop a maze of systems and decisions. Indeed, recent developments in Web 2.0 services, cloud computing, APIs, and increased mobility have pushed some capabilities back onto shared services to which powerful yet lightweight devices will presumably have reliable and secure access. Therein lies a paradox: The empowered user is, in fact, enmeshed in and dependent upon an interlaced technical and organizational environment. And the operation of that environment – like the operation of the library – is essential yet unseen.

So what do libraries and other memory institutions do? Certainly they provide integrative functions and they become, more than ever, stewards of the collected knowledge. The same survey that found little support among faculty for librarians as gatekeepers also found widespread support for the library's preservation function. The importance of data curation, archiving and preservation functions to the research library was a recurrent theme in a conference at CLIR on the future of the research library in the 21st Century in late February 2008,[12] and the solicitation for Sustainable Digital Data Preservation and Access Network Partners (DataNet) at the U.S. National Science Foundation Office of Cyberinfrastructure placed substantial importance on the contributions of library and archival science, calling for organizations that "integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise" to provide reliable and long data management, meet user needs and expectations, support research, and serve as components in an interoperable network of data preservation and access.[13] Ominous predictions of a data deluge within the scientific community are almost commonplace, and reports from market intelligence firm IDC on the growth of digital information primarily from the standpoint of the data storage community echo and amplify these statements. The most recent report projected a compound annual growth rate of information between 2001 and 2006 of about 60 percent and an excess of data over storage capacity in 2007, as predicted in an earlier report. Together with more regulatory requirements for information retention in the commercial sector, these developments, the

[10] Joel M. Smith and Jared L. Cohon, Managing the Digital Ecosystem, Information Technology and the Research University, Issues in Science and Technology (Fall 2005). http://www.issues.org/22.1/smith.html

[11] Roger Schonfeld and Kevin M. Guthrie, The Changing Information Services Needs of Faculty, Educause Review (July/August 2007), p. 9. http:// connect.educause.edu/Library/EDUCAUSE+Review/ TheChangingInformationSer/44598?time=1205866916

[12] Core Functions of the Research Library in the 21st Century, February 27, 2008, http://www.clir.org/ activities/registration/08R21.html

[13] U.S. National Science Foundation, Office of Cyberinfrastructure, Sustainable Digital Data Preservation and Access Network Partners (DataNet), last updated, January 8, 2008. http://www.nsf.gov/funding/ pgm_summ.jsp?pims_id=503141

authors conclude, put "greater pressure on those responsible for storing, retaining, and purging information on a regular basis."[14]

Storing, retaining and purging sound a lot like appraisal, selection, acquisition, and weeding – functions well known to librarians and archivists. While perhaps the visibility of the reference function may be diminished (and I am not entirely sure that it is), the collection management functions, including curation, preservation, and archiving, are increasing substantially in volume and difficulty. The information is highly heterogeneous, combining analog and digital as well as a range of digital formats, standards and platforms. The technical issues of preservation are well known: hardware and software obsolescence of content, operating systems, applications, storage and access devices; error identification and management; intellectual property requirements, and data confidentiality and security. Users, however, expect data to be accessible and interoperable, and distinctions between "data" and "information" or "books" and "manuscripts" or "primary" and "secondary" sources, not to mention "libraries" and "archives" seem to melt away. But not entirely.

Different regulatory and access regimes are still tied to such distinctions, and although revisions to the copyright statute in the U.S. are slowly lurching forward, these terms do not simply denote an arbitrary container for sequences of bits. They convey tangible and intangible information meaningful to researchers that applies to use of the objects and the trust they instill. A biography of Jane Austen is simply not equivalent to her letters even if both are rendered in digital form, and a publication in a peer reviewed journal carries weight independent of the pre-print or conference paper that might have become available six months earlier. More problematic are new kinds of objects, like websites that are intentionally dynamic or may not be wholly self-contained, posing challenges for the librarian or archivist who seeks to enable users 10 or 15 years hence to re-experience objects as authors intended and to understand the context in which to employ or interpret them. Finally, data, unlike publications, frequently carry restrictions to protect personally identifiable information or to restrict access to data, like the locations of archaeological sites or mineral resources, that are deemed sensitive and therefore considered confidential. Scientists, moreover, whose publications may rest on unique access to their hard-won data, have often proved reluctant to release that data to more general use. Thus, the notion of re-using and re-purposing data is not widespread in scientific research and the value of preservation (not to mention the ease with which it might be accomplished from the desktop) may be admired in the abstract but not necessarily relevant to an investigator.

Not surprisingly, then, building the digital collections for the future requires attention to discovery and access, and active management of digital archives is more than keeping track of formats, platforms, and machines. In particular, requirements for interoperability have both technical aspects and implications for discovery and use. Especially among the social and behavioral scientists, it is unlikely that designers of individual systems will know ahead of time all the kinds of information that an investigator might wish to assemble and analyze with a single set of tools. Epidemiological research, which may engage data from the cellular to the societal as well as studies of climate change, public health, environment and ecology are obvious examples of integrative domains (or systems science) in which highly heterogeneous information ranging from 16th century travelers' accounts to 21st century sensor measurements might be relevant. Although librarians might not be the researchers' gatekeeper, they are most certainly most likely to be well positioned as recommenders, to understand the range of information, how to get to it, and how to use it in a blend of skills formerly associated with cataloging, reference and subject specialties so that the information

[14] John F. Gantz, Chrisopher Chute, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgang Schlichting, Anna Toncheva, The Diverse and Exploding Digital Universe; An Update Forecast of Worldwide Information Growth Through 2011(IDC, March 2008), pp. 2, 3, 4.

is seamlessly there for the end user who does not need to know all the magic taking place behind the curtain.

CLIR's role

The Council on Library and Information Resources (CLIR) is the successor organization to two older entities: The Council on Library Resources, formed in 1956, and the Commission on Preservation and Access, organized in 1986. CLIR itself was created in 1997 and inherited from both parent organizations a commitment to preservation, access and stability and management of research library collections. With the appointment of Charles Henry as president in 2007, CLIR's programmatic agenda has been organized into six major related and mutually reinforcing topics:

(1) Cyberinfrastructure: *What are the technical and organizational systems, services and relationships required to support an extensible, scalable network of data and services?*

(2) Preservation: *What is required to manage data for effective use over the long term?*

(3) Digital Scholarship: *How does the new environment allow us to ask new questions not otherwise possible?*

(4) Emerging Library: *What will the role of the library be in the future?*

(5) Leadership: *How do we prepare students for a future in libraries and information, and what will a career path look like over the course of a professional lifetime?*

(6) New Models: *What kinds of organizations and frameworks should we build in the U.S. and abroad?*

In addition and in pursuit of a coordinative role in the developing cyberinfrastructure to support higher education and advanced research, CLIR seeks formal and informal partnerships and collaborations with other entities. Major efforts have been directed toward working with the federal agencies, notably, the Institute of Museum and Library Services, National Science Foundation, National Endowment for the Humanities and Library of Congress, where successful collaborations have been achieved either by contributing to and participating in their programs; obtaining awards for funding under one of their grant programs; or creating a joint project. CLIR also seeks to support graduate students where possible by articulating well-defined tasks that can be handed off to students, thus providing support, integrating them into our program and hence into the profession, and building partnerships with the leading schools of library and information science. Such arrangements have been put in place with University of North Carolina at Chapel Hill and with University of California at Los Angeles. These informal efforts amplify our programs directed toward students and post docs (Zipf, Rovelstadt, Mellon and CLIR post doctoral fellowships) and toward future library leaders (Frye Institute).

Historically, CLIR's interests have resided more in the humanities than in the sciences and social sciences. In the last year, however, we have increased our participation in programs that emanate from the scientific research agencies. These take two principal forms: participating directly in the research agenda, in particular, where decisions concerning data management draw on library and archiving expertise; and convening activities across agencies and disciplines where there is common cause to be found among a broad swath of researchers. Not surprisingly, these are both complementary avenues into a matrix of related issues known variously as digital scholarship, preservation and cyberinfrastructure, which are three of our program areas. The very fact that it is so difficult to separate the strands speaks to the tight integration of the substance of the scholarship and the environment of data, services and systems on which that scholarship is predicated. That is why these topics interest us and why we believe that the roles of libraries, archives and other stewardship institutions are simultaneously intuited to be important yet remain difficult to parse. That ambiguity is likely to remain with us for some time to come, precisely because the larger context of scholarship and higher education is itself unsettled.

With respect to the first dimension of our work with the basic science research agencies, the flagship effort is probably our support for the NSF Blue Ribbon Task Force on Economically Sustainable Digital Preservation and Access. By now, it is fairly well understood that more data is created by computationally intensive science than can be easily managed and that over the long term, management of that data cannot be born by the research agencies. In addition, valuable data are created outside of the scope of traditional research by entities that have no apparent motive for the long term preservation of this material.

For example, economists and sociologists who are studying information technology deployment in general, have trouble finding suitably detailed information to complement the aggregate data compiled by the U.S. federal statistical agencies, which also tend to be very conservative in their sampling and analytical methodologies. The limitation in access to data affects certain microeconomic studies of firms, salaries and wages, and pricing. One solution is using private sources of information, but these are expensive and can be limited by issues of confidentiality and proprietorship. Leading scholars have employed data supplied by the business intelligence and marketing company Harte-Hanks. But not only are the datasets are expensive, but they naturally reflect the interests of the company's clients who have paid for the initial surveys. The content of the data files is geared toward marketing and not necessarily toward the questions that investigators may have, and there is no guarantee that the data will be preserved, constraining both long term longitudinal comparisons as well as the ability of later scholars to validate earlier results by re-examining the data.[15]

This is a huge problem. Validating prior results, whether it takes the form of re-running the experiment or checking the sources, is central to scholarship across the disciplines and goes to the heart of the trust model that supports advanced research. Thus, data preservation is integral to computationally intensive research, whether in the science or the humanities, and libraries, archives and museums, as stewards of the data are lynchpins in the infrastructure that supports and enables the conduct of the research and managing its products in whatever form those products shall take. The realities may be somewhat harsh: data management outstrips the capacity of the basic research agencies to pay for it indefinitely; important data lie outside the scope of the federal agencies; and owners of the data may have not obvious interest in sustaining the data indefinitely or in providing access to it. The Task Force does not propose to address all of these issues, but it does propose to come to terms with some of the basics, namely, how much does it cost? And who should pay? Among the early conclusions is a recognition that there will be a life cycle in the curation of information, which the LIFE project in the UK has also concluded, as the various interests of the owners and custodians of data evolve. Data may be initially preserved to comply with regulatory requirements, such as the Sarbanes-Oxley Act, which governs financial and accounting information held by institutions subject to federal regulation. But when the period mandated by Sarbanes Oxley expires, there may exist an historical value in the information and hence a public interest, which might lead to a transfer of custody of the material with necessary safeguards to protect confidential and proprietary data.

[15] Kenneth Flamm, Amy Friedlander, John B. Horrigan, William F. Lehr, Measuring Broadband: Improving Communications Policymaking through Beter Data Collection (Washington, DC: Pew Internet & American Life Project, 2007), pp. 19, 22.

Under the rubric of convening symposia and workshops to identify common research challenges are two events. The first took place in late November 2007[16] and addressed what Gregory Crane of Tufts University has called the "Million Books" problem. When text corpora become very large, in some measure as a result of mass digitization projects, only the computer can "read" the text. Working effectively with these collections begins to push computer science research in the areas of multilingual services (embracing both information retrieval as well as machine translation), semantic disambiguation (in multiple contexts ranging from individual and place recognition to more abstract meaning), and document structure. Thus, we begin to see how a common agenda arising from the convergent interests of humanists and computer scientists is possible with the active engagement of librarians and archivists who are responsible for management and long term viability of the data.

We are continuing to examine the shape and form of this convergent agenda in a second workshop co-sponsored with the National Endowment of the Humanities for which two program officers of the National Science Foundation and representatives of the Institute of Museum and Library Services, the Coalition for Networked Information, and the Andrew W. Mellon Foundation are advisors. Again, the domains represented are those traditionally associated with the humanities: history, art history, literature, and so on, but the research challenges are posed as much for the computer scientists as for the humanists and as one social scientist told me, these questions could be asked equally well of his community. The goals of the one-day symposium are two fold:

- to explore how advanced technologies enable new, deeper and richer analysis and interpretation of text, video, sound and other forms of creative expression traditionally grouped under the rubric of the humanities and to understand the resulting implications for the research agenda in the information technologies; and

- on the basis of that discussion to formulate questions and topics that may represent the convergence of research issues and are distinct to the digital environment.

Underlying these goals is the acknowledgement that technology development is not a linear process of basic research, technology transfer, applications and deployment. Rather, the application layer itself is iterative and dynamic, and research entails engagement and re-engagement with diverse user communities, especially user communities that pose hard problems. Moreover, as the information universe expands, many of the research techniques associated with the humanities will flow over into the sciences as investigators navigate the millions of pages in pre-print archives, journals, lab notes and so on.

Finally, earlier this year (March 2008), the Andrew Mellon Foundation awarded CLIR a significant grant to support a competition for cataloging so-called "Hidden Collections." This generous grant addresses a problem that the research library community has studied for about a decade, namely the existence of unprocessed, uncataloged and essentially undiscoverable yet tremendously valuable materials held in the special collections of libraries, archives, museums, and historical societies. Estimates of the sizes of these so called "hidden collections" vary from 15 percent of the printed volumes in university special collections, to an average of 27 percent of manuscripts, and 35 percent and 37 percent for video and audio respectively, according to surveys conducted by the Association of Research Libraries.[17] The goal of CLIR's five-year program is to create a distributed, multi-institutional web-accessible

[16] Many More than a Million: Building the Digital Environment for the Age of Abundance, November 28, 2008, http://www.clir.org/activities/digitalscholar/Nov28final.pdf.

[17] Winston Tabb. 'Wherefore Are These Things Hidden'? A Report of a Survey undertaken by the ARL Special Collections Task Force, RBM vol. 5, no.2. (Fall 2004): 123-126.

catalog to materials of scholarly value that are presently inaccessible except to those specialists who may stumble across them in their course of their research or who may be directed to them by skilled librarians and archivists.

The grant is for one year with four one-year renewals, based on our performance in the first year. We expect to open the competition in June with awards in the fall. Right now, the eligible collections must be owned or held in the U.S., but in future years, we hope to expand the scope to include international participants. We are completely open on format, technical platform, and schema, requiring, however, that the resulting records be findable by future search engines and hence compliant with existing standards and protocols. Primary weight will be given to the research merit of the collections and then to innovation in cataloging and description that will enable increased discovery and access. Finally, CLIR will not hold the cataloged records in a centralized repository, although it is possible that a third party aggregator may assume that responsibility. Rather, we insist that the collections that own the materials must also own the catalog records and accept responsibility for the sustainability of the web-accessible catalog.

We view this project as both a research experiment and an element in the evolution of the cyberinfrastructure to support research. The reasons are clear: We expect to learn a lot about efficient cataloging of rare materials, which has long been a time consuming and labor intensive process. We will probably learn more about how organizations build cyberinfrastructure through cooperation and shared resources while minimizing the free rider problem that has historically plagued centralized construction and delivery of infrastructure systems. Finally, the purpose of research infrastructure is precisely that: to support, enable and foster research for which discovery of information is critical. In particular, we have hopes that rare materials will be findable and that a series of small collections may form a critical mass of information that may be highly relevant to various small science fields like enthnography and cultural linguistics where the scale of the projects have traditionally been focused on the individual investigator and the research collections frequently end up in archives, where they may be perceived more in terms of their biographical value than in terms of current science. In this sense, the Hidden Collections effort is similar to the institutional repository movement where, again, the research collections of students and faculty might be collected and preserved at the local level and but found, re-used and re-purposed in future investigations that span multiple collections.

I recognize that the discovering and re-purposing small collections that are widely distributed over many institutions is more a dream than a reality and much needs to change before that vision becomes commonplace, namely, agreed upon and interoperable metadata, distributed search engines, tools and policies to support deposit so that archiving truly does begin at creation, plus a mind-set among researchers that appreciates the value of preserved information to current research and research designs that rely on access to such information rather than on newly collected data. It requires a cultural change in the way that much science is done. But attaining the vision speaks to CLIR's purpose: to clarify the issues, ask hard questions, convene the right people in the right room at the right time to articulate and pursue common causes, and in the end to do our part in the long transformation of higher education and advanced research in which we find ourselves. At CLIR, we work with a very broad range of people from library directors and chief information officers at small liberal arts colleges who balance the costs of a new roof for the gymnasium against investments in the IT infrastructure to senior researchers at the basic science research agencies, who envisage the possibilities thirty years from today. Although my head may be in the clouds and my heart is in research, my boots, like CLIR's, are always on the ground.

Thank you.