Illinois Digital Scholarship: Preserving and Accessing the Digital Past, Present and Future

A White Paper prepared jointly by
The University of Illinois Library and CITES
University of Illinois at Urbana-Champaign

Michael Grady, CITES
William Mischo, Library
Beth Sandore, Library

7 April, 2004

Executive Summary

Since the University's establishment in 1867, its scholarly output has been issued primarily in print, and the University Library and Archives have been readily able to collect, preserve, and to provide access to that output. Today, technological, economic, political and social forces are buffeting all means of scholarly communication. Scholars, academic institutions and publishers are engaged in debate about the impact of digital scholarship and open access publishing on the promotion and tenure process. The upsurge in digital scholarship affects many aspects of the academic enterprise, including how we record, evaluate, preserve, organize and disseminate scholarly work. The result has left the Library with no ready means by which to archive digitally produced publications, reports, presentations, and learning objects, much of which cannot be adequately represented in print form. In this incredibly fluid environment of digital scholarship, the critical question of how we will collect, preserve, and manage access to this important part of the University scholarly record demands a rational and forward-looking plan—one that includes perspectives from diverse scholarly disciplines, incorporates significant research breakthroughs in information science and computer science, and makes effective projections for future integration within the Library and computing services as a part of the campus infrastructure.

This report recommends that the campus take action now to do two things: 1) create a reliable and easy to use repository service to preserve, manage, and provide persistent and widespread access to the digital scholarship faculty and students now produce; and, in parallel, 2) initiate with faculty, students, departments, and colleges the discussions that will enable them to make changes in publication models that involve institutional and disciplinary archiving and the retention of their own copyright to their scholarship, thereby maintaining the authority of scholarship within their respective disciplines.

The greater part of this report focuses on the development of a repository service and its technical underpinnings. We recommend that the Library and CITES serve as trusted agents in the development and implementation of this service, recognizing also that there will be a number of rich opportunities for technology research collaborations with units on the UIUC campus as well as the Chicago and Springfield campuses. To accomplish this goal, the University Library and CITES pledge $1.3 million of in-kind and cash resources over a six-year period, and we request an equivalent amount of support from the campus to implement the repository, for a total investment of $2.6 million over six years.

In this report we recommend that the initial collection efforts focus on digital materials that do not pose copyright or other intellectual property issues. However, we strongly urge that the University now begin to address the more challenging issues associated with developing new models for faculty and institutional ownership and widespread access to their own peer-reviewed digital scholarship. If the campus chooses to address these challenges, scholars here and at other academic institutions will reap much greater intellectual rewards in the long-run. We urge the campus to consider seriously this proposal to catalyze a faculty-driven initiative to re-shape

scholarly publishing and the mechanisms used for its dissemination, as well as a much-needed understanding of the role of technology in digital archiving.

Challenges and Benefits

The University Library and Archives have been responsible for collecting, preserving and providing access to the scholarly output of the University since its inception in 1867. The majority of this output has been in print. With the advent of digital scholarship, faculty and students at academic institutions world-wide are re-defining their output relationships within the world of scholarly communication (with publishers and professional societies). They increasingly post publications, working papers, and research reports on Web sites, or make their preprints available in digital form through professional society Web sites. Scholars, academic institutions and publishers are engaged in debate about the impact of digital scholarship and open access publishing on the system of recognizing significant research in the disciplines, as well as on the promotion and tenure process. Publishers are examining their value-added role in the refereeing and editing process, as well as issues related to the ownership, management of, and access to the archival record of digital scholarship that is created by individual scholars. Further, all parties seek reliable, permanent places to archive these digital publications, as well as the supportive information for this research--datasets, instructional materials, field notes and interviews, performances and creative works, interviews, simulations--all of which comprise the scholarly record of a career, regardless of institutional affiliation, as well as the provenance of the research.

The upsurge in digital scholarship has left the Library with no ready means by which to archive digitally produced publications, reports, presentations, video, audio, and learning objects. It is insufficient in most cases to preserve only the print version of a digital work because increasingly print cannot represent the interrelationships among documents that can be created using digital works. In this incredibly fluid environment of digital scholarship, the critical question of who is responsible for collecting, preserving, and managing access to this important part of the University scholarly record demands a rational and forward-looking plan—one that includes content from diverse scholarly disciplines, incorporates significant research breakthroughs in information science and computer science, and makes effective projections for future integration within the Library and computing services as a part of the campus infrastructure.

In this paper we outline a collaborative plan aimed at enabling the University of Illinois to preserve, manage, and provide access to the digital works and learning materials created by scholars on the University of Illinois at Urbana-Champaign campus. This report recommends that the Library and CITES serve as trusted agents for the University in the joint development and management of a repository service for the campus, recognizing also that there will be a number of rich opportunities for collaboration with units like GSLIS, Computer Science, Electrical Engineering, and NCSA that have the potential to greatly enrich the repository service model. We also recommend that the campus initiate with faculty, students, departments, and colleges the discussions that will enable them to play a key role in shaping publication models that involve institutional and disciplinary archiving, thereby maintaining the authority of scholarship within their respective disciplines.

The first phase of this initiative would involve the development of a digital repository as a proof of concept that would provide a suite of underlying services. This model would be developed using existing repository software that is available from other institutions under open-source licenses. The development of a testbed repository architecture would present numerous opportunities for externally funded applied and basic research in data mining, secure knowledge management, information architectures, information retrieval, and metadata creation and processing. The lessons learned in the first phase would provide a basis for a second phase, where a quality production service would be developed, along with the definition of requirements to make the service permanent. The third and final phase of this project would involve institutional broad-based commitment and permanent adoption. Our long-term target (six years) is to produce a useful service that is widely deployed and actively used across campus. In achieving this goal, the University Library and CITES would pursue collaborations with a variety of partners on campus, and also the Chicago and Springfield campuses, including content providers and technology developers such as the Graduate School of Library and Information Science, NCSA, Computer Science, and Electrical and Computer Engineering. We have also initiated a conversation with the UIC Center for Data Mining, which can prove fruitful in future phases of the proposed work with research datasets.

Setting in place a process to preserve digitized and born-digital research and other individual and institutional output will require several programmatic activities:

- Involving faculty in determining what types of output ought to be included in a repository;
- Working with faculty to develop new organizational models for preserving and providing access to their peer-reviewed publications and other forms of scholarship;
- Conducting a systematic survey across campus of colleges, units, and programs that could potentially contribute to an institutional repository;
- Developing criteria for the selection and inclusion of digital content;
- Differentiating between access and preservation in repository setup and digital object life cycle;
- Developing methods to simplify the deposit, description, and location of materials within the repository;
- Collaborating with publishers and government agencies in content preservation efforts and metadata harvesting activities.

The benefits of a repository service for digital scholarship are many, as already outlined by a number of peer institutions. The University of California eScholarship program[1] cites a number of benefits that would be directly transferable to the University of Illinois environment:

- **Free to the University of Illinois**: Research units, centers, or departments would be able to use this technology to make their publications widely accessible and to ensure that they would be preserved in digital form.

[1] The University of California eScholarship Repository. "Repository benefits." URL: http://repositories.cdlib.org/escholarship/benefits.html.

- **Promising alternative** to commercial ventures or self-publishing.
- **Permanence** from the University's commitment to maintaining persistent access to content that is stored in the repository.
- **Increased visibility** of faculty research and the department or unit. The repository would bring many new readers to the content, and to the related faculty or unit's web site(s). Persistent links to the publication as well as the related faculty or unit web sites would be provided as part of the repository service.
- **World-wide accessibility** using the Open Archives Initiative (OAI protocol for metadata harvesting). This protocol would make the content discoverable from a variety of locations with no extra work on the part of the author.

The Work Plan in the final section of this paper provides more information about the ways in which we propose to address the above activities. Clearly, librarians and archivists cannot achieve the goals represented in a repository effort alone—rather this must be a concerted effort involving scholars and their research, as well as the expertise of technologists and information science researchers, and institutional policy makers. This is as much a social change as it is a technical development. The goal of enabling the creation of scholarship in preservable form will require leadership at all levels, across the subject domains. The development of a repository for digital scholarship will provide the capability for the University to play an integral and new role in the distribution and provision of access to faculty and student scholarship. With the technical capability will come the concomitant need for faculty discussion across the disciplines to develop selection as well as access policies, and to determine what level of significance the repository will play in the dissemination of scholarly output apart from traditional publishing channels.

We approach this endeavor with the knowledge that simply building a service for the deposit of digital scholarship will not ensure that it will be used by faculty and students for the purpose of preserving and providing access to their works for the long-term. To be more specific, faculty and student investment in the concept must be based on the assumption that the content in the repository has undergone some sort of widely recognized vetting process, in order to ensure its value to one or more discipline, and to the University. For this reason, this report recommends that campus-wide discussions focusing on how to get high quality scholarship into the repository need to take place simultaneous to the work on building the service that Library and CITES propose here.

Background

Defining Systems and Methods for Preserving Digital Scholarship

In her recent work *New Model Scholarship*, Abby Smith warns that academic institutions are in danger of not being able to preserve important digital scholarship across the disciplines because the digital documents and media that faculty and students develop cannot in their current forms be preserved by librarians and archivists with the tools that we currently possess. The questions posed by Smith in her recent report are being echoed throughout academic libraries and archives world wide:

> *How do we know what the value of these digital objects is and may be decades hence?*
> *How do we anticipate and address the technical needs of fragile digital objects over time?*
> *Who is responsible for preservation, and how is it financed?*

Smith points out that while most scholars rely on librarians and archivists to collect, preserve, and provide access to important resources upon which they base their research, the practice of digital scholarship has changed the interdependencies in this traditional model, placing the burden on the scholar for the creation, delivery, and management of "preservable" digital objects:

> "…the task is not only to invent tools that foster productive use of the Web as a medium of scholarship and teaching but also to create material in preservable form."

Further, once digital objects exist in a preservable form, they need to be archived in some type of system that will allow the content of the files to be managed and accessible so that it can be used over time, regardless of the software application that must be used in order to view, interact with, or otherwise experience the digital content. Recently the phrase "institutional repository" has emerged to describe the handful of software systems (both open source and commercially produced) that are geared at the archiving and long-term management of digital content. Some of these products are well-known and have been developed by academic institutions or by professional societies to address the very problem that is outlined in the preceding paragraphs (e.g., DSpace, co-developed by MIT and Hewlett-Packard; FEDORA, co-developed by the University of Virginia and Cornell University; EPrints, developed by a faculty member at the University of Southampton).

As Smith and others involved in digital archiving have indicated, librarians and archivists cannot achieve this goal alone—rather this must be a concerted effort involving scholars and their digital output, as well as the expertise of technologists and information science researchers, and institutional policy makers. This is as much a social change as it is a technical development. The goal of enabling the creation of scholarship in preservable form will require leadership at all levels, across the subject domains.

Institutional Repositories: Definition and Significance

In the ARL Bi-monthly newsletter of February, 2003, Clifford Lynch defines the role of repositories in the academic setting, and he identifies groups who are responsible for implementing them, as well as the concerns and caveats that institutions must bear in mind when developing these structures. Lynch views institutional repositories as "…a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members." Lynch emphasizes, however that the institution's commitment to the stewardship of these materials is perhaps more important than the actual service model:

> "[A]n institutional repository is a recognition that the intellectual life and scholarship of our universities will increasingly be represented, documented, and shared in digital form, and that a primary responsibility of our universities is to exercise stewardship over these riches: both to make them available and to preserve them. An institutional repository is the means by which our universities will address this responsibility both to the members of their communities and to the public. It is a new channel for structuring the university's contribution to the broader world, and as such invites policy and cultural reassessment of this relationship."

Lynch also reinforces the points that institutional repositories must be easy to use or contribute to, and that once a repository is established, faculty, staff, and students would view it as both an essential and continuing commitment by the institution to the stewardship of digital materials of enduring value.

> "Faculty who choose to rely on institutional repositories to disseminate and preserve their work are placing a great deal of trust in their institution and in the integrity, wisdom, and competence of the people who manage it. We need to ensure that our institutional repositories are worthy of this trust. "

Institutional Repository Development Efforts

Once created, digital collections can be daunting to manage. Simply storing discrete digital objects in the computer's file system and providing access to those objects through hand-made web pages or a manually maintained database may be an adequate strategy for smaller projects, but it is not a practical model for larger collections. With larger collections come a number of new problems: multiple communities of users, complex relationships among digital objects, compound digital objects, shared behaviors and other types of object-class attributes.

To address these issues, there has been in recent years considerable investigation into and development of digital object repositories—but even so, digital object repositories are still in early stages of development, their architectures are still being specified, and as a community, we have relatively little experience with them. Projects are already under way in a number of our peer institutions.

According to a report issued by Mark Ware in January 2004 for the UK-based Publisher and Library/Learning Solutions (PALS) group, there are approximately a dozen digital object repository software systems that use different hardware and software platforms as well as different operating procedures and strategies. While a few commercially-developed long-term archiving solutions exist (e.g., Documentum[2]), the majority of software developed in the past several years is freely-available as open source, and has been developed and implemented in either academic or not-for-profit settings. Some systems (e.g., EPrints[3]) focus on enabling institutional self-archiving of publications and working papers. EPrints was first made available for download in 2001, and it is reported to be the most widely used repository system. Others

[2] URL: http://www.documentum.com

[3] URL: http://www.eprints.org

(DSpace[4], FEDORA[5], Greenstone[6]) provide a mix of functions that enable the archiving of text, video, audio, and other media. Early reports from our peer institutions suggest that one repository system may not fit all needs, and that different formats of digital content (text, audio, video, still images, data sets, simulations, etc.) may require management using different repository tools.

DSpace is a digital repository system that was developed jointly by MIT Libraries and Hewlett-Packard to capture, store, index, preserve, and redistribute the intellectual output of a university's research faculty in digital formats. DSpace is now freely available to research institutions world-wide as an open source system that can be customized and extended. Subsequent funding from the Andrew Mellon Foundation in 2003 has supported the D-Space Federation, a group of seven institutions implementing DSpace and participating in its further development (Cambridge University, Columbia University, Cornell University, MIT, Ohio State University, the University of Rochester, the University of Toronto, and the University of Washington).

The Fedora project, jointly developed by the University of Virginia and Cornell University, was funded in 2001 by the Andrew W. Mellon Foundation to build an open-source digital object repository management system based on the Flexible Extensible Digital Object and Repository Architecture (Fedora). The new system demonstrates how distributed digital library architecture can be deployed using web-based technologies, including XML and Web services, and it supports such applications as institutional repositories, digital libraries, content management, digital asset management, scholarly publishing, and digital preservation.

The use of institutional repositories and the breadth and depth of their content have been topics of considerable speculation. The PALS report by Ware surveyed approximately 45 institutional repositories in existence world-wide that provide information about their holdings in a standard format, using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The survey revealed that the total number of documents on the 45 sites was approximately 42,700, which were divided roughly into the following categories:

- 22% e-prints
- 20% theses and dissertations
- 58% other documents—including "grey literature"—technical reports and working papers—and a collection of digital images.

The repository implementations to date have clearly focused their efforts on collecting the "grey literature," e-prints for which there are few or no rights issues, and theses and dissertations. Ware also noted that no research datasets were found in this survey, although colleagues at several early adopter institutions indicate that they preparing to address this issue. The subjects covered by the 45 repositories surveyed by the Ware report include physics, mathematics, computer science and economics, with small amounts of documents in linguistics, philosophy and some humanities.

---

[4] URL: http://www.dspace.org

[5] URL: http://www.fedora.info

[6] URL: http://www.greenstone.org

Open Access Publishing and Archiving Peer-Reviewed Scholarship

The results of the Ware report, as noted above, are somewhat disappointing in terms of the slow uptake and use of repositories among academic and other research institutions. One of the fundamental drawbacks of institutional repositories from the faculty perspective is that many do not contain peer-reviewed scholarship, and therefore they do not represent significant research archives. The Budapest Open Access Initiative (BOAI) is to date the major international movement, supported by the Soros Foundation, that serves to promote the provision of open and free access to the refereed scholarly literature.[7] At its core, the BOAI proposes that scholars and research institutions world-wide form an alliance to make peer-reviewed journal articles and other support materials (unreviewed preprints, working papers) freely-accessible through the Internet:

> By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

Proponents of the Open Access Initiative suggest that the provision of free access to peer-reviewed journal literature ought not to be equated with "costless" production, but that the true cost of online production of these materials is far less than current pricing suggests,[8] and that scholars, along with academic institutions and research organizations, now ought to work together to reduce production costs as well as increase accessibility of their works.

Preparatory Technical Work

Several activities currently under way on the Urbana campus are aimed at exploring the development of digital archiving capability. The University Library and CITES began a series of conversations late in 2002 about building a suite of repository architectures that could preserve, on a long-term basis, UIUC scholarly research output, significant educational materials, and historically significant institutional events and information. These conversations culminated in an agreement between the two units to develop this white paper. Further, Library and GSLIS faculty have been engaged in externally funded research projects that focus on areas where prerequisite knowledge would be required--repository creation, automatic metadata generation, and data mining. Through informal contact we know of faculty in other units who are engaged in related research. The proposed work has the potential to encourage collaborative research to solve a number of challenges that have yet to be addressed.

---

[7] Budapest Open Access Initiative: http://www.soros.org/openaccess/index.shtml.

[8] Odlyzko, Andrew. "The Economics of Electronic Journals." *First Monday: Peer Reviewed Journal on the Internet*. Vol.2 No.8 - August 4th. 1997; URL: http://firstmonday.org/issues/issue2_8/odlyzko/index.html

One of the outcomes of these joint conversations is the understanding that we need to develop a reliable technical infrastructure to support the management, storage and delivery of materials that are in current use, as well as addressing the long-term storage and preservation of scholarship and learning materials that are deemed to have significant institutional value. The ability to share and re-purpose (within and beyond the University of Illinois community) scholarly and learning content is a concurrent need that must be considered in constructing a repository. Work on various aspects of the technology that supports inter-institutional sharing is under way in a number of major regional and national research and educational support organizations, including the NSF Internet 2 Middleware initiative, the Committee on Institutional Cooperation's (CIC) interest in creating a repository for Native American Indian materials, and the Digital Library Federation's (DLF) recently announced Distributed Open Digital Library (DODL) initiative. Building a robust institutional archiving program is already a critical factor in the University's ability to support faculty participation in these cutting-edge initiatives.

In the fall of 2003, the University Library, GSLIS, NCSA and several external partners, including OCLC, an alliance of seven state libraries, and several academic institutions, submitted a proposal to the Library of Congress NDIIPP (National Digital Information Infrastructure Preservation Program) to support a three-year grant to develop automatic data harvesting methods and to test them with current open-source digital repository architectures (FEDORA, D-Space, Greenstone) and one commercial system (OCLC Digital Archive). These awards will be made some time in the early spring of 2004. Should the University receive this award, the proposed research, evaluation, and tool development would significantly advance any local institutional repository efforts the campus might initiate.

The University Library has also been working with NCSA and NARA (National Archives and Records Administration) to develop automatic data mining and extraction methods for full-text archival documents (e.g., email). These techniques will be of critical importance in making it easy for faculty, students, and staff to participate in and contribute to an institutional archiving service that represents rich locally-developed content. Both NARA and the San Diego Supercomputing Center (SDSC) have collaborated with increasing funding to develop storage and retrieval models for long-term data archiving, and we believe that there is significant potential for the University and NCSA to explore data archiving and data mining partnerships with the SDSC.

Related Campus and University Efforts

The proposed effort can leverage the current work of several initiatives that are implementing technologies that are either key to an institutional repository's operation, or will serve as convenient conduits for depositing content. There are two technology building blocks that must be in place for an institutional repository to be implemented—secure, flexible, and reliable storage, and robust identity management. In addition to this, an institutional repository will have an impact on the planning for future information architecture and networking. A potential campus portal, and the Illinois Compass learning management systems could serve as highly visible and easy-to-use conduits for depositing content into a repository. The specific CITES services and projects that might be leveraged are NetFiles, Illinois Compass, the directory

services effort, a possible portal pilot, cross-campus efforts to better align IT strategies and leverage resources across the entire University, and planned networking infrastructure upgrades.

NetFiles is the recently deployed centrally supported file storage system for students, faculty. Individuals have their own file storage area, and can control access to their files through a web browser or a specific software interface. The focus is on space for individuals, although providing file space for "groups" (e.g. a Registered Organization, a research group, a unit) is under consideration.

Illinois Compass, the Urbana campus enterprise deployment of the WebCT Vista learning management system, will provide the framework, service platform, and delivery mechanism for online resources for courses. This will include both traditional (timetable) and non-traditional courses.

The campus and University have also been investigating the needs and requirements for a portal deployment, including several committees that have gathered much input from significant constituencies, and have evaluated portal software alternatives. A portal pilot effort beginning some time in the next few months seems likely. An IR implementation need not be linked to a campus community portal. However, the efforts could be coordinated to work seamlessly--the portal could serve as one of several convenient points of entry for searching the content in the institutional repository. A campus portal could serve as one of several high profile points where faculty, staff and students could submit content to the repository.

Through the use of NetFiles and Compass, faculty and students have the potential to build up large collections of digital content. Similarly a portal effort is also likely to encourage faculty, staff and student creation and sharing of digital content.

Moving from ubiquitous file storage to the secure and well-managed storage environment will require both planning and new resources. *None of the services discussed above provides information retrieval or management functions—either short-or long-term*. While most of these digital objects will not be of long term institutional value, some number will be of enduring value to the campus.

An institutional repository deployment will require a sufficient middleware infrastructure to support it, particularly in managing and controlling access and access rights. CITES is planning to re-design its directory services infrastructure to provide more timely and accurate data that is used to authorize appropriate access to services and systems at both the campus and unit levels. Key goals are to create a flexible, scalable, framework that is lower in management cost than the current structure, and to support key interoperability standards and a rich variety of standard information access protocols.

Further, there are a number of current cross-campus committees investigating ways to better align our IT strategies and leverage our diverse resources. The Common Architectural Vision and Road Map (CAV) committee and the Data Centers committee are particularly of note in relationship to an IR effort. The coordination of cross-campus data management, storage and networking efforts could be the key to a successful and reliable institutional repository deployment that is both effective and

disaster tolerant. Now is an appropriate time to investigate the needed infrastructure to support the storage systems that would ensure long-term viability of digital objects managed within an institutional repository.

Needs Assessment and Recommended Starting Point

The experiences of MIT and the other DSpace implementers, and the results of the PALS survey suggest that there are compelling needs for archiving a number of categories of digital scholarship, including peer-reviewed journal publications, working papers, research reports, web sites, databases and datasets, theses and dissertations, audio and video of performances and creative works. While the list may seem endless, virtually all institutions currently implementing institutional repositories have taken a similar approach that has targeted print publications for their initial efforts, following in subsequent phases with multimedia objects, research datasets, and other materials comprised of complex formats.

A recent informal needs assessment carried out in the fall of 2003 by the Library suggests that the initial pilot study ought to focus on collecting scholarly output—and more specifically, the "grey literature"-- publications, reports and working papers that emanate from scholars and programs at the University. This category would include publications from centers, institutes, or initiatives with an outreach (public, scholarly) component that publish (or self-publish) on a consistent basis their research or promote their work using print and electronic publications. This group recommends that the pilot study focus on preserving published digital materials, primarily full-text documents (encoded with a standard schema or not encoded), html documents, or Adobe Acrobat .pdf documents, where access rights have been cleared. This would include discrete works that fit the specific metaphor of a "publication."

An informal survey of the UIUC Web pages yielded a number of examples of initial target areas for seeking document contributions to a digital archive. The Web site "Research Centers, Institutes and programs" provides a starting point that includes both web sites and publications for campus units (http://www.publications.uiuc.edu/info/research.html). There are also a number of college or departmental publications and technical reports that summarize or provide in-depth information about research programs (e.g., Summary of Engineering Research-- http://www.engr.uiuc.edu/Publications/engineering_research/2003/);

Further conversations with peer institutions indicate that it must be easy for faculty to identify the objects that are deposited, otherwise the repositories will not be utilized effectively. This suggests that we need to develop the means to make it easy to generate the information about objects, whether that be routines that extract and generate metadata automatically, or the use of desktop tools that simplify the process of description and deposit for faculty, students, and administrators.

Although we recommend that the initial testbed repository development be oriented toward textual materials, we recognize that a fully-developed repository service would need to expand in later phases to accommodate a variety of digital content in standard formats, including video, audio, still images, computer simulations, and numeric data. In particular, because of the increased requirements by federal agencies related to data archiving, research data sets for federally funded projects

ought to be considered for inclusion in a repository as soon as it is feasible. We recommend that the scope of the first phase be limited to contain costs, but also to be able to evaluate whether the repository adequately meets a limited set of requirements before expanding its functionality and scope.

One of the difficult paradoxes of preserving digital scholarship is the fact that the material that is at the highest risk is that which is often the most difficult to preserve (e.g., multimedia materials, research datasets, performances, simulations.) In our investigation of the ground-breaking work on establishing institutional repositories, we have found that most institutions have made the initial investment in text documents, for which reasonable digital preservation guidelines already exist. For this reason, the University could adopt the perspective that our organizational investment in an institutional repository not be one that is self-contained, but rather one in which we seek to develop a network of partners with expertise in the preservation and management of different types of digital content—geospatial datasets, video and audio, encoded texts, still images, etc. The University is in a unique position to forge partnerships with NCSA and the San Diego Supercomputing Center that have the potential to enhance the digital preservation services we can offer to the University of Illinois community, and we have initiated informal conversations with these organizations to investigate the development of a common agenda for the preservation of digital scholarship.

We also wish to acknowledge that one of the oft-stated goals of institutional repository systems is to collect and archive the locally produced scholarly works of the institution's faculty and students that are at present typically published in refereed journals and conference proceedings. Indeed, institutional repositories have been proposed as an alternative scholarly communication infrastructure to the present publisher and professional society based scholarly publishing system that is responsible for the dissemination and archiving of research and scholarly literature. However, this has broad implications for promotion and tenure and raises questions with regard to copyright and intellectual property rights and institutional responsibility for multi-authored works. Up to this point, institutional repository systems have had limited success in attracting the journal/conference scholarly works, and the role of institutional repositories within the evolving scholarly publishing model remains an open question. One of the recommendations of this report is that the campus take up this question and look to identify faculty in those disciplines who are both willing and interested to make substantive changes in their approaches to producing peer-reviewed scholarly publications. There are many potential solutions to this challenging problem, and not all disciplines will arrive at the same solution. Although the MIT DSpace implementation did not begin with peer-reviewed journal publications, MIT is experimenting with a mechanism called "journal overlay" that tracks the actual publication in a peer-reviewed journal of a document that was deposited in the DSpace repository prior to publication.

Work Plan

This report recommends that the Library and CITES jointly develop and manage a repository service for the UIUC campus, recognizing also that there will be a number of rich opportunities for technology research collaborations with units like GSLIS, Computer Science, Electrical Engineering, and NCSA that have the potential to greatly

enrich the repository service model. In this model, the Library would serve as the campus agent and point of contact for developing the repository and its content, and CITES would provide the support for scalable storage solutions, a flexible identity management framework (authentication and directory services), and advice on integration with related campus systems, including a portal and the Illinois Compass learning management system. We envision that a campus advisory group would oversee the development of policies governing the deposit of content into the repository. At this stage, there is not a substantial amount of cost data available from the implementations at peer institutions. Early reports from MIT, which is now in the second full year of their DSpace repository implementation, as well as informal discussions with peer institutions, suggest that full implementation and widespread participation will require a multi-year commitment. Based on the facts that UIUC represents a similar, distributed environment, with a substantial focus on research both within and across a variety of disciplines, we have outlined a six-year work plan, which is detailed below. The budget detail for the project is included in Appendix A.

Several common needs have emerged from the information we have gathered from peer institutions that are implementing one or more institutional repository software systems in campus-wide efforts:

- Governance: The project must receive guidance from faculty groups whose discussions determine the core content that is included in the repository. These groups would also provide advice on user needs, policy, and operations;
- Coordination: The implementation requires someone to coordinate the contribution of content from academic units to the repository;
- Technical support: The project requires dedicated technical support to implement the repository software and to scale up the pilot to a production service;
- Storage solutions: The computing centers will require additional resources to develop flexible, scalable, and reliable storage solutions;
- Reduce contribution barriers: Technical resources must be devoted to developing methods that make the contribution of content as simple as possible;
- Find content easily in the repository: The project needs to develop automatic methods for capturing and generating metadata—information that describes the digital objects in the repository. This will enhance our ability to manage the information and make it more accessible to the user community.

Phase 1: Duration--2 years

The first phase of the proposed project will focus on two critical components, which we recommend be carried out during the same two-year time frame: 1) developing the underlying infrastructure of an institutional repository service for the campus; and 2) conducting campus-wide discussions focusing on how to get high quality scholarship into the repository. The focus will be on identifying content, formulating collection development, selection, and appraisal policies and submission standards, selecting and configuring the system(s), coordinating how we get digital objects into the repository, demonstrating the function of the system, preparing documentation, and evaluating

the pilot study. A number of specific activities will take place during this time, including the following:

- Campus-wide discussions with faculty to determine what types of output ought to be included in a repository;
- Appoint groups to advise on policy, content selection, and operational activities;
- Early adopter faculty groups/departments/colleges develop new organizational models for preserving and providing access to their peer-reviewed publications and other forms of scholarship;
- Library discussions of the role of the Library and Archives in providing a repository, and the development of a set of working principles and policies for content selection and workflow, in conjunction with campus faculty discussions;
- Conduct a systematic survey across campus of colleges, units, and programs that could potentially contribute to an institutional repository;
- Identify content to be included in pilot study; provide user support to enable the submission of the content and generation of metadata to discover the content in an online retrieval system;
- Focus on the development of underlying or "core" services:
  - Implement one or more digital repository systems;
  - Develop capabilities for University of Illinois pilot groups to submit and access materials in the repository;
  - Host and preserve pilot faculty materials—identify needs for expanding this capability;
  - Identify a baseline for creating ongoing support for UIUC contributors, monitor and back up systems, respond to user questions and suggestions;
  - Create data entry templates and mechanisms to make metadata creation simple for contributors
  - Develop and test automatic metadata extraction schemes to simplify metadata creation and information discovery and retrieval in the repository;
- Develop training and informational programs about the purpose and use of the repository for students, faculty, librarians, CITES, and staff in campus units who are involved with the content submission process;
- Develop and refine evaluation questions:
  - Will this solution work across disciplines to support preservation of and access to digital scholarship?
  - How do we evaluate the ways in which faculty use these systems and how do we determine what kind of finding tools and archiving functions are both valued and useful services?

Phase 2: Duration—2 years

Activities:

- Build and expand campus-wide quality production service for digital materials based on outcomes from Phase 1;
- Expand the scope of the services policy on content selection to include additional formats, based on ongoing priorities identified by the community advisory process;

- Encourage externally funded research projects using the testbed;
- Develop a service and a cost model for managing large-scale research datasets and multimedia content (incorporate the cost of preserving research data, where required by sponsors or desired by researchers, into grant proposal budgets at the campus level);
- Investigate a potential service and a cost model to support open access digital archiving for partner institutions;
- Report on the viability of the repository service model, selection policies, systems, refresh schedule.
- Investigate fundamental digital archiving issues, including semantic and functional migration requirements, version control, rights management, scholarly communication issues.
- Investigate partnership arrangements for a digital archiving network—archiving back-ups, content sharing.

Phase 3: Duration 2 years

Activities:

- Define partnership requirements for a digital archiving network.
- Develop cost and service models for ongoing support.
- Institutional commitment to permanent service
- Test and evaluate available and sensible solutions to fundamental digital archiving issues.
- Develop ongoing service level agreements.
- Review storage model and revise plans where necessary.

Background Readings:

Barton, Mary R. and Julie Harford Walker. *MIT Libraries' DSpace Business Plan Project: Final Report to the Andrew Mellon Foundation*. (July, 2002) URL: http://libraries.mit.edu/dspace-fed-test/implement/mellon.pdf.

Lynch, Clifford A. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," ARL Bimonthly Report 226 (February 2003), 1-7. URL: http://www.arl.org/newsltr/226/ir.html.

Smith, Abby. *New Model Scholarship* Council on Library and Information Resources (2002) URL: http://www.clir.org/pubs/abstract/pub114abst.html.

Ware, Mark. *Pathfinder Research on Web-based Repositories. Final Report*. Publisher and Library/Learning Solutions (PALS) Group, January, 2004; URL: http://www.palsgroup.org.uk/.