Review of Digital Repositories August 22, 2005 Integrated Library Technology Services, Research and Planning Yale University Library

Contents:

- I. Introduction
- II. Four Factors Influencing the Development of Repositories
- III. What is a Digital Repository?
- IV. Multiple Digital Repositories within a Tiered Architecture
- V. Assessing Needs and Participation Potential
- VI. Recommendations
- VII. Appendices
 - 1. Digital Repository Examples
 - 2. Bibliography
 - 3. Outline of themes covered in digital repository examinations

I. Introduction

In September 2004, the Integrated Access Council <u>http://www.library.yale.edu/iac/</u> requested that the newly formed Integrated Library Technology Services Research and Planning group (ILTS Research and Planning) undertake a research project to investigate emerging digital repository models. This study was commissioned in an effort to develop a better understanding of how digital repositories fit into the planning for Yale University's future information technology infrastructure. Several significant environmental trends and internal campus developments made the year-long investigation timely.

The environmental trends that have contributed to the development of digital repositories are all driven by the ubiquity of digital information including increased capacity of the personal workstation, the broadening of bandwidth and strengthening of network capacities, reduced costs and learning curves associated with digitization and storage, and the proliferation of both commercially and privately produced digital content. Economic pressures associated with the rising costs of journal publications, especially in STM (science, technology, and medicine), along with the development of enabling technologies, have led to open access movements with preprint and postprint digital repositories offering a different paradigm for the distribution of scholarly output. The open source software movement has led to several initiatives that continue to develop repository software and in turn, several library vendors have begun to develop commercial systems for repository management. Learning or course management

This report [dated 27 July 2005] was prepared by Ann Green (Director of Social Science Research Services of Yale ITS Academic Media and Technology) while on assignment to the Library's Integrated Access Council and ILTS Research and Planning. The author wishes to thank Julie Linden, Meg Bellinger, Fred Martz, and David Gewirtz for their contributions. [Futher amenedments to this document were made by M. Bellinger and incorporated into a Final PDF version dated 27 September 2005.]

systems have matured as campus-wide enterprise systems and have enabled faculty to use digital technologies in creating digital learning objects.

Here at Yale, there is evidence of increased faculty creation and use of digital content. In the fall of 2002, then provost Alison Richards sponsored an inventory of digital activity at the university to identify the key issues involved in a wide range of digital production activity. The inventory had several goals, but the primary one was to help gain a fuller understanding of Yale's digital panorama in order to aid the university in its planning for the future. A simple online tool was developed to gather information on a sample of digital projects at Yale. More than 170 projects were described in the inventory. The results confirmed, for example, that digital projects at Yale are continually increasing in number, that they can be found in every corner of the university, that they vary enormously in size and type of content and that they use a wide variety of software tools. Records in the inventory were submitted by libraries, units in ITS (primarily Academic Media & Technology), museums, research centers, professional schools, and individual research projects. All types of projects were represented in the inventory, including catalogs, collections of images, statistical and geospatial databases, digital reference works, websites with backend databases and complex applications, electronic journals, and multimedia projects.

The inventory did not collect information about the life span of projects beyond initial support for producing, distributing, and hosting the digital projects. Half of the projects, however, are listed as "ongoing" and only 20% gave a completion date. This suggests that most projects assume they will have a future and are planning on continuing. Yet it can also be assumed that the majority of the projects are not certain of their future support or long-term support needs, and there is little evidence that projects have begun to plan for ongoing support and maintenance. Clearly, ongoing "support" (in every sense of the word—staff, financial, preservation, etc.) is a critical need for a substantial number of projects. In addition to the inventory findings of 2002, there is evidence of serious interest in the preservation of digital collections embodied in inquiries this past year from several projects and faculty members (e.g. Jonathan Edwards Papers, Yale Daily News, the Cambodian Genocide Project, the Yale Indian Papers Project, etc).

The responsibility for that stewardship, however, is not clear, nor is it apparent how long the producers of the digital data intend the projects to last and remain accessible. Repositories have an obvious role in relation to the long-term maintenance of these distributed digital collections, and could provide a common infrastructure for housing and maintaining long-term access to the valuable digital resources scattered across units, disciplines, and communities of interest.¹

In the spring of 2005, the provost's Teaching and Learning Portal Committee voted to support the migration of Yale's locally developed course management system to the open source based Sakai. The development of Sakai on campus and the support of this initiative by ITS, the Library, and SFAS (Student Financial and Administrative Services)

¹ Summary of the Digital Inventory Working Group's findings have been taken from the Yale Digital Projects Inventory Draft Final Report. (Internal document, October 2004.)

ensure that it will be a more robust environment that will both demand and create more digital content of lasting value.

This report on digital repositories seeks to help clarify how digital repositories are developing to support the academic enterprise. This review articulates the different interpretations of institutional repositories and illustrates them with examples from other academic institutions. The report places digital repositories within a framework of services and interfaces, and presents needs assessments and use cases as ways of evaluating roles of digital repositories. The report also makes recommendations about further assessment of faculty-driven requirements.

This report considers what digital repositories offer to Yale's central mission of creating and disseminating knowledge, how institutional repositories benefit the institution itself and users outside institutional boundaries, and what external funding and policy initiatives will affect Yale's decisions regarding the development of a digital repository infrastructure.

This report makes four recommendations. The Integrated Access Council has initiated recommendations 1 and 2 while recommendations 3 and 4 require broad campus-wide support.

Recommendations:

There is pressing need to find **homes for digital collections** that have proliferated throughout campus and for which the Library has been given or must take stewardship responsibility in order to organize and provide access to digital collections and to ensure **infrastructure for long-term preservation.** To do so, the following actions should be taken:

1 Implement a digital repository testbed using FEDORA and VITAL, and 2. Develop policies, tools, workflows, and good practices for digital collection building in order to promote interoperability and sustainability by multiple creators and compilers across the university, including faculty projects and research center based digital productions.

Although several peer institution have developed repositories as **safe harbors for faculty output** as well as **mechanisms for supporting scholarly publishing, open access, and institutional "branding,"** the needs of Yale faculty and the longterm institutional benefits of such repositories require further analysis. In order to develop deeper understanding of the needs, the following studies should be undertaken:

 Develop a process and program to gain deep understanding of Yale faculty requirements based upon use case studies and other assessment methods.
 Support the development of an Information Architecture Plan for Yale University, with digital repositories as an element of the overall information management strategy.

II. Four Factors Influencing the Development of Repositories

Among the factors that are driving the development of digital repositories are the need for infrastructure to manage the long-term curation of digital collections, the increasing levels of at-risk faculty resources, the need for alternative scholarly publishing models, and the need to repurpose and collaborate on digital assets.

Long-term Curation of Digital Collections

It has become increasingly clear that not only is the volume of digital research output and digital learning content growing at a rapid pace, but that the challenges of managing and providing access to that digital legacy is of considerable concern to funding agencies. The draft of the 2005 report by the National Science Board (NSB) entitled *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* emphasizes that

long-lived digital data collections are powerful catalysts for progress and for democratization of science and education. Proper stewardship of research requires effective policy in order to maximize their potential... Data collections provide more than an increase in the efficiency and accuracy of research; they enable new research opportunities. They do this in two quite different ways. First, digital data collections provide a foundation for using automated analytical tools, giving researchers the ability to develop descriptions of phenomena that could not be created in any other way. While this is true for science that studies natural physical processes, it is particularly enabling for the social scientists.²

Beyond proclaiming the potentially profound cultural value of digital collections, the report also seeks to "frame the issues and to begin a broad discourse. Specifically, the NSB and NSF working together—with each fulfilling its respective responsibilities—need to take stock of the current NSF policies that lead to Foundation funding of a large number of data collections with an indeterminate lifetime and to ask what deliberate strategies will best serve the multiple research and education communities."³

The NSB report also proposes that a policy framework be developed in regard to research collections:

The National Science Board and the National Science Foundation are uniquely positioned to take leadership roles in developing a comprehensive strategy for long-lived digital data collections and translating this strategy into a consistent policy framework to govern such collections. Policies and strategies that are developed to facilitate the management, preservation, and sharing of digital data

² United States. National Science Board. Long-Lived Digital Data

Collections,*H*<u>http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf</u>H, 5, 8-9. ³ Ibid., 10.

will have to fully embrace the essential heterogeneity in technical, scientific, and other features found across the spectrum of digital data collections.⁴

Recommendations from the NSB clearly indicate that the National Science Foundation and other funding agencies "should require that research proposals for activities that will generate digital data, especially long-lived data, should state such intentions in the proposal so that peer reviewers can evaluate a proposed data management plan."⁵

At the University of California, Santa Barbara, the report from a Mellon-funded investigation (2004) into the characteristics of current data-intensive research projects articulates the need for "an infrastructure that enables faculty to preserve their research which would result in more of their time being devoted to research and teaching."⁶

It became clear from the frank discussions and roundtables held with faculty, staff researchers, and technologists that UCSB researchers have reached a critical period. With increasing levels of computing power and complexity and advanced skill sets of experts, scientific data has increased in depth, breadth, and volume as never before. Faculty and staff manage these data with varying degrees of success given the limited resources and competing obligations. Customized systems and tools are proliferating but it is not obvious whether these could be leveraged more efficiently. It is more worrisome that long-term preservation especially for data has become sidelined. Recent data loss, large-scale interdisciplinary collaborations, and projects requiring historical analysis have exposed how extremely valuable, and at risk, UCSB research data is.⁷

Academic institutions must position themselves to respond to emerging policies regarding the preservation of research data as well as to the challenges of establishing effective long-range curation solutions both locally and collaboratively. Yale's Provost Andrew Hamilton has talked about the primary mission of Yale being the creation and dissemination of knowledge, and digital repositories have the potential of playing a critical role in that mission.

Faculty Resources at Risk

Faculty-produced digital materials currently reside on personal desktops, e-mail systems, multiple server environments and websites (personal, departmental, and research center based), each with varying levels of backup, disaster recovery, security, and long-term support. Very rarely are these materials managed in ways that will protect them from media decay, technological migration, and software and operating system changes. And

⁴ Ibid., 5.

⁵ Ibid., 6.

⁶ Pritchard, Carver, and Anand, "Research Information Management,"

Hhttp://www.library.ucsb.edu/informatics/informatics/documents/UCSB Informatics White paper July04 .docH, 5.

⁷ Ibid., 2.

as importantly, most of these "assets" are not discoverable or searchable and are not easily integrated into the workflow of teaching and research.

Clifford Lynch, director of the Coalition for Networked Information (CNI)⁸ has discussed the potential of institutional repositories in this context:

Most individual faculty lack the time, resources, or expertise to ensure preservation of their own scholarly work even in the short term, and clearly can't do it in the long term that extends beyond their careers; the long term can only be addressed by an organizationally based strategy. Institutional repositories can address both the near-term questions about continuity of access ... and also the longer-term questions about preservation by creating an institutional commitment to such preservation.⁹

But creating digital repositories, even with low barriers for participation, does not ensure that faculty will choose to use them as a means of protecting digital materials. One recommendation in a study sponsored by the UK Joint Information Systems Committee (JISC) stated that "Repositories need to be positioned within the workflow of personal content creation and the 'personal information environment' of users." ¹⁰ But even if materials were easily accessioned into safe harbors, we do not necessarily understand what to protect and preserve and what to allow to disappear into the digital graveyards of technological obsolescence. How much of the material at risk is of long-term value to the faculty member, the research center, the university, the academic discipline, and scholars of the future? Further understanding of the digital life span requirements of digital materials from a faculty member's perspective is needed to assess the demand for a range of stewardship commitments and services.

Neil Beagrie writes about the increasing awareness of this risk to digital assets:

As digital content in personal collections continues to grow, particularly content that has been paid for such as digital music or video, it seems likely that

⁸ "CNI is an organization dedicated to supporting the transformative promise of networked information technology for the advancement of scholarly communication and the enrichment of intellectual productivity. Some 200 institutions representing higher education, publishing, network and telecommunications, information technology, and libraries and library organizations make up CNI's Members." HHhttp://www.cni.org/

⁹ Lynch, "Institutional Repositories" http://www.arl.org/newsltr/226/ir.html.

¹⁰ "The Joint Information Systems Committee (JISC) supports further and higher education by providing strategic guidance, advice and opportunities to use Information and Communications Technology (ICT) to support teaching, learning, research and administration. JISC is funded by all the UK post-16 and higher education funding councils." Hhttp://www.jisc.ac.uk/H. "The JISC is bringing together a programme of work relating to digital repositories. Its aim is to bring together people and practices from across various domains (research, learning, information services, institutional policy, management and administration, records management, and so on) to ensure the maximum degree of coordination in the development of digital repositories, in terms of their technical and social (including business) aspects." JISC, "Digital Repositories Programme," http://www.jisc.ac.uk/uploaded_documents/rep-review-final-20050220.pdfH, 4.

individual and public consciousness of and concerns over digital continuity will also increase. ... For any collection intended for access and use over a decade or more, the incremental accumulation of risk will become unacceptable. Its mitigation may become more inherent and automated in systems. Similarly, public awareness of and resistance to all but essential format migrations and associated costs may increase. ... To date there has been little ... activity in major research repositories on personal digital collections—although one might expect them to be a major focus of future research. ... Personal digital collections should become a major area of interest for research collections.¹¹

Scholarly Publishing, "Open Access," and Institutional Repositories

With the increasing presence of web-based publishing and with the availability of publicly accessible collections of scholarly resources, the open access initiative has gained support from numerous library and scholarly professional organizations to establish alternatives to high-cost scholarly publishing models and improve access to research and teaching outputs.

Yale faculty are taking part in open access publishing as discussed in the Forum on Scholarly Publishing in February 2005 (sponsored by the Cushing/Whitney Medical Library, the Library's Integrated Access Council, and the Science Libraries). In her opening remarks, University Librarian Alice Prochaska noted, "Yale is an institutional member of the Public Library of Science (PLoS)¹² and BioMed Central,¹³ both open access publishing initiatives. Our membership in these organizations covers all or part of the article publication fees for Yale authors." To date, more than 80 articles by Yale faculty have been published by PLoS and BioMed Central. The arXiv pre-print server is "an e-print service in the fields of physics, mathematics, non-linear science, computer science, and quantitative biology." Yale authors deposit pre-prints in arXiv; to date, 765 submissions are from yale.edu in Physics.¹⁴

Beyond providing "open" access to publications, most recent developments involve the right for authors to "self-archive" and for communities to build alternative publishing and dissemination mechanisms via the web.¹⁵ Institutional repository systems play a significant role in these developments by providing services for new models of scholarly communications and online publishing solutions. Repositories also offer content

¹¹ Beagrie, "Plenty of Room at the Bottom? " http://www.dlib.org/dlib/june05/beagrie/06beagrie.html ¹² Public Library of Science (PloS) is "a nonprofit organization committed to making the world's scientific and medical literature freely available online, without restrictions on use or further distribution, free from private or government control." The PLoS journals are run by professional editors, trained scientists, and physicians, in close collaboration with editorial boards. Hhttp://www.plos.orgH

¹³ "BioMed Central is an independent publishing house committed to providing immediate free access to peer-reviewed biomedical research." (http://www.biomedcentral.com/info/) As of today it has more than 100 journals.

 ¹⁴ H<u>www.arxiv.org</u>H. Counts are from e-mail from arXiv administrator Michael Fromerth, July 11, 2005.
 ¹⁵ Swan and Brown. *Open Access Self-Archiving*,

Hhttp://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archivingan%20author%20study.pdfH.

management software and technological support for indexing, searching, and accessing networked repositories, thus lowering access barriers and offering wider dissemination of scholarly output. SPARC – Scholarly Publishing and Academic Resources Coalition – is an alliance of academic and research libraries that seeks to facilitate the emergence of systems that capitalize on the networked environment to disseminate research. SPARC's Enterprise Director, Richard Johnson, writes:

Institutional repositories build on a growing grassroots faculty practice of posting research online, most often on personal web sites, but also on departmental sites or in disciplinary repositories. ... [I]nstitutional repositories merit serious and immediate consideration from academic institutions and their constituent faculty, librarians, and administrators. ... While institutional repositories centralize, preserve, and make accessible an institution's intellectual capital, at the same time they will form part of a global system of distributed, interoperable repositories that provides the foundation for a new disaggregated model of scholarly publishing. ¹⁶

Clifford Lynch has also emphasized the important role that digital repositories can play in scholarly communication:

Institutional repositories can encourage the exploration and adoption of new forms of scholarly communication that exploit the digital medium in fundamental ways. This, to me, is perhaps the most important and exciting payoff: facilitating change not so much in the existing system of scholarly publishing but by opening up entire new forms of scholarly communication that will need to be legitimized and nurtured with guarantees of both short- and long-term accessibility. Institutional repositories can support new practices of scholarship that emphasize data as an integral part of the record and discourse of scholarship. They can structure and make effective otherwise diffuse efforts to capture and disseminate learning and teaching materials, symposia and performances, and related documentation of the intellectual life of universities.¹⁷

A recent survey of 1,296 scholars in the UK explored the recent trends of depositing journal articles (usually peer-reviewed post-prints or as pre-print drafts) in repositories and websites, a practice known as "self archiving." Institutional repositories provide one type of solution for this form of "archiving." The authors of the study's report write:

Almost half (49%) of the respondent population has self-archived at least one article during the last three years in at least one of the three possible ways — by placing a copy of an article in an institutional (or departmental) repository, in a subject-based repository, or on a personal or institutional website. More people (27%) have so far opted for the last method — putting a copy on a website — than have used institutional (20%) or subject-based (12%)

¹⁶ Johnson, "Institutional Repositories: Partnering with Faculty," H<u>http://www.dlib.org/dlib/november02/johnson/11johnson.html</u>H.

¹⁷ Lynch, "Institutional Repositories" http://www.arl.org/newsltr/226/ir.html.

repositories, though the main growth in self-archiving activity over the last year has been in these latter two more structured, systematic methods for providing open access. Use of institutional repositories for this purpose has doubled and usage has increased by almost 60% for subject-based repositories. Postprints (peer-reviewed articles) are deposited more frequently than preprints (articles prior to peer review) except in the longstanding self-archiving communities of physics and computer science. There are some differences between subject disciplines with respect to the level of self-archiving activity and the location of deposit (website, institutional or subject-based repositories). Selfarchiving activity is greatest amongst the most prolific authors, that is, those who publish the largest number of papers.¹⁸

Repurposing Content, Collaborative Environments

Highly customized stand-alone digital collections and web spaces do not provide the best environment for identifying, locating, or repurposing digital content. Common infrastructure and the use of open standards and best practices optimize unified access to diverse collections (discovery, cross-collection searching), and enable open collaboration across units, academic disciplines, and institutions. Communities can reformat, repurpose, republish, and build upon investments in research and teaching.

An institutional repository ... becomes the mediator for a one-input, manyoutputs scenario, where a researcher can retrieve whichever elements of his or her own research record are needed for a task-in-hand (perhaps writing a paper, a lecture, preparing teaching materials, preparing a CV). It can also provide the home for research data that cannot be published in traditional journal format but which supports research findings and which the author would like to make available to peers and colleagues, data such as very large datasets, video files, graphical files of various formats, audio files and mixed media output.¹⁹

We see evidence of this type of research collaboration in numerous applications on Yale's research.yale.edu service and the repurposing of course materials from semester to semester on classes.yale.edu.

In addition to faculty-generated digital collections, Yale's libraries, galleries, and museums have made significant investments in the acquisition, licensing, and production of digital collections. These include, for example, the Avalon Project, the Beinecke Digital Collections, the Divinity Digital Image and Text Library, the Economic Growth Center Digital Library, Insight Digital Collections of images, and the Manuscripts and Archives Digital Image Database. These and others are listed on the Library's website at

¹⁸ Swan and Brown. "Executive Summary," *Open Access Self-Archiving*,
 H<u>http://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archiving-</u>
 an%20author%20study.pdfH The study was commissioned and funded by the Joint Information System

¹⁹ Ibid., 5.

an%20author%20study.pdfH, The study was commissioned and funded by the Joint Information Systems Committee (JISC) in the UK and the Open Society Institute.

http://www.library.yale.edu/libraries/digcoll.html. Repurposing content, maintaining that content in persistent and reliable formats, and broadcasting the availability of the digital objects within these collections are central to the goals of digital repository developments.

Summary of Incentives for Establishing Digital Repositories:

- The university will be well positioned to address funding agencies' longrange curation requirements as those policies are developed. The university can take a proactive role in developing infrastructure and influencing policies and standards for the long-term management and preservation of digital assets.
- New modes of dissemination, new modes of publication, wide dissemination of knowledge influence the development of digital repositories.
- Sharing and re-use of digital resources for collaborative environments enable pre-publication exchanges among researchers; teaching environments can support the reuse and development of learning objects such as images, datasets, multimedia productions, reading lists, digitized texts, etc.
- Different types of digital assets can be brought into a managed environment that attends to the life-cycle process; repositories facilitate efficient storage and management of resources.
- Enhanced discovery of and access to the institution's own digital assets and those acquired from other digital producers can be supported and integrated.
- Preservation of digital resources: without a repository structure, migrating digital material through technological change is considerably more difficult and prone to loss of digital assets
- Quality assurance: digital repositories have the potential to remove questions of provenance and substance.

The various incentives for creating digital repositories as described above illustrate the complexity of environments, policies, and content that must be considered as a digital strategy is developed for Yale University. Strategies are needed for dealing with complex technological challenges, multiple user expectations and user requirements, and an extremely complicated array of digital resources. As Campbell, Blinco, and Mason state, "institutional repositories and their management strategies will need to deal with a mixture of both 'raw' and 'processed' materials, including datasets, 'un'published yet persistent material, dynamic resources, preprints, scholarly publications, resource lists, and e-portfolios."²⁰

²⁰ Campbell, Blinco, and Mason. *Repository Management and Implementation*, H<u>http://www.jisc.ac.uk/uploaded_documents/Altilab04-repositories.pdf</u>H, 7.

III. What is a Digital Repository?

There is a confusing array of different interpretations of what an institutional digital repository is; there are many interpretations about what it contains, what services it offers to whom, who can participate, and what it promises over the long term.

In the JISC-funded report entitled Digital Repositories Review:

Repositories are "collections of digital objects' but what makes repositories distinctive from other collections of digital objects such as directories, catalogues, databases? What are the defining characteristics of a 'repository'? As with other terms that have been popularised in the digital world (portal, architecture...) some qualification is required: is the repository managed as an *institutional repository*? or a *subject repository*? What is the content of the repository? an *e-prints repository*? a *data repository*? Is the underlying purpose of the repository for *preservation, access, or data management*?²¹

In general, a repository has been defined by JISC in the following way:

- content is *deposited* in a repository, whether by the content creator, owner or third party on their behalf
- the repository architecture manages content as well as metadata
- the repository offers a minimum set of basic services e.g. put, get, search, access control
- the repository must be sustainable and trusted, well-supported and wellmanaged²²

However, not all of these conditions are consistently part of the goals of every digital repository, as illustrated in the examples in Appendix 1. Not all repositories house metadata, not all repositories provide search capabilities, and not all repositories have the funding to be sustainable or pass an official certification as being "trusted." Adding to the complexity and confusion regarding what digital repositories are and what their purposes may be is the fact that they can dramatically evolve over time. As noted in some examples below, new functions are added, policies evolve, participation rates ebb and flow, standards evolve, and communities redefine the functions and goals of the repositories.

This investigation moves beyond these characteristics and builds a typology of repositories based upon the functions and goals of multiple repositories and takes the broadest possible interpretation of digital repository models. This allows us to explore multiple content types (publications, learning objects, datasets, images, theses and dissertations, videos, websites, etc.) and diverse potential participation (includes faculty producers and collection builders) and include a wide range of issues, from long-term

²¹ http://www.jisc.ac.uk/uploaded_documents/rep-review-final-20050220.pdf, 1.

²² Ibid., 2.

stewardship of faculty-produced knowledge assets to the intersection of course management systems and personal collection management, as well as exploring the complexities of digital life-cycle management within faculty research and teaching contexts.

Digital Repository Typology

The following areas of focus help clarify the range of digital repository implementations:

- Safe harbors for faculty output: digital repositories built to house the scholarly output of teaching and research activity, repository content and use are dependent upon faculty participation and their willingness and ability to contribute digital materials. Content may be of multiple types and in varying states of completion and availability for sharing or publishing.
- Mechanisms for supporting scholarly publishing, open access, and institutional "branding": digital repositories whose objectives are related to supporting new models of scholarly publishing that usually offer public access; can have objectives related to the institutional "branding" of digital assets. Content is usually pre- and post-publication materials.
- **Homes for digital collections**: digital repositories that focus upon bringing together digital content (either within a common repository or within a repository framework) for enhanced search, retrieval, and repurposing for teaching and research; objectives may include moving stand-alone digital collections into integrated access environments.
- **Infrastructure for long-term preservation**: digital repositories that are trusted to preserve the formats and intellectual content in perpetuity with long-term stewardship commitments. These systems can be independent from search, display, and other applications and services.

A. Digital Repositories as Safe Harbors for Faculty Output

Many universities are investing in digital repositories that serve as safe harbors for faculty-driven research and teaching output (pre- and post-prints, working papers, research reports, datasets, course materials, personal image collections, etc.) These repositories offer a means for locating and disseminating the intellectual output of universities. They not only allow local branding of digital assets but also have the potential of offering collaborative environments for cross-discipline and cross-institutional scholarly activity. Functional requirements include storage, access, and digital content management. Detailed policies, practices, and technological solutions are required regarding intellectual property rights, access restrictions, and version control.

B. Mechanisms for Supporting Scholarly Publishing, Open Access, and Institutional 'Branding'

Some digital repositories have been developed to replace or supplement the peerreviewed scholarly journal publication process. "Enhancing access to scholarly communications has been a main driver for establishing repositories, both institutional repositories (in particular e-print archives) and subject based archives. Many, though by no means all, repositories" support open access at least in part to the digital assets and most all provide open access to metadata for harvesting.²³

C. Homes for Digital Collections

Digital repositories also house collections of digital materials grouped by type, subject, purpose, or ownership. Digital content from these collections is made available for multiple analysis and research applications; these collections hold the raw materials that faculty and students can reuse, repurpose, analyze, and recompile in numerous teaching, learning and research environments. The digital resources held in these collections are created through digitization processes or are collections of "born digital" materials. The collections are often discipline-specific, built for varying purposes and restricted by locally determined access policies. Content is not necessarily limited to "locally" produced digital assets but may include digital materials obtained through licensing and purchase. Some of these digital repositories are developed and supported within the context of a single institution, government agency, academic discipline, or consortium (for example ICPSR, the Inter-university Consortium for Political and Social Research). This report focuses primarily upon those digital repositories developed within institutions, but acknowledges they are part of a much broader knowledge landscape.

The distributed nature of these digital collections and the idiosyncratic characteristics of their contents, metadata, and policies present a challenging and complex landscape for digital repository planning and integration. Within this context, repositories and repository frameworks are developed with emphases upon digital content management, integrated access, and the use and promotion of common standards and protocols. Integration solutions include cross-collection federated searching, harvesting metadata into central search and locating services, and software-based federated repository systems (e.g. Fedora²⁴).

D. Infrastructure for Long-term Preservation

This type of digital repository focuses primarily upon supporting the digital life-cycle perspective of providing support for digital assets from birth through use to long-term access and careful preservation.

²³ JISC. *Digital Repositories Review*, H<u>http://www.jisc.ac.uk/uploaded_documents/rep-review-final-20050220.pdf</u>H.

²⁴ Fedora stands for "Flexible and Extensible Digital Object and Repository Architecture." http://www.fedora.info/

The long-term survival, value and usability of the information stored within digital repositories depends on numerous criteria such as the formats selected for storage, the capture of associated metadata, proactive preservation measures, and the perceived trust in the repository itself. These represent just a few of the factors that may affect the long-term viability of digital information held within repositories. Factors such as economic sustainability, populating digital repositories and rights management are also integral to the long-term usability of digital information²⁵

Funding for research programs that produce digital output do not normally carry any provisions for preparing or preserving the data, descriptive research materials, or laboratory results, i.e. the "raw materials" of research projects. As noted above, the National Science Board and National Science Foundation are beginning to address the issue of "long-lived digital data" at a policy level; institutional consideration of and infrastructure for long-term preservation may be driven in part by these external funding and policy initiatives. The economic challenges of documenting and preserving research data recently have become more apparent. Universities have a critical role to play in the articulation of these fiscal challenges and in the design and provision of technological infrastructure to support the activities necessary to preserve these research materials.

In 2003-2004, PREMIS (Preservation Metadata: Implementation Strategies, a working group jointly sponsored by OCLC and RLG) conducted a survey of academic institutions "aimed at gathering information on key aspects of planned and existing preservation repositories for digital materials." ²⁶ The survey is especially notable in its goal to discover how preservation repository systems are actually being implemented. Responses were received from twenty-eight libraries, seven archives, three museums, and eleven other types of institutions. Included in the findings:

- Most repositories serve the two goals of preservation and access. Less than a fifth could be called "dark archives."
- All respondents offered "secure storage" as a service; 92% offered or planned to offer preservation treatments, defined as normalization, migration, emulation, or other strategies designed to ensure long-term usability.
- The majority of institutions chose more than one strategy for preservation. Most (85%) are offering bit-level preservation. Beyond that, restrictions on submissions, normalization, migration and migration-on-demand are the four most popular strategies, in that order.
- According to the respondent's future plans, the four most popular strategies, in order, will be migration, normalization, restrictions on

²⁵ Joy Davidson. Digital Curation Center Training Coordinator and ERPANET British Editor. Announcement of the DCC Workshop on the Long-term Curation within Digital Repositories. E-mail June 23, 2005.

²⁶ OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS), Implementing Preservation Repositories for Digital Materials H<u>http://www.oclc.org/research/projects/pmwg/surveyreport.pdfH, i.</u>

submission, and migration-on-demand. Emulation is being used now by only 10% of respondents, but that doubles when future plans are considered.

Although digital preservation repositories are in the early stages of development, there is a very strong international commitment to building standards and best practices for the development of digital preservation systems and digital preservation metadata.

Recognizing that digital materials will have varying life spans, the Duke University Digital Archive has developed three basic tiers of custodianship for which they will develop services.

Archived – Materials which have been deemed of significant and widespread value, and which will be given a high level of care and attention, including provision of complex, normalized metadata, control over chain-of-custody and authenticity, sophisticated access tools, and periodic migration to current formats. Materials in this category will require significant time and resources to maintain at this level, therefore stringent selection criteria will be applied to determine which materials merit this level of effort and commitment, and cost-recovery models may be used to finance this level of care for certain materials or sources. Preserved – Materials which have enduring value, but for which resources are not currently available to provide complex metadata, customized access tools, or make ongoing migration commitments. The majority of the materials in the system will likely be in this category. The Library makes a commitment to preserve and provide long-term access to the materials in the formats in which they were submitted (preserving the bits), but makes no commitment to migrate to new formats or to add metadata beyond the basic level provided by the individual or organization submitting the materials to the Archive. Materials in this category may be upgraded to the Archived category where appropriate as funding to do so becomes available.

Stored – Materials not owned or managed by Duke or over which Duke does not control the copyrights, but which have significant and ongoing value to Duke scholarship. Copies will be kept of these materials in order to provide local access and as a backup copy in the event of damage to or loss of the originals, which are stored and managed elsewhere. Materials in this category might include CD-ROMs that arrive in the back of books, or data from proprietary vendor databases or copies of materials from the archival systems of peer institutions, publishers, or scholarly societies who maintain copyrights over the materials. The Library makes no commitments regarding materials in this category, other than to provide access to and periodically refresh the data and metadata from the source systems as agreements with the original sources permit.²⁷

²⁷ Duke University Libraries. "Duke University Digital Archive: Tiers of Custodianship," H<u>http://www.lib.duke.edu/its/diglib/digarchive/custodianship.html</u>H, accessed July 2005. "Approximately a dozen systems were evaluated, with DSpace eventually being selected as the system that would work best for Duke. ITS is currently pilot-testing DSpace on a development server, and plans to work closely with colleagues in Arts & Sciences Computing who have already implemented DSpace in production for the

The Yale University Library has begun to build the policies and infrastructure for longterm stewardship of its digital collections.

The practice of storing human knowledge in electronic formats has forever changed the way librarians and archivists think about the collection and preservation of the cultural record. The University Library aggressively creates and acquires electronic collections, but these holdings are as vulnerable as sand castles on a beach unless protected by a digital preservation program. Access to digital objects is as short lived as the next technological wave. Technology waves therefore are a direct challenge to the Library's mission to be a custodian of human knowledge. To remain a good steward of the cultural record the Library must establish a digital preservation infrastructure.²⁸

Based upon this commitment to developing a digital preservation infrastructure, the Library-sponsored Integrated Access Council charged a Digital Preservation Committee to "develop a preservation program that ensures digital information of continuing value will remain persistent, accessible and usable. The committee will evaluate, compile, document and articulate policies, procedures, best practices and systems in order to establish a digital preservation infrastructure at Yale University Library. The Committee will work from a base of clearly articulated policies, then will focus on preservation program planning and, finally, will make recommendations for implementation through digital preservation projects, initiatives, and system development." Deliverables for the first year include: a digital preservation mission statement; the publication of a digital preservation program; and time lines for digital preservation projects, initiatives, and implementation strategies.²⁹ As its first work, the committee has a preservation policy draft in production with an expected release date of late summer 2005.

In addition, the Library has established a Rescue Repository to provide a centrally supported system for the short-term safekeeping of materials in immediate danger of permanent loss through media decay, physical damage, technological obsolescence, or difficulties in archival management. This repository constitutes the first step toward the evolution of a long-term, OAIS-compliant digital preservation archive.³⁰

Duke Student Portfolio as well as with colleagues at NCSU and UNC who are also developing DSpace systems for their institutions. More information will be available here as pilot testing and further planning progresses." "Duke University Digital Archive: Technical Infrastructure,"

Hhttp://www.lib.duke.edu/its/diglib/digarchive/tech-infrastructure.htmlH, accessed July 2005. ²⁸ Paraphrased from Integrated Access Objectives Planning Report to Library Management Council, September 21, 2002.

Hhttp://www.library.yale.edu/lmc/public/Integrated%20Access%20OPG%20Report%20(Intro)%2009.25.0 2.pdfH

^{2. put 1}
²⁹ Yale University Library, Integrated Access Council. "Digital Preservation Committee Charge," H<u>http://www.library.yale.edu/iac/DPC/digpreschg.html</u>H, October 2004.

³⁰ Yale University Library Rescue Repository Requirements Task Force, "Requirements Document for the Rescue Repository," May 2004.

http://www.library.yale.edu/iac/documents/RescueRepositoryRequirements.pdf

Finally, the IAC's Metadata Committee and Digital Preservation Committee have appointed a Joint Task Force on Preservation Metadata whose charge is "to propose or customize an existing preservation metadata model and accompanying data dictionary for implementation at Yale University that identifies and defines preservation activities, processes, entities, elements and their relationships. In addition, the PMTF will make preliminary recommendations ... on a second phase of digital preservation planning in order to serve designated communities at Yale that need to use preservation metadata."

Examples at Other Universities

As specific examples of digital repositories were investigated, it was clear that each implementation focused to a greater or lesser extent upon each of the repository types listed above. To aid the review of repository efforts at other universities, an extensive list of questions informed the discussions and findings about content flows, policies, participation, and priorities. (See Appendix 3 for an outline of the themes covered.) These questions were based upon a questionnaire from the OCLC/RLG PREMIS Working Group on Preservation Metadata Implementation Strategies entitled "Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community."³¹

Who chooses content? What is required to participate? What is the goal of the repository? The questions also dealt with a range of issues including the responsibility for long-term stewardship of faculty-produced knowledge assets, and the intersection of course management systems and personal collections management. Information was collected about policies, funding, user requirements, needs assessments, and expectations for participation.

The review was not intended to be a survey of all institutions or types of software, but rather it seeks to build an understanding of what other institutions are doing and to work toward a common vocabulary based upon repository examples. Each repository implementation developed under the influences of varying communities and functional objectives. It should be noted that this review was not technology focused; rather than build this inquiry as a review and comparison of different software systems (DSpace, Fedora, Digital Commons, ContentDM, Eprints, DigiTool, etc) this is a glimpse at a small set of universities that share common challenges. For an overview of repository software solutions, see the Open Society Institute's publication: *A Guide to Institutional Repository Software*, particularly the table on features and functionality.³²

Each of the four types of digital repository focus mentioned above (faculty focused, scholarly publication focused, digital collections focused, and preservation focused) has specific functional requirements. Some of the requirements are common across all the

³¹ <u>http://www.oclc.org/research/projects/pmwg/surveyreport.pdf</u>

³² Open Society Institute. A Guide to Institutional Repository Software v 3.0. Feature & Functionality Table. H<u>http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_Table_v3.pdf</u>H

types and some are specific to a particular type. The following table illustrates how a small number of selected digital repositories compare in the generalized focus areas. Clearly this table oversimplifies the roles and long-term capabilities of the sample repositories, but moves us closer to understanding the multiplicity of purpose, participation, and content flows among the various examples.

A Few Examples of Digital Repositories									
and their Primary Initial Functions									
	Safe Harbors	Scholarly	Digital	Preservation					
	for faculty	Publishing,	Collections	and Life Cycle					
	output	Open Access	Repository	Management					
MIT: DSpace	•	X	X						
CDL:		Х							
eScholarship									
Repository									
Harvard		Х	Х						
Science Digital									
Library									
Georgia Tech	Х	Х	Х						
Univ of			Х						
Michigan:									
DSpace									
Univ of			X						
Virginia:									
FEDORA									
Rutgers Univ:			X	D					
FEDORA									
Harvard Univ:				X					
Digital									
Repository									
Service									
X = Current Function									
D = Function in Development									
Yellow = Scholarly workspace									

There is a great deal of digital activity by scholars in faculty workspaces but little digital repository support for that activity (see the yellow column above). Scholars are involved in creating digital productions, they may contribute to open access journals, online journals within their academic disciplines, and can "self-archive" their pre- and post-print publications in institutional repositories. They might also be creating digital collections of numerous content types, including personal collections, blogs, e-mail, images and multimedia, numeric datasets, econometric models, complex relational databases, output from experiments, texts, etc.

The following table explores the potential of the three digital repository types to act as targets for the management and dissemination of these various sorts of scholarly output. Further investigation needs to be made in regard to the potential for participation and the role of repositories in the workflow and long-term access to the digital materials in the scholarly workspace. How much can and should go to the "Homes for Digital

Collections" type of repositories? What is the role of a digital collection repository for learning objects and collections used primarily for instruction?

The last row in the next table begins to explore the role that digital life-cycle management, and the policies, tools, best practices and services that support the life cycle of digital materials, might move more of those materials of value from the scholarly workspace into the different repository types to the right.

Content Flows

From scholarly workspaces to digital repository types By type of material

And with digital life cycle view and support

	Scholarly workspace	Scholarly Publishing, Open	Digital Collections Repository	Preservation and Life Cycle Management
Instructional materials	x	Access	9	
Research datasets and databases	X	?	?	
Images, multimedia productions	Х	?	?	
Texts	Х	?	?	
E-mail, blogs, collaborative interactions	X			
Dynamic websites with applications and backend databases	Х		?	
Selected materials produced and managed w/ digital life cycle view and support	X	x	X	X

The last table, below, adds columns with some examples of digital asset content storage and dissemination systems that currently exist or are proposed as pilots at Yale. The examples are Sakai (the developing course management system called classesv2), the proposed pilot of Fedora and VTLS, Luna Insight image collection, and the Social Science Data Archive (SSDA). Again, the life-cycle oriented support, policies, and best practices are represented in the final row.

CONTENT FLOWS								
With sample digital collection repository targets								
	Scholarly		Scholarly	Fedora	Sakai	Luna	SSDA	Preservation
	workspace		Publishing,					and Life
			Open					Cycle
			Access					Management
Instructional	Х			?	X	Х	Х	?
materials								
Research	Х		?	?			Х	
datasets and								
databases								
Images,	Х			?		Х		?
multimedia								
productions								
Texts	Х		?	?	Х			?
E-mail,	Х				Х			
blogs,								
collaborative								
interactions								
Dynamic	х			?				
websites w/								
applications								
and backend								
databases								
Selected	Х		Х	Х	?	?	?	X
Materials								
from above,								
produced								
and								
managed w/								
life cycle								
view and								
support								

IV. Multiple Digital Repositories within a Tiered Architecture

A digital repository framework will need to support a wide variety of materials, and as Clifford Lynch states, the repository structure needs to allow for multiple access systems layered on top of an underlying repository layer:

My personal view is that we're going to start seeing a two-layered model. We're going to see digital collections that are presented and

managed in a passive way. They will function similar to a repository where stewardship is the major theme. Then you're going to find access systems layered on top of these, which may be more volatile. They may have shorter lives than the underlying collections. You may see the same collections presented through multiple access systems. These access systems will be not just retrieval tools, but analysis environments in some cases. We'll see a great diversity in these access systems...³³

There are evolving complexities and challenges that will have significant impact upon the information architecture and the design and development of institutional repositories. For example, it is very important to work on the challenges of integrating digital repositories with learning/course management systems (e.g. Sakai).

Flecker and McLean (2004) discuss in detail the need to "increase the integration of existing digital resources into the working environments of instructors of higher education."³⁴ They recommend that repositories be developed and operate in ways that enable other systems to interoperate with them. They developed a checklist against which repository developers could measure themselves in regard to the interoperability criteria. The results were that "no repository comes close to satisfying all the criteria."³⁵ Since that report, there has been significant increase in the awareness of these challenges and much more interaction among course management developers and digital library/institutional repository implementers.

Yet the challenges persist, as pointed out in the 2005 JISC study:

At present there is very little interoperability between repositories. For example, e-print institutional repositories are unlikely to be linked to or interact with repositories for teaching and learning. Software does not facilitate sharing services between repositories, or provide the full range of functionality that users might require - users in the broadest sense to mean those submitting content, those managing content, and those using content.³⁶

The framework of interconnected repositories, services, and applications is defined by the functional requirements of particular use cases and digital asset content flows among multiple systems. The demands for repurposing digital assets in multiple systems and for multiple scholarly applications should drive the planning of digital repository systems. Not one digital repository fits all requirements of every community or goal. Not one of the particular types of digital repositories (faculty focus, publication focus, collections focus, or preservation focus) has emerged in isolation; these different models indeed coexist within a common architecture and a shared digital landscape

³³ Lynch, "Check Out the New Library,"

http://www.acm.org/ubiquity/interviews/pf/c lynch 1.html

³⁴ Digital Library Content and Course Management Systems, Hhttp://www.diglib.org/pubs/cmsdl0407/H,

^{2.} ³⁵ Ibid., 9.

³⁶ Digital Repositories Review, Hhttp://www.jisc.ac.uk/uploaded_documents/rep-review-final-20050220.pdfH, 14.

The generalized diagram below, from the Johns Hopkins study of repositories and services,³⁷ builds on this idea of multiple digital repositories operating in a layered architecture with a "repository interface layer." In this case, multiple services operate through an agnostic repository layer to pull resources from multiple repositories. The three tiers are repositories on the bottom layer, a "repository interface layer" in the middle, and a range of "services" at the top layer.



The services on the top layer (the boxes) illustrated in this example are Sakai (a course management service), Chandler (a personal information management application), OSP (an e-portfolio application), and LionShare (a peer-to-peer architecture for sharing educational materials in a secure manner). The Repository Interface Layer mediates connectivity between applications and repositories. In theory, applications could write to this layer without knowing what repositories might access it, and repositories used as examples here are DSpace, Fedora, ContentDM, and a collection of science data.

Obviously, the diagram includes no library content, Orbis, web applications, or licensed content. Repositories also need to interface with other services within institutions such as portals and library catalogs. It isn't a view of the complete landscape by any means but it helps to illustrate the idea that applications should access repositories through an abstract, repository-agnostic layer. Content can thus be managed in multiple repositories external to applications, so that the same content can be used by several systems and

³⁷ Choudhury and Martino. *A Technology Analysis of Repositories and Services*, http://www.cni.org/tfms/2005a.spring/abstracts/presentations/CNI_martino_technology.pdf.

support multiple services. This concept is being tested with content that is moved through repositories into applications as defined against a set of use cases that reflect various services. The focus is on moving content between repositories and applications; the tiered architecture allows content integration and diversity of repositories, applications, and digital objects.

The following diagram, by Fred Martz (Director, Integrated Library Technology Services, Yale Library) illustrates the capabilities of cross-platform metadata harvesting where metadata harvesters aggregate data from individual Yale collections shown in the yellow and green boxes (e.g. Beinecke Digital Library of images, library catalogs and finding aids, digital collections of statistical tables, museum and gallery collections).

Included in the diagram is a digital repository pilot of Fedora and VTLS's VITAL software, which could hold faculty output, library and research based digital collections, research databases, course materials, image collections, etc. The blue boxes outside the repository hold personal materials and course materials that are linked to Sakai and the university portal, which in turn can be integrated into the metadata harvesting scheme; the Fedora digital repository has the ability to ingest materials (both digital objects and metadata) from Sakai. Future enhancements could enable Fedora/VITAL to harvest metadata from other digital collections and catalogs (shown here by the small boxes inside Fedora labeled O for Orbis, S for SSDA, V for the Visual Resources Collection, B for the British Art Center, and A the Yale University Art Gallery.)

The orange OAI-PMH³⁸ service providers and the red external portals serve as examples of the "multiple access systems and analysis environments" envisioned by Clifford Lynch.

The Open Archives Initiative Metadata Harvesting Protocol gives us the tools for an institutional repository to act as an entry point for redistributing works to systems of disciplinary repositories. It is desirable to make this as simple as possible, and to insulate faculty from having to deal with the details of a constantly evolving multiplicity of disciplinary services. Better to present the faculty with a simple and stable submission interface to the institutional repository. In this sense institutional repositories can be an infrastructure upon which disciplinary services and repositories can build.³⁹

³⁸ OAI-PMH is the Open Archives Initiative Protocol for Metadata Harvesting. H<u>http://www.openarchives.org/</u>H

³⁹ Lynch, "Institutional Repositories." http://www.arl.org/newsltr/226/ir.html.



Diagram courtesy of Fred Martz, July 2005.

V. Assessing Needs and Participation Potential

Thorough needs assessments will help determine what types of digital repositories and services are most appropriate to Yale, will guide development priorities, and will inform efforts to encourage faculty participation. Questions to be addressed in a needs assessment from that perspective include:

Will faculty contribute materials to a centrally coordinated safe haven for digital resources? Do faculty value long-range stewardship and institutional support for their digital assets? Will participants follow proven standards and best practices? What kinds of support services are necessary for adding faculty driven content to repositories?

Needs assessments for digital collections built mostly by the library through digitization and collecting "born digital" materials will also need to be undertaken.

Examples from three methodological approaches for assessing needs and participation potential illustrate the ways that other institutions have approached the development and planning of digital repositories. Surveys, workplace studies, and use case analysis all offer methods of understanding local needs, user preferences, and technical requirements.

MIT Faculty Survey

MIT Libraries surveyed 93 faculty to determine interest in publishing materials in DSpace.⁴⁰ Questions focused upon scholarly publishing, making research available, searching, and preservation. There were considerable concerns and confusion about the role of institutional repositories as alternative publishing models: worries that submitting to DSpace would "constitute prior publication" and prevent researchers from submitting their work to journals (48% of responses); reluctance to assign MIT the distribution rights (46%); and the desire to have only formally published works available for public consumption (50%). (Interestingly, 43% reported distributing their preprint articles by posting them to their own websites; 18% posted to their department's or a disciplinespecific preprint site.) Yet 57% said there were important benefits to making research available "with very little effort on my part and without my having to maintain a website" and "69% said there were important benefits to "provid[ing] long term preservation of my digital research materials." And 72% felt it was important to "make it easy for other people to search for and locate my work." Faculty emphasized that they need a "safe, cost-effective searchable publishing solution" and they see a critical need for "legacy and recognition of leadership in their field."

University of Rochester Workplace Study

This workplace study was undertaken to better understand the work practices of faculty, and as a means of establishing better alignment of the institutional repository at the university. The project team, which consisted of an anthropologist, a computer scientist, a graphic designer, a programmer, catalogers, and librarians, carried out interviews and videotaped observations. Particular attention was paid to desktop, office, and computer use. They also took a survey of faculty concerns, finding that "very few" cared about the concept of an institutional repository or metadata or an open source approach to digital asset management. Overwhelmingly, however, they were concerned about controlling access to and ownership of their digital content. Only about 25% indicated that they were concerned about long-term preservation of their digital materials. Through observing workflow and content management problems, the team concluded that faculty need an authoring tool that functions as a document management system (versioning, coauthoring, backups, access control, accessible from multiple locations) and it was not clear that an institutional repository can provide solutions for the personal content management challenges observed in the study. Other conclusions: Faculty have copyright worries and err on the side of caution in regard to scholarly publishing and sharing resources. Not everyone understands the ability to self-archive post prints.

⁴⁰ MIT Libraries. "MIT Libraries' DSpace Business Plan Project," H<u>http://libraries.mit.edu/DSpace-mit/mit/mellon.pdf</u>H.

Academic disciplines are the primary orientation for faculty, not the institution. They think outside their departmental hierarchy and outside their institutional affiliation.

In response to the findings, the University of Rochester Library now deposits materials into the institutional repository for faculty. They also have added researchers' own private pages to the DSpace interface to allow for customization and personal branding.⁴¹

Use Case Study at Johns Hopkins (funded by the Andrew Mellon Foundation)

The Digital Knowledge Center (DKC) at Johns Hopkins University, working with the University of Virginia (UVA), the Massachusetts Institute of Technology (MIT) and the Sheridan Libraries, have been funded to make an evaluation of repository software and services such as e-publishing, e-learning, and digital preservation. Each system will be evaluated against a series of use cases.

Use cases are a tool that should show the "what" of the interactions between the users and the computer system. In producing a small number of use cases for functionality, the analysts and the users are forced to abstract the activities of the system until they truly represent what the system must accomplish.⁴²

Use cases are helpful in illustrating a range of practical challenges and illustrate the complexity of content flows and user requirements for integrating content and operating in an integrated landscape. The use cases will provide a realistic, comprehensive set of experiments using the repository software and services systems. The following list from Johns Hopkins⁴³ of possible repository roles illustrates the range of possible roles for repositories:

- Self-archiving: A researcher wants a place to put working papers and • similar documents, some of which would be available to the public, some of which would not. She will provide links to the content from her web site. The content will also be available via OAI harvesting.
- Species image repository: An organization wishes to provide a service with which field biologists can upload and make searchable images of various plant and animal species. Included in the field data will be georeferencing information. Information from this repository will be harvested to support a georeferenced species finder service.
- Library digital collections: The Sheridan Libraries at Johns Hopkins University wishes to ingest a large amount of sheet music (metadata and associated digital images) into a digital repository. The metadata uses a

⁴¹ Gibbons, "Aligning Content Recruitment Strategies,"

http://www.arl.org/sparc/meetings/ir04/presentations/gibbons.html.

⁴² Johns Hopkins University Digital Knowledge Center, "A Technology Analysis of Repositories and Services," H<u>http://dkc.mse.jhu.edu/repository.html</u>H⁴³ Ibid.

sheet music-specific format, which can be transformed to oai_dc for OAI harvesting, but should be directly accessible for OAI harvesting using a different metadataPrefix. It should be possible to see thumbnails and some metadata on a summary search results page and to view the complete sheet music on an 800x600 monitor.

- Learning objects: A university runs a learning management system. From that system, the University wishes to link to or access content in the repositories in scenarios (1), (2), and (3). Additionally, the University would like to store the learning objects created in the LMS into another repository, which also contains content to which it would like to link.
- E-learning: An instructor wants to store problems for a web based homework system in a repository. The problems themselves are expressed in a markup language that requires external applications to render. The instructor would like to be able to efficiently search for problems in order to create problem sets, and to be able to have the problems render appropriately when delivered to students' browsers.
- E-portfolio: A university has a policy that requires that students retain meaningful control over work that they produce for courses. The student wants to grant access privileges to various entities for some material she has created. These privileges may have different expiration times for the different entities.
- Publishing: a scenario similar to the self-archiving example from (1), except that there is a review (peer or otherwise) process introduced before the content enters the repository.
- Repository management: Operators of the repositories listed in these use cases need to be able to manage these facilities and the content. To do so, they will need to undertake activities such as integration with external services, format migrations, replication and/or backup, and inventory. For example, a repository manager may need to identify portions of an archive that might be at-risk, perhaps because a commercial entity makes an intellectual property claim regarding a file format. The repository manager then needs to develop tools to deal with a large-scale format migration. Repositories need to provide reporting facilities and interfaces that will support these activities.
- Distributed file organization: A researcher uses a personal information manager and P2P applications to manage her individual and academic files, email messages, calendar notices, research files, etc. and wishes to consider long-term archiving of these materials through repositories.

These use cases illustrate the complexity of the digital landscape with multiple content stores and multiple applications. The array of activities within the digital landscape requires a framework of interconnected repositories, services, and applications. The study emphasizes the need to support multiple repositories for different audiences, purposes, and disciplines.

A report by Christine Borgman on the 2003 NSF Workshop on Post-Digital Libraries Initiative Directions⁴⁴ emphasizes "the highly individualized ways of seeking and using information" and the "need (for) real data on users and uses that can be used to design a new generation of systems." They recommend the use of case studies "in multiple disciplines to determine what behaviors and requirements can be generalized across user groups and what requirements are individual and group specific."

A 2004 study sponsored by the Digital Library Federation by the Use Case Working Group chaired by Hoebelheinrich, Greenbaum, and Fern,⁴⁵ lists functional requirements based upon use cases exploring the scholar's workspace between content repositories and educational technologies. They explore the activities that faculty undertake to gather, create, and share digital sources and call for demonstration projects, collaborative partnerships, and explorations of the "end-to-end process of gathering materials from multiple repositories, creating new teaching and learning products from the materials, ... and then sharing/publishing materials for reuse in some sort of repository." The emphasis is upon demonstrations of user-driven needs that inform the cycle of use and reuse of digital resources (especially learning objects) within a complex repository and learning landscape.

The Johns Hopkins digital repository review project is applying specific functional requirements from use cases to evaluate specific interface protocols and repository systems. Among the systems being evaluated (according to the proposal to the Mellon Foundation) are: DSpace, Fedora, Greenstone, ePrints, WebWare, LOCKSS, DiVA, Virginia Tech electronic theses and dissertations software, Sakai, WebCT, WebWork, Internet Scout, Open Source Portfolio Initiative, and uPortal. Perhaps most importantly, this effort will create a greater understanding of the relative merits of these systems and provide a roadmap for enhancing interoperability among their services.

Their methodology is shown in the following table in which 10 specific functional requirements are extrapolated from four different use cases. ⁴⁶ Each green box is a required function; red boxes are not required for a specific use case. Thus, use case 2 requires Functions 3 and 7 through 10. The specifics of these are still in development.

 ⁴⁴ Borgman, "Personal Digital Libraries," H<u>http://www.sis.pitt.edu/~dlwkshop/paper_borgman.html</u>H.
 ⁴⁵ Flecker and McLean, "Digital Library Content and Course Management Systems." Appendix 4.0: Use

Case Working Group: Report and Recommendations. Chaired by Nancy Hoebelheinrich, David Greenbaum, and Jay Fern. Digital Library Federation. http://www.diglib.org/pubs/cmsdl0407/.

⁴⁶ These diagrams are from Choudhury and Martino, *A Technology Analysis of Repositories and Services*, http://www.cni.org/tfms/2005a.spring/abstracts/presentations/CNI_martino_technology.pdf.

	Use Case 1	Use Case 2	Use Case 3	Use Case 4
Function 1				
Function 2				
Function 3				
Function 4				
Function 5				
Function 6				
Function 7				
Function 8				
Function 9				
Function 10				

Functional Requirements of Use Cases

The next table shows how the functional requirements and use cases are put into a decision grid for evaluating specific applications. The Johns Hopkins team produced the following matrices to show which use cases require which functions, and also which functions the various repositories and interfaces support.

	UC1	UC2	UC3	UC4	170	ΟΚΙ	IMS	DS	FED	CDM
Function 1										
Function 2										
Function 3										
Function 4										
Function 5										
Function 6										
Function 7										
Function 8										
Function 9										
Function 10										

Functional Requirements, Use Cases, Interfaces, and Repository Systems

The first 4 columns are four different use cases; the next three columns are examples of interfaces standards. (The ones here are the JSR-170 interface; the OPen Knowledge Initiative, and the IMS Meta-data Standard for Learning Object Metadata). The rows are functions (not defined yet– these have to come from use cases which are still in

development.) Use case 1 can use any of the three repositories and any of the three interfaces, since its required functions are just the ones numbered 1-3. Use case 2 requires functions 3,7,8,9 and 10, and as such is only supported (in this example) by the JSR-170 interface and the Fedora repository.

The challenges with this approach are that the scenarios evolve quickly, tools are developing quickly, there is an explosion in the number of applications and complexity of web based systems, protocols are not mature, and it doesn't cover content standards or content flows. However, the strategy of combining user informed cases of content flows and system requirements with evaluation of interfaces and repository systems should produce useful evaluation outcomes and serve as a model for ways that Yale might apply particular scenarios from its specific learning and research landscapes.

VI. Recommendations

1. Set up a test digital repository using FEDORA and VITAL. (in progress)

Use content and a participation model to test the implementation of two digital repository goals: providing access for repurposing for diverse content types and for addressing long-term preservation strategies. Issues to be explored include file formats; metadata for discovery, rights management, and preservation; storage for access and preservation; and cooperative development.

"The library should engage in the development and implementation of digital repositories that support the repurposing of content in portal and course management systems university wide." (Library's Integrated Access Council's Portal Opportunities Group Final Report, Recommended Policies, June 2005)

By establishing a repository solution along with associated services that have the flexibility to fulfill needs as they develop, the library can entice projects and individuals to participate in the central repository in part because it is more efficient for them than developing their own solutions. If we can show the benefits in an operational Fedora system, the recommended practices will more likely be adopted along with the convenience and cost-savings associated with shared tools.

- This development should extend, be informed by, and be coordinated with the Library's Rescue Repository.
- Develop the ability for courses to use digital objects from the test DR in classes*v2 (Sakai).
- Develop integration of Fedora and Sakai with external partners; for example, seek a research and development project with efforts at Johns Hopkins University.
- Test and develop user-friendly mechanisms for self-service creation of collections that encourage recommended practices for content and metadata production.

- Explore the quality of OAI-PMH services provided by the system.
- Potential participants:
 - Yale Finding Aids (Beinecke, Divinity, MSSA, Music, etc.)
 - Yale Daily News archive (including newspaper interface)
 - Images (backend to Insight and the DL and possible Medical image collections)
 - Digital content from social sciences and sciences (especially datasets and digital objects from research at Yale, focusing on large-scale data management and custom web applications)
 - Yale Indian Papers Project
 - Human Relations Area Files
 - OACIS/AMEEL Middle East project
 - McClintock Collection (TEI and images)
 - Audio and Video from MSSA
 - Learning objects from classes*v2 and/or ELI Program
- Build expertise and models of support.
- Coordinate development with existing research infrastructure and services for faculty that are already in place, e.g. research.yale.edu and institution file space (Pantheon storage, SAN storage, blogs, e-mail services, etc).
- Coordinate test repository development and use cases with ELI and Davis grant projects and AM&T's Instructional Technology Group support services
- Investigate the inclusion of University administrative research programs digital output (e.g. customer satisfaction surveys, user assessment studies)
- 2. A cross-unit team should develop policies, tools, workflows, and good practices for digital collection building in order to promote interoperability and sustainability by multiple creators and compilers across the university, including faculty projects and research center based digital productions.
 - Identify and develop the types of services that are needed for building digital objects and collections that can be integrated into the larger teaching and research landscape. Digital asset production design and policies should look beyond the production of separate silos. Storing content in generalized, non-fixed formats ("simple" or "dumb" repositories) gives us a better chance of handling that content over time. We must plan for content being used within and among multiple interfaces and repository environments and by several applications. At the same time, lower barriers for participation, encourage entrepreneurial development, and offer customized solutions.

- Articulate the benefits and methods of participating in an integrated access view of digital objects including cross-domain search and discovery, reuse of digital assets in multiple applications, and the ability to move digital assets into preservation contexts as required.
- Articulate standards and best practices for digital assets, repository interfaces, and applications. Articulate the benefits of and build tools that support the life-cycle view of digital assets. Develop best practice guidelines for digital production processes for digital assets. Explore ways of optimally partnering with faculty from the beginning of their data collection and digital production efforts.
- Address the long-range stewardship requirements and costs of taking on an institutional commitment to the long-term preservation of repository materials. Consider a continuum of stewardship requirements, from minimum to maximum.
- Review the life-cycle process of different types of digital assets, and build a model of when and how they can be integrated into a more managed environment—and where that managed environment might best be developed and managed. The model should address centralized versus decentralized aspects of the environment. Legacy data—that is, digital assets that currently exist at Yale but are in environments of varying stability—should be included in this review.
- Create a digital collections registry where faculty, the Library, ITS, etc can communicate information about special collections, web applications, databases, etc.

3. Develop deep understanding of user requirements based upon use case studies and other assessment methods.

- The Library should engage in usability studies and establish metrics using survey / feedback tools that support user centered design of all new integrated library services.
- Build use cases and scenarios to understand multiple users' needs and content flows, to evaluate repository options, to articulate interface requirements, and to define functional requirements.
- Explore the expectations of the faculty and campus units regarding the life span requirements of digital resources (including preservation and access as required by funding sources) through use case studies.
- Mine the expertise of projects and programs across the university that interact with and include faculty in creating and maintaining digital assets and collections.
- Assess activity at Yale in e-scholarly publishing, open access journals, etc.

- Develop a clearinghouse to record and share the information gathered from the various assessment methods described above.
- Clarify what support services and applications could be provided to faculty during the research and teaching processes in order to evaluate potential participation in the digital repository and create the potential to migrate and preserve Yale's scholarly materials.
- Develop a multi-dimensional model of requirements based upon the expected life span of digital resources. The model should extend from personal collections without long-term preservation requirements to fully archived resources, and should address stewardship requirements and support costs.

4. Support and participate in the development of an Information Architecture Plan for Yale University, with digital repositories as an element of the overall information management strategy.

For those organizations within the university concerned with stewardship—we think immediately of libraries, archives, and museums but should recognize there are also huge numbers of academic units that curate collections of information—it should be clear that institutional repositories raise complex and nuanced questions about organizational roles, responsibilities, resources, and strategies. (Lynch)⁴⁷

Further targeted analysis is necessary to understand the role of digital repositories and their related services within the overall teaching and research landscape at Yale. A digital repositories strategy is an essential component in an integrated information architecture that incorporates the workflows and digital assets of research and teaching—the creative activity that includes the output, sharing, and reuse of the components of knowledge.

Repositories must be considered within the context of the wider integrated information environment and the complex interactions between repositories and other components of the information environment. We need to analyze and map workflows among services, digital storage, interfaces, applications, and other components of the information landscape and develop future services that can interoperate within well-structured workflows between these components. Efforts need to be made to organize communities of interest and articulate priorities, functional requirements, and overall goals for developing digital repositories.

Without an information architecture plan and vision, digital repositories will not operate within the larger landscape, they may not be used, and they will not provide the institutional context for the benefits outlined in this report. A primary goal should be to integrate information into a common digital landscape.

⁴⁷ Lynch, "Institutional Repositories." http://www.arl.org/newsltr/226/ir.html.

As described in Flecker and McLean's *Digital Library Content and Course Management Systems: Issues of Interoperation*, repositories and other digital resources should:

- Make themselves known to operators of learning applications in expected ways
- Follow standards and best practices in terms of access, search, metadata practices, and download support (such as the Open Archives Initiative Protocol for Metadata Harvesting).
- Document their systems and policies so that others can configure their systems appropriately to interoperate with them.

Current initiatives in the Library are addressing these issues of interoperability:

- the embedding of library metadata and content in the Sakai learning environment through various means described in the IAC's Portal Opportunity Group's report, for example, the MetaLib X-Server;
- coordinating the adoption of emerging standards and best practices in the implementation of OAI for the many independent initiatives interested in pursuing this technology at Yale.

VII. Appendices

Appendix 1: Digital Repository Examples

Massachusetts Institute of Technology⁴⁸

Model: Faculty focused publications, centralized curation, institutional branding. Content is identified and added to the repository by "communities" (e.g., departments, labs, and research centers). Primary goal is to provide access to MIT digital resources through a centralized institutional repository; to "capture, distribute and preserve the intellectual output of MIT and to offer the opportunity to provide access to all the research of the institution through one interface."

Technology used: DSpace

Content: Guidelines specify that material must be "education-oriented," in digital format, and produced by an MIT faculty member. The communities choose collections that are made up of items. Items are sets of contents (can be multiple formats) and accompanying metadata. Items are similar in some dimension, e.g. purpose, subject, audience. Items can be working papers, video clips, materials for a course lecture, a research paper with a dataset or media images, etc. Content items are in various formats

⁴⁸ In part based upon notes from a meeting with MacKenzie Smith and Margret Branschofsky at MIT. January 10, 2005.

(text, images, datasets, videos, etc) and generally are bundled by some common dimension (purpose, like working papers, subject or intended audience).

Funding: MIT and Hewlett Packard have invested a significant amount of effort in developing DSpace. The repository is organized to accommodate the varying policy and workflow issues inherent in a multi-disciplinary environment. It is based upon participation by a wide variety of organizations across the institution. Submission workflow and access policies can be customized to adhere closely to each community's needs.

Roles: MIT Libraries working with numerous communities at MIT.

Progress to date: As of June 2005, there were over 15,000 titles in the MIT DSpace repository. MIT Libraries received a grant from the Andrew W. Mellon Foundation in the fall of 2002, and the DSpace Federation project was launched in January 2003. This project had as its primary goal to establish the DSpace Federation by close collaboration with a small number of universities who would act as testers, advisors, collaborators, and hopefully adopters of the DSpace platform. The Federation has the task of learning whether and how well DSpace might work outside the MIT environment. The initial partners included:

- Columbia University
- Cornell University
- Massachusetts Institute of Technology
- Ohio State University
- University of Rochester
- University of Toronto
- University of Washington

"A brief study of the content represented by those fifty live DSpace sites shows that there are approximately sixty thousand digital items available online with this technology, most of it free to the public, and the number is rising steadily. For a project that is not yet two years old, this is quite remarkable progress and demonstrates both the clear demand for such technology, and the level of interest from the academy to share its research and teaching material, and to get it under long-term curatorial control."⁴⁹

California Digital Library: eScholarship Repository

Model: Explicit strategy to "influence scholarly communication and provide a publishing platform for electronic journals, and leverage library buying power."⁵⁰ Commercial software installation is used to support the scholarly publishing efforts among a consortium of universities. Notification services are offered to authors regarding the number of downloads and alerts from readers, and alerts to the contributing units.

⁴⁹ MIT Libraries, "Final Report on the Initial Development of the DSpace Federation," <u>Hhttp://www.DSpace.org/federation/mellon-DSpace.pdf</u>H

⁵⁰ Ober, "IR Content and Service Expansion: the Case of eScholarship," http://www.arl.org/sparc/meetings/ir04/presentations/ober.html.

Technology used: The system uses Berkeley Electronic Press software (http://www.bepress.com) licensed by the University of California.

Content: Faculty-produced content across multiple institutions with centralized curation responsibilities. All current faculty, staff, or students at the University of California who own the copyright for their works or are in compliance with the original publisher's policy on the electronic distribution of postprints, or have express permission from the publisher may register for an account and deposit items into the repository. "The University of California's eScholarship Repository offers faculty on the 10 UC campuses a central facility for the deposit of research or scholarly output. Individual research centers, departments, and sponsoring units set the policies for acceptance of content. Determination of acceptable content is in the hands of researchers and faculty."⁵¹ Research and scholarly output is selected and deposited by the individual University of California units. Also, any University of California research unit, center, or department may receive an account and training in using the system. Direct marketing is done via personal invitations and encouragement from Vice Chancellors. The project is also harvesting citations for post-prints and sending invitations to authors to submit their materials to the repository via a simple submission form.

Funding: Sponsored by the California Digital Library within the University of California Office of the President.

Roles: The eScholarship repository has been developed by the California Digital Library in close collaboration with the University of California Press. Each of the 10 UC campuses has an eScholarship liaison.

Progress to date: As of June 2005, there were approximately 7,500 scholarly papers in the eScholarship Repository.

Harvard Science Digital Library (HSDL): DSpace implementation⁵²

Model: Local faculty-produced content, scholarly publishing initiative, central curation, library and departmental support. The HSDL focuses upon supporting alternative scholarly communications. Communities work with professional librarians who add content to DSpace. The model relies on collaboration with department-based technical staff and is hosted on departmental servers. This is a pilot project using DSpace that emphasizes support for the transformation of scholarly communication, increased access to scholarly materials, promoting the continuing existence of scholarly objects, and adding Harvard branding to the digital materials. The HSDL also is operating as a publishing platform by setting up "virtual journals" in DSpace. They also are developing a desktop tool for faculty to help with the processes of preparing articles for publication. Some peer review templates are also supported.

⁵¹ Drake, "Institutional Repositories," H<u>http://www.infotoday.com/searcher/may04/drake.shtml</u>H.

⁵² Based upon notes from a meeting with Michael Leach and David Osterbur at MIT. January 10, 2005.

Technology used: DSpace. The system provides search support and harvesting via the OAI-PMH protocol and provides links to other digital collections.

Content: Content types that are included are theses, videos from colloquia and seminars, research articles, datasets, and electronic journals. To acquire more material for the repository, the staff plans to run open access searches to identify Harvard faculty publications, acquire permission to include them in the HSDL, and add the publications and metadata to the HSDL. Access to the digital materials is open to the public now, but they will eventually also offer "Harvard only" and research community limitations.

Funding: The business model is based upon free services offered through a Science Libraries consortium; funding from the Physics Department provides technical support (including equipment).

Roles: This is in an environment where there already is a great deal of partnering with faculty, strong local library presence and collaboration with departmental resources. Existing staff was reorganized to add support into regular operations; the Library produces metadata for all materials going into the DSpace repository.

Progress to date: By early 2005 they expect to have about 1,000 objects in the HSDL. Future plans include collaborations with DSpace, incorporating ICommons learning objects and using the METS integration for structural metadata. HSDL has also expressed interest in "intellectual genealogy" by tracing faculty and graduate students over time and generating "trees" of connections among dissertations and publications. HSDL plans future collaboration with the Harvard Digital Repository Service (see below). They have found that the biggest hurdles have been policy development, in regard to community based policies and intellectual property issues.

Georgia Institute of Technology: DSpace aggregation

Model: Local faculty based content; library harvests metadata from decentralized DSpace servers; gathers metadata and selected content from multiple DSpace instances on campus and "pulls" selected objects for Library's repository.

Technology used: DSpace

Content: A number of Georgia Tech's departments and labs are using DSpace for relatively ephemeral materials, in some cases more as workspace than as a repository. Some of their materials would not be appropriate candidates for long-term access and dissemination (such as problem solutions and incomplete data sets). However, they will be producing and collecting materials and also repackaging materials into discrete learning objects that the Library does want to collect.

Funding: Service provided by Georgia Tech's Library and Information Center.

Roles: The Library's role is to serve as the aggregator for these materials, and the Library will provide long-term access and dissemination and serve as the repository of record for campus. Librarians select materials from the harvested metadata from local DSpace repositories, and then pull objects that they select for collections. For objects they don't ingest, they plan to use OAI-PMH to link to the item still residing on the other DSpace servers. Note that they are also setting up a usability lab with their Human Centered Computing Lab to assess the DSpace interface.⁵³

Progress to date: As of July 23, 2005, there were 4,984 titles in the SMARTech repository.

University of Michigan: DSpace Pilot

Model: Central DSpace institutional repository pilot being used for managing formal digital collections, primarily those produced by Library digital production services. Not focused upon faculty participation or faculty publications.

The digital repository effort in the University of Michigan Library is a prototype pilot test of a single repository instance of DSpace. Primary purpose is to "gather the output of ongoing campus initiatives." It is not intended to "replace the peer-reviewed, scholarly journal model for publication (a departure from SPARC's vision.)"

Technology used: DSpace

Content: Carefully selected categories of materials, based upon well-defined existing digital collections. Personal faculty or student collections are not being considered at this time. Selectors were asked to nominate content; these submissions were sorted into a short list for implementation. They are seeking crossover with archives and course management resources, and a wide discipline range. The institutional repository is "format blind"—they are not limiting content by any particular format requirements but rather are including a range of digital format types. They will be producing "modest" metadata. Access to the IR content may be restricted to particular communities (i.e. it is not all publicly accessible and it is not all University of Michigan accessible) but all metadata will be exposed for harvesting. The IR will not require exclusive ownership of intellectual property, but will seek shared ownership. They are focusing upon the following categories of materials that have a high level of uniqueness to the university and strong relationships to campus activities:

- dissertations (PhD and Masters)
- Research Institute papers
- Business School videos/multimedia
- Performing Arts Technology performances
- Library collections built from departmental collections (research and technical reports)

⁵³ Jannik, "SMARTech: Georgia Tech's Institutional Repository,". H<u>http://www.arl.org/sparc/meetings/ir04/presentations/Jannik.html</u>H The DSpace website may be viewed at: http://smartech.gatech.edu:8282/dspace/

• Digital Library Production Services (DPLS) content⁵⁴

Funding: They have two-year initial start-up funds.

Progress to date: Test installation of DSpace is in place, initial content has been added.

University of Virginia

Model: Focus is on building a federated repository management system. "The University of Virginia Library is attempting to solve four problems with their Fedora implementation: management of complex objects that are organized in potentially multiple hierarchical structures; management of highly disparate data types and their preservation requirements; building virtual collections by recording and identifying relationships between objects in the repository; and the collection of born-digital faculty projects that incorporate new and reused materials into new scholarly contexts. Fedora was chosen because it was architected to facilitate handling of complex objects. UVa is using Fedora to build a digital asset management system for their locally digitized collections and a discovery service that integrates with the repository by calling disseminators underneath the interface to present the objects and exploit their associated behaviors."⁵⁵ According to Tim Sigmon, ITC's Director of Advanced Technology: "The beauty of Fedora is that it allows you to store and manage your digital assets for multiple purposes. Information stored in the Tibetan website, for example, can also be used for a course. That's what we're heading toward: one repository that multiple environments point to."56

Technology used: Fedora. "The UVa repository uses Cocoon as its XML pipeline and XPAT for indexing. At the moment, the digital library at UVa is a standalone system, but in the course of the next year, they will be implementing a metasearch application (SingleSearch from SirsiDynix Corp., using a custom plug-in) that will include the digital repository as a search target alongside the Library's OPAC and licensed e-journals."⁵⁷

⁵⁴ Information based upon meetings at the University of Michigan. October 20, 2004. Digital Library Production Service (DLPS) at the University of Michigan: This service is involved with digitization projects and preservation reformatting, including scanning printed volumes, and optical character recognition of 4-5 million pages per year with TEI encoded page breaks. The Scholarly Publishing Office (SPO) at the University of Michigan is an e-publishing center developing tools and methods in regard to publishing services, subscriptions and licensing, intellectual property, and local control of university assets. They offer highly functional services, including full text search and retrieval based upon the locally developed DLXS system. There is currently no connection with the institutional repository pilot, but that is being planned. The DLPS at the University of Michigan reformats Library content into DLXS. SPO publishes new content into DLXS. Also at the University of Michigan is the OAI-ster project: They harvest from over 350 providers and have the ability to test OAI data provider compliance with official registry. They are developing best practices for data providers in creating OAI metadata and getting help with OAI harvesting.

⁵⁵ The Fedora Project. About Fedora Users: The University of Virginia Library. http://www.fedora.info/about/uva.shtml

⁵⁶ Church, "Hands Across the World," http://www.itc.virginia.edu/virginia.edu/spring05/hands.htm

⁵⁷ The Fedora Project. http://www.fedora.info/

Content: The UVa Fedora repository currently contains legacy image and text collections, but plans are under way for additional content types including data sets and digital video. "The three current collection types are: art and architecture images, EAD finding aids, and TEI electronic text. Challenges still to come include tool development for the presentation of data sets and their associated code books, and designing content models for virtual collections, which will be a factor in how they collect faculty projects—collecting both the content and the logic behind the intellectual collections." ⁵⁸

Progress to date: The official release of the Repository is expected for Fall 2005.⁵⁹

Northwestern University

Model: Federated repository management, including complex objects and websites.

"Northwestern University's Academic Technologies group has been partnered with the Fedora Development Team since phase 1 of the Fedora project. The Academic Technologies group's Fedora efforts center around managing complexity. They have web sites or collections, in which the objects themselves are complex, or in which the relationships between objects are complex, and the presentation of these objects is varied, but they need to provide consistent access. ... They liked the opportunities afforded by the formalisms that the Fedora object model provides based on standards, as well as the opportunity that the software gave them to exploit existing and future functionality, in particular, Fedora's dissemination mechanism. In general, Fedora afforded them room to grow."⁶⁰

Technology used: Fedora. "In general, the Academic Technologies Group has written middleware layers above the Fedora core software which interact with the Fedora APIs, including Java programs that do bulk ingest and behaviors for images that drive zooming clients and allow users to retrieve specific parts of an image at a specific size."⁶¹

Content: "There are three Fedora deployments. The first is a general deployment brought up for test purposes. This deployment is used to test their object models, is being used for a class, and contains a collection being used for development on a project with three other institutions. The second is specific to a project with the Chicago Historical Society, in which the *Encyclopedia of Chicago* is being presented via the Web. And third, Chris Karr has set up a repository for a joint project between NWU-Academic Technologies, Michigan State University, the National Archives and Records

Presentation at the Fedora Users Conference, May 2005.

⁵⁸ Ibid.

⁵⁹ Information about the Fedora implementation at UVA may be found in: Staples, Thornton. "Fedora: Digital Library Use Cases and Applications" Presentation at the Danish Technical University Copenhagen August 27, 2004. <u>Hhttp://defxws.cvt.dk/events/seminar040827/FedoraThornton.ppt</u>H and Johnston, Leslie. "Digital Library Repository Development with Fedora at the UVa Library,"

Hhttp://www.scc.rutgers.edu/fedora conf 2005/charts/Johnston.pdfH

⁶⁰ The Fedora Project. About Fedora Users: Northwestern University's Academic Technologies Group. <u>Hhttp://www.fedora.info/about/nwu_at.shtml</u>H

⁶¹ Ibid.

Administration, Glasgow Caledonian University, and the BBC -- Information and Archives which will make digital audio archive materials available for use by students and faculty at the partner institutions."⁶² "Northwestern University has been working with FEDORA over the past year, creating object models for images and spoken word audio. We are currently in the process of integrating Fedora with our Project Pad system. The vision is that Project Pad will interoperate with a FEDORA archive and provide a web-based graphical view and search of the repository and its contents."⁶³ "Project Pad integrates tools for searching for - and organizing materials from digital repositories; for annotating and analyzing those materials; for integrating those materials and analyses into a scholar's (or a student's) web-based projects; and for group exchange and decision making. Unlike other web-based collaborative systems Project Pad is designed to support synchronous interaction between users and lets users download, keep copies of, and distribute their work."⁶⁴

Roles: Academic Technologies and Northwestern University Library

Rutgers University

Model: Build a trusted digital preservation process within a digital preservation architecture based upon Fedora.

Technology used: Fedora. "The Rutgers University Libraries digital repository is made up of cutting-edge SAN hardware and Hierarchical Storage Management (HSM) software, which essentially allows the repository server to address both the disk array and the tape library as though they were connected disks. The HSM software permits repository administrators to more effectively manage the storage system, optimizing usage of the disk and ensuring data is backed up via policies to store all presentationform datastreams on disk and release rarely-accessed archival datastreams to tape. The hardware and software work together to ensure lost or mangled data may be regenerated to form a 'suitable facsimile' should such an event occur."⁶⁵

Content: Text, images, sound recordings, numeric data sets. The New Jersey Digital Highway (NJDH).

Roles: Scholarly Communication Center (SCC) of the Rutgers University Libraries (RUL)

Harvard University: Digital Repository Service 66

⁶² Ibid.

⁶³ Information about the Spoken Word and Fedora project at Northwestern may be found at http://www.at.northwestern.edu/spoken/p05fedora.html

⁶⁴ Information about Project Pad may be found at

http://www.at.northwestern.edu/spoken/p06architecture.html

⁶⁵ Information about the Fedora work at Rutgers University is based upon: Jantz, Ron, and Michael J. Giarlo. "Digital Preservation: Architecture and Technology for Trusted Digital Repositories." *D-Lib Magazine* 11, no.6 (June 2005). http://www.dlib.org/dlib/june05/jantz/06jantz.html

⁶⁶ Based upon notes taken at a meeting with Dale Flecker, Harvard University. December 3, 2004.

Model: Digital repository and digital preservation infrastructure based services for formal collections; no e-publishing focus. The Digital Repository Service will accept a wide range of formats but preservation services are limited to specific formats. The repository is a separate object management service, distinct from search services, and separates the discovery and rendering of objects from the storage of objects. The DRS provides a digital storage facility, management of administrative and structural metadata, preservation policies and procedures, and delivery of an object to a registered application (like an online catalog).

Technology used: The system currently provides a "home grown" preservation storage layer and does not include search, display, metadata support, etc. "Trusted services" are layered on top of the digital repository infrastructure; trusted delivery apps check access rules stored in the DRS then deliver objects to applications. Communities and services outside the DRS develop trusted search and delivery applications that pull resources from the DRS in their own specific ways for specific purposes.

Content: The repository will accept a wide range of formats but preservation services are limited to specific formats. The levels of service range from "no loss" to "some loss" and to "no transformation." Currently materials go in and do not come out; there is no ability to delete or remove objects. Museums are the largest contributors to the repository (TIFFs and JPEGs). This is a storage service to keep bits safe, provide an environment and tools for digital preservation activity, and enforce security measures for digital objects. There are no descriptive metadata in the repository; it does contain administrative metadata (ownership and access information), and structural metadata by format (i.e. the object relationships needed for applications to render the digital assets, for example a page turning service).

Funding: The business model is based upon cost recovery; pay by the gigabyte. Only "depositing agents" may put materials in the DRS on behalf of content providers; these are labs and library units.

Roles: Harvard University Library, Office for Information Systems manages the service, curators and collection managers from Harvard organizational entities contract for use of DPS. "Any Harvard organizational entity is eligible to use the DRS. An organizational entity might be a library, a museum, an archive or a department within such an organization. DRS documentation refers to these organizational entities as 'object owners.' Individual members of the Harvard community are welcome to make inquiries, but DRS participation requires a Harvard organizational sponsor."⁶⁷

Progress to date: DRS current size is 8-10 terabytes.

⁶⁷ Harvard University Library. Office for Information Systems. "Overview: Digital Repository Service (DRS)." http://hul.harvard.edu/ois/systems/drs/

Appendix 2: Bibliography

Beagrie, Neil. "Plenty of Room at the Bottom? Personal Digital Libraries and Collections." *D-Lib Magazine* 11, no. 6 (June 2005). http://www.dlib.org/dlib/june05/beagrie/06beagrie.html.

Besser, Howard. "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries." *First Monday*7, no. 6 (June 2002). <u>http://firstmonday.org/issues/issue7_6/besser/index.html</u>.

Borgman, Christine L. "Personal Digital Libraries: Creating Individual Spaces for Innovation." Paper presented at the NSF Post Digital Library Futures Workshop, Chatham, MA, June 15-17, 2003. http://www.sis.pitt.edu/~dlwkshop/paper_borgman.html.

Campbell, Lorna M., Kerry Blinco, and Jon Mason. *Repository Management and Implementation*. DEST (Australia) and JISC-CETIS (UK), 2004. <u>http://www.jisc.ac.uk/uploaded_documents/Altilab04-repositories.pdf</u>.

Cervone, H. Frank. "The Repository Adventure." *Library Journal* 129, no. 10 (June 1, 2004).

Choudhury, Sayeed, and Jim Martino. *A Technology Analysis of Repositories and Services*. Presented at theCNI Task Force Meeting, Washington DC, April 5, 2005. http://www.cni.org/tfms/2005a.spring/abstracts/presentations/CNI_martino_technology.p df

Choudhury, Sayeed, Tim DiLauro, and James Martino. "Choice and Empowerment." *Campus Technology: Newsletters: Technology-Enabled Teaching/eLearning Dialogue*, n.d.<u>http://www.campus-technology.com/news_article.asp?id=10507&typeid=156</u> (accessed July 5, 2005).

Church, Melinda. "Hands Across the World." *virginia.edu: The Information Technology Magazine at the University of Virginia*, Spring 2005. http://www.itc.virginia.edu/virginia.edu/spring05/hands.htm

Cornell University Libraries. *Report of the CUL Task Force on Open Access Publishing*. Presented to the Cornell University Library Management Team August 9, 2004 http://DSpace.library.cornell.edu/bitstream/1813/193/3/OATF_Report_8-9.pdf.

Crow, Raym. *The Case for Institutional Repositories: A SPARC Position Paper*. The Scholarly Publishing & Academic Resources Coalition 2002. http://www.arl.org/sparc/IR/ir.html.

———. SPARC Institutional Repository Checklist & Resource Guide. Release 1.0.

The Scholarly Publishing & Academic Resources Coalition, 2002. http://www.arl.org/sparc/IR/IR_Guide_v1.pdf.

Dempsey, Lorcan, Eric Childress, Carol Jean Godby, Thomas B. Hickey, Andrew Houghton, Diane Vizine-Goetz, and Jeff Young. "Metadata Switch: Thinking About Some Metadata Management and Knowledge Organization Issues in the Changing Research and Learning Landscape." In *E-Scholarship: A LITA Guide*, ed. Debra Shapiro. February 2005 preprint available at

http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf.

DiLauro, Tim. "Choosing the Components of a Digital Infrastructure." *First Monday* 9, no.5 (May 2004). <u>http://www.firstmonday.org/issues/issue9_5/dilauro/index.html</u>.

Drake, Miriam A. "Institutional Repositories: Hidden Treasures." *Searcher: The Magazine for Database Professionals* 12, no. 5 (May 2004). http://www.infotoday.com/searcher/may04/drake.shtml.

Feltner-Reichert, Melanie, Cindy Gray and Thura R. Mack "Training the University Community for Self-Archiving in Institutional Repositories." Paper presented at the EDUCAUSE Southeast Regional Conference, Atlanta, GA, June 6-8,2005. http://www.educause.edu/LibraryDetailPage/666?ID=SER0544.

Flecker, Dale. "Digital Archiving: What is Involved?" *EDUCAUSE Review*, January/February 2003. http://www.educause.edu/ir/library/pdf/erm0316.pdf.

Flecker, Dale, and Neil McLean. *Digital Library Content and Course Management Systems: Issues of Interoperation*. Digital Library Federation, 2004. Last updated Wednesday, September 1, 2004. <u>http://www.diglib.org/pubs/cmsdl0407/</u>.

Flecker, Dale, Leslie Johnston and David Greenbaum. *Digital Library Repositories and Instructional Support Systems. Project Briefing: Spring 2004 Task Force Meeting.* Coalition for Networked Information, 2004. <u>http://www.cni.org/tfms/2004a.spring/abstracts/PB-digital-flecker.html</u>.

Foster, Nancy Fried, and Susan Gibbons. "Understanding Faculty to Improve Content Recruitment for Institutional Repositories." *D-Lib Magazine* 11, no. 1 (January 2005). http://www.dlib.org/dlib/january05/foster/01foster.html.

Gibbons, Susan. "Aligning Content Recruitment Strategies with Faculty Work Practices." Presented at the SPARC/SPARC Europe Workshop on Institutional Repositories: The Next Stage. November 18–19, 2004. Washington, D.C., http://www.arl.org/sparc/meetings/ir04/presentations/gibbons.html.

Harvard University Library. Office for Information Systems. *DRS Policy Guide*, revised October 5, 2001. <u>http://hul.harvard.edu/ois/systems/drs/policyguide.html</u>.

Hedstrom, Margaret. "Digital Preservation: a Time Bomb for Digital Libraries." University of Kentucky, n.d. <u>http://www.uky.edu/~kiernan/DL/hedstrom.html</u>.

Hey, Jessie M. N. "Targeting Academic Research with Southampton's Institutional Repository." *Ariadne* 40 (July 2004). http://www.ariadne.ac.uk/issue40/hey/.

IMS Global Learning Consortium. *IMS Digital Repositories v1.0 Final Specification*. January 30, 2003. <u>http://www.imsproject.org/digitalrepositories/</u>.

Jannik, Catherine. "SMARTech: Georgia Tech's Institutional Repository." Presentation at the SPARC/SPARC Europe IR Workshop "Institutional Repositories: The Next Stage," Washington, DC, November 18-19, 2004. http://www.arl.org/sparc/meetings/ir04/presentations/Jannik.html

Jantz, Ron, and Michael J. Giarlo. "Digital Preservation: Architecture and Technology for Trusted Digital Repositories." *D-Lib Magazine* 11, no.6 (June 2005). http://www.dlib.org/dlib/june05/jantz/06jantz.html

Johns Hopkins University Digital Knowledge Center. A Technology Analysis of Repositories and Services: A Proposal Submitted to the Mellon Foundation. http://dkc.mse.jhu.edu/repository.html. Accessed July 23, 2005.

Johnson, Richard K. "Institutional Repositories: Partnering with Faculty to Enhance Sholarly Communication." *D-Lib Magazine* 8, no.11 (November 2002). http://www.dlib.org/dlib/november02/johnson/11johnson.html.

Johnston, Leslie. "Digital Library Repository Development with Fedora at the Uva Library." Presentation at the Fedora Users' Conference, Rutgers University, May 13-14, 2005. http://www.scc.rutgers.edu/fedora_conf_2005/charts/Johnston.pdf

Joint Information Systems Committee (JISC), Digital Repositories Programme. *Digital Repositories Review*. February 2005. <u>http://www.jisc.ac.uk/uploaded_documents/repreview-final-20050220.pdf</u>

Digital Repositories Review: Annex 1: Focus Group Report

http://www.jisc.ac.uk/uploaded_documents/rep-review-Annex1-fg.pdf.

Digital Repositories Review: Annex 2: JISC Digital Repositories Review Software Developers Survey. <u>http://www.jisc.ac.uk/uploaded_documents/rep-review-Annex2-</u>software.pdf.

Digital Repositories Review: Annex 3: Repository Issues...from a Teaching and Learning Perspective. <u>http://www.jisc.ac.uk/uploaded_documents/repos_issues_cetis_feb05.pdf</u>

Jones, Richard. "DSpace vs. ETD-db: Choosing Software to Manage Electronic Theses and Dissertations." *Ariadne* 38 (January 2004). <u>http://www.ariadne.ac.uk/issue38/jones/</u>.

Lagoze, Carl, Sandy Payette, Edwin Shin, and Chris Wilper. "Fedora: An Architecture for Complex Objects and their Relationships." Preprint, version 4, March 7, 2005. <u>http://www.arxiv.org/abs/cs.DL/0501012</u>.

Lynch, Clifford. "Check Out the New Library." Interview in *Ubiquity: an ACM IT Magazine and Forum* 4,no. 23 (July 30-August 5, 2003). http://www.acm.org/ubiquity/interviews/pf/c_lynch_1.html

———. "Where Do We Go From Here? The Next Decade for Digital Libraries." *D-Lib Magazine* 11, no. 7/8 (July/August 2005). http://www.dlib.org/dlib/july05/lynch/07lynch.html.

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

Lynch, Clifford, and Neil McLean. "Interoperability between Library Information Services and Learning Environments—Bridging the Gaps." A Joint White Paper on behalf of the IMS Global Learning Consortium and the Coalition for Networked Information, May 10, 2004. http://www.imsglobal.org/digitalrepositories/CNIandIMS_2004.pdf.

Mackie, Morag. "Filling Institutional Repositories: Practical Strategies from the DAEDALUS Project." *Ariadne* 39 (April 2004). http://www.ariadne.ac.uk/issue39/mackie/.

Marcum, Deanna B. "Preservation of Scholarship: The Digital Dilemma." EDUCAUSE, Forum for the Future of Higher Education, 2002. <u>http://www.educause.edu/ir/library/pdf/FFPIU028.pdf</u>.

McCord, Alan. "Institutional Repositories: Enhancing Teaching, Learning, and Research." EDUCAUSE Evolving Technologies Committee, October 16, 2003. http://sitemaker.umich.edu/dams/files/etcom-2003-repositories.pdf.

McLean, Neil. "The Ecology of Repository Services: A Cosmic View." Presented at ECDL 2004, the 8th European Digital Library Conference, Bath, England: September 12-17, 2004. <u>http://www.ecdl2004.org/presentations/mclean/n-mclean.pdf</u>.

Metros, Susan. "Learning Objects: A Rose by Any Other Name." *EDUCAUSE Review* 40, no. 4 (May/June 2005). http://www.educause.edu/apps/er/erm05/erm05410.asp?bhcp=1.

MIT Libraries. "MIT Libraries' DSpace Business Plan Project. Final Report to the Andrew W. Mellon Foundation." July 2002. http://libraries.mit.edu/DSpace-mit/mit/mellon.pdf.

—. "Final Report on the Initial Development of the DSpace Federation." June 2004. http://www.DSpace.org/federation/mellon-DSpace.pdf.

Ober, John. "IR Content and Service Expansion: the Case of eScholarship." Presented at the SPARC/SPARC Europe Workshop on Institutional Repositories: The Next Stage. November 18–19, 2004. Washington, D.C.,

http://www.arl.org/sparc/meetings/ir04/presentations/ober.html.

OCLC. "Research and Learning Landscape: Institutional Repositories, Scholarly Communication and Open Access," The 2003 OCLC Environmental Scan: Pattern Recognition: A Report to the OCLC Membership. Dublin, OH: 2004. http://www.oclc.org/membership/escan/research/repositories.htm.

OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS). Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community. Dublin, Ohio: OCLC Online Computer Library Center, Inc., September 2004. http://www.oclc.org/research/projects/pmwg/surveyreport.pdf

Open Society Institute. A Guide to Institutional Repository Software v 3.0. New York: Open Society Institute, 2004. http://www.soros.org/openaccess/software/.

Ottaviani, Jim, and Cory Snavely. Towards a University of Michigan Institutional Repository: A Study and Prototype Proposal. Ann Arbor, MI: University of Michigan, November 12, 2003. http://www.umdl.umich.edu/pubs/inst-repos20031112.html.

Parod, Bill. "FEDORA Project," Presentation at A Symposium on the Crisis in Scholarly Communication: "For the Greater Good: Open Access and Institutional Repositories," McGill Libraries, May 17, 2004 http://www.library.mcgill.ca/symposium/parod.ppt.

Pritchard, Sarah M., Larry Carver, and Smiti Anand. Collaboration for Knowledge Management and Campus Informatics. Based on a report to the Andrew W. Mellon Foundation. Santa Barbara, CA: University of California, Santa Barbara, August 23, 2004.

http://www.library.ucsb.edu/informatics/informatics/documents/UCSB Campus Informa tics_Project_Report.pdf.

—. Research Information Management: Common Concerns of UCSB Faculty, Librarians, and Technologists. Draft version. Based on a study funded by the Andrew W. Mellon Foundation. Santa Barbara, CA: University of California, Santa Barbara, July 22, 2004.

http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Informatics_Whi te_paper_July04.doc.

Prudlo, Marion. "E-Archiving: An Overview of Some Repository Management Software Tools." Ariadne 43 (April 2005). http://www.ariadne.ac.uk/issue43/prudlo/intro.html.

Research Libraries Group. *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report*. Research Libraries Group: Mountain View, CA: May 2002. <u>http://www.rlg.org/longterm/repositories.pdf</u>.

Simpson, Duncan. *Digital Preservation in the Regions. Sample Survey of Digital Preservation Preparedness and Needs of Organisations at Local and Regional Levels.*London: Museums, Libraries and Archives Council, 2005. <u>http://www.mla.gov.uk/documents/mla_dpc_survey.pdf</u>.

Suber, Peter. "Open Access Overview: Focusing on Open Access to Peer-Reviewed Research Articles and their Preprints." *Open Access News*, a blog by Peter Suber. <u>http://www.earlham.edu/~peters/fos/overview.htm</u> (accessed June 22, 2005).

Swan, Alma, and Sheridan Brown. *Open Access Self-archiving: An Author Study*. Study commissioned by JISC Committee for the Information Environment (JCIE). Cornwall, UK: Key Perspectives Limited,, May 2005.

http://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archiving-an%20author%20study.pdf.

Tennant, Roy. "Institutional Repositories." *Library Journal* 127, no. 15 (September 15, 2002): 28, 30.

United States. National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Pre-publication draft. Approved by the National Science Board May 26, 2005, subject to final editorial changes. Arlington, VA: National Science Foundation, 2005.

http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf.

Van de Sompel, Herbert. "Challenges Ahead." Keynote address, Olybris Ex Libris Seminar, Kos, Greece, April 18, 2005. http://public.lanl.gov/herbertv/presentations/olybris_keynote_2005.pdf.

Van de Sompel, Herbert, Jeroen Bekaert, Xiaoming Liu, Luda Balakireva, and Thorsten Schwander. "aDORe: A Modular, Standards-based Digital Object Repository". Draft of submission to *The Computer Journal*. Los Alamos, NM: Los Alamos National Laboratory, Research Library, 2005.

http://public.lanl.gov/herbertv/papers/aDORe_20050128_submission.pdf.

Van de Sompel, Herbert, Sandy Payette, John Erickson, Carl Lagoze, and Simeon Warner. "Rethinking Scholarly Communication: Building the System that Scholars Deserve." *D-Lib Magazine* 10, no.9. (September 2004). http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html.

Wheatley, Paul. *Institutional Repositories in the Context of Preservation*. DPC Technology Watch Series Report 04-02. York, England: Digital Preservation Coalition, March 2004. <u>http://www.dpconline.org/docs/DPCTWf4word.pdf</u>.

Zandonella, Catherine. "UC to Launch Open Access Journals." *The Scientist*, (June 16, 2003. <u>http://www.biomedcentral.com/news/20030616/03</u>.

Appendix 3: Outline of themes covered in digital repository examinations⁶⁸

- 1. Background information
 - Mission statement, overview of services, funding, operational status, governance structure and administrative policies
- 2. Participation
 - Who deposits, what types of materials and formats, agreements for obtaining materials, contracts, permissions, etc.
- 3. Participation feedback and repository evaluation
 - o Assessments of client satisfaction, quality assessment, ongoing evaluation
- 4. Access
 - Open access, restrictions, authentication, finding content (search, browse), transfer of content, etc.
 - o Connections between other resources, learning management systems, etc.
- 5. Relationship between preservation and access copies
- 6. Preservation strategies
 - Format specific, bit-level, normalization, migration, emulation, redundancy, etc.
- 7. Software
 - Open source, commercial, locally developed, layered software applications
- 8. Metadata
 - OAIS categories, how obtained, requirements, automated metadata production, metadata storage, relationships among objects, event metadata, etc.
- 9. Copyright and intellectual property
 - Content providers, repository, institution
- 10. Documentation
 - o Policies, metadata conventions, technical profiles, etc.

⁶⁸ Based upon a questionnaire from: OCLC/RLG PREMIS Working Group. 2004.

[&]quot;Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community." Report by the joint OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS). Dublin, O.: OCLC Online Computer Library Center, Inc. <u>http://www.oclc.org/research/projects/pmwg/surveyreport.pdf</u>