Draft Paper for *Encyclopedia of Library and Information Science*
Joseph Branin, the Ohio State University Libraries
Summer 2003

INSTITUTIONAL REPOSITORIES

KEYWORDS

Institutional Repositories, Disciplinary Repositories, Digital Repositories, Digital Archives, Digital Libraries, Open Archives, Digital Asset Management, Knowledge Management, Digital Preservation, Digital Stewardship, Metadata Standards, OAIS Reference Model.

INTRODUCTION AND DEFINITION

In a general sense, an "institutional repository" can mean many things. A library, an archive, a museum, or even a warehouse that stores for use and safekeeping an organization's records or artifacts falls under the broad definition of an institutional repository. In recent years, however, an institutional repository has taken on a more specific, new, but still evolving, meaning that refers to the storage and preservation of an organization's digital information or knowledge assets. As more and more information is created in digital formats at the individual, desktop level, institutions are beginning to turn their attention to how to identify and manage for long-term use, and for the common good of the organization or a larger public, important digital assets. In 2002, the Massachusetts Institute of Technology (MIT) in collaboration with the Hewlett-Packard Corporation launched Dspace,[1] which is a highly publicized, open source, institutional repository system, and The Scholarly Publishing and Academic Resources Coalition (SPARC) issued "The Case for Institutional Repositories: A SPARC Position Paper."[2] These events put institutional repositories in the spotlight as an interesting new development in librarianship and information management.

Although institutional repositories are still evolving and taking on differing manifestations in specific institutions, they can be defined in general as systems and service models designed to collect, organize, store, share, and preserve an institution's digital information or knowledge assets worthy of such investment. This may, of course, sound very much like a library, and in many cases an institution's library should and is taking responsibility for developing and operating such a digital repository. But while the mission of an institutional repository coincides nicely with that of a library, the technical infrastructure and the types of material collected in such a repository present new challenges and extended responsibilities for the traditional library.

GENESIS OF INSTITUTIONAL REPOSITORIES

Repositories, in their broadest sense, have existed ever since humans began collecting and storing important information and artifacts for safekeeping and long-term use. The long and rich history of libraries, museums, and archives provides the foundation for any type of repository program, but two contemporary developments in particular have helped shape the nature of today's institutional repositories: the emerging knowledge management movement; and the maturing, but still rapidly advancing, technology of content or asset management in the digital information system.

Knowledge Management Movement

In 1988, Peter Drucker published an influential paper entitled "The Coming of the New Organization" in the *Harvard Business Review*.[3] In this paper Drucker argued that for the modern organization, knowledge had become its most important asset, and that those organizations best able to manage and exploit their corporate knowledge assets would be the most successful in the marketplace. Drucker did not explain in any detail what he meant by "knowledge," but he was clear that knowledge manifested itself in many forms in the organization, ranging from patents and trade secrets, to operational routines, to the expertise inside the heads of employees. Subsequent books, articles, and conferences by researchers in the multi-disciplinary fields of information science and business management have elaborated on Drucker's ideas about the importance of knowledge management in the organization. These researchers have pursued the meaning and implications of knowledge as it is differentiated from data and information, and they have developed models, conceptual structures, and best practices for managing knowledge in the modern organization.[4]

Repositories, and that is the name commonly used in the knowledge management field, play an important, but supporting, role in a knowledge management system. Davenport and Prusak, for example, in their 1998 book *Working Knowledge: How Organizations Manage What They Know*, list "knowledge repositories" first in their review of knowledge management projects in practice. They claim to have "come across three basic types of knowledge repositories:

1. *External knowledge* repositories (example: competitive intelligence)
2. *Structured internal knowledge* repositories (example: research reports, production oriented marketing materials and methods)
3. *Informal internal knowledg*e repositories (example: discussion databases full of know-how, sometimes referred to as 'lessons learned')."[5]

While a definite part of a knowledge management system, repositories play a supporting role, for as Davenport and Prusak point out, they tend to treat knowledge as an "it," while in reality knowledge is "a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In

organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms."[6]

The knowledge management movement of the 1990's influenced the development of institutional repositories in a number of significant ways in addition to establishing the nomenclature. Management consultants and senior administrators endorsed the movement's emphasis on the competitive value to an organization of paying attention to its knowledge assets. And the movement's broad view of knowledge as diverse and dynamic made the identification, capture, and management of knowledge assets much more complex and challenging. Books, articles, or any types of published documents were viewed as only one obvious manifestation in a wide range of explicit and tacit knowledge assets that needed to be managed in an organization.

Maturing Digital Asset Management Technology

By the year 2000, several broad technological developments in the digital information system were also pushing individuals and their organizations towards the creation of institutional repositories. First, it was becoming easier for individuals or small groups to create and to disseminate digital assets through the use of microcomputer desktop tools and computer networking. At a university, for instance, it would not be uncommon by 2000 to find faculty members or small disciplinary centers around campus creating digital text documents, digital multimedia, web sites, or online courses. While the highly decentralized and distributed nature of microcomputer desktop publishing was empowering to the individual, it presented some managerial or stewardship challenges for organizations interested in coordinating, sharing, and preserving its units' or employees' digital assets. Clifford Lynch, for example, in his 2003 report, *Institutional Repositories:*

NASA's Consultative Committee for Space Data Systems, which offers "a comprehensive logical model describing all the functions required in a digital repository."[8]

- The Open Archives Initiative from the library and scientific community, which has developed an Open Archives Metadata Harvesting Protocol (OAI-PMH) that defines a mechanism for harvesting XML-formatted metadata from repositories.[9]
- A Metadata Encoding and Transmission Standard (METS) developed under the sponsorship of the Digital Library Federation, which provides a schema for encoding descriptive, administrative, and structured metadata in a digital repository or library.[10]
- Shareable Courseware Object Reference Model (SCRORM) developed by the federal government agency Advanced Distributed Learning to provide guidance for the preparation and storage of digital educational material so that such material is "reusable, accessible, interoperable, and durable."[11]
- Publishing Requirements for Industry Standard Metadata (PRISM), a schema under development by the publishing industry to create a common language for the metadata that describes published digital assets.[12]
- Open source and proprietary software systems such as Dspace, ePrints, FEDORA, bepress, Documentum, CONTENTdm, IBM's Content Management, and Artesia's TEAMS that offer technical infrastructure options for implementing all or part of an institutional repository.[13]

Institutional Stewardship and the Culture of Sharing

While the proliferation of digital assets and the maturing of digital asset management systems were pushing towards and making possible the creation of digital repositories, the jurisdictional boundaries for such repositories remain open to a range of organizational options. A digital repository in fact can be owned and managed by an individual, a small group, an institution or commercial organization, a consortium of organizations, or a government entity. A digital repository can be defined by its jurisdictional scope but also by the type of assets it collects. Disciplinary repositories focus on the collection of digital assets in a subject area, for example, arXiv, which collects e-prints from around the world in the fields of physics, mathematics, non-linear science, and computer science.[14] Institutional repositories organize themselves along organizational or political jurisdictional lines, and they collect and manage digital assets in a variety of formats and subjects for the constituents within that jurisdiction. MIT's Dspace is an example of an institutional repository, for its mission is to collect in digital form the intellectual output of the MIT faculty.

Does an institution provide a logical or effective framework for a digital repository? Someone like Peter Drucker, of course, would see competitive value in a commercial organization's careful control over the sharing and restricting of its knowledge assets. Employees will be more productive if they can identify, share, and pass on expertise within an organization; and, in turn, the commercial organization through centralization

or careful coordination can better control the flow of its knowledge to outside competitors and consumers. In the non-profit and public sectors, however, ownership and control of information within the organization are usually more diffuse. In higher education in the United States, for example, certain types of technical research and knowledge are tightly controlled by the institution through patent, trademark, and invention policies; while creative work, which usually takes the form of books, articles, or course material, remains largely unregulated, at least for the time being.[15] If students and faculty in an academic setting are "free agents" rather than "workers for hire" when it comes to the ownership and control of their creative work, will they deposit their digital assets in an institutional repository?

Academic administrators and librarians appear motivated to create institutional repositories for two primary reasons: improving access to and preservation of unpublished digital assets, and reforming the scholarly publishing system. The Council on Library and Information Resources 2003 report *New-Model Scholarship: How Will It Survive* by Abby Smith is an excellent summation of "the growing problem of digital stewardship" in the academy.[16] Smith describes and characterizes the growth of new digital scholarship in the academy but worries about its ephemeral nature. Will these new, digital models of scholarship – "digital objects that are created outside the library and seldom developed expressly for publication"[17]– be available for long-term use? Smith finds a small but growing number of universities, academic disciplines, publishers, and government or non-profit agencies trying to address this preservation and access challenge through the deployment of digital repositories. Raym Crow, the author of the SPARC paper on *The Case for Institutional Repositories*, believes the rational for creating intuitional repositories centers on an interest in a "new scholarly publishing paradigm" and on "institutional visibility and prestige."[18] While Smith identifies preservation and access as basic, altruistic reasons for creating digital repositories, Crow emphasizes the more self-interested motivations of a creator or an institution gaining control of its digital content assets for reasons of economics and prestige. As librarians find it increasingly difficult to buy back the output of faculty who have given it to publishers, why not create new, less expensive venues for sharing scholarship through institutional repositories and open access? Academic institutions and their faculty are the well springs of much new knowledge, and these same institutions can do more to collect, share, and preserve this knowledge at its very source through institutional repositories.

In the end, institutional repositories in the academic setting are likely to succeed only if there are shared values and motivations among administrators, faculty, and students. Certainly, many creators of digital assets would appreciate an institution's assistance with the long-term management of these assets if they could trust the institution to protect and honor their property and privacy rights. As a broad social goal, faculty and academic institutions both value the creation and dissemination of new knowledge and the preservation of this knowledge for future generations. If an institutional repository can be shown to advance these values, it might become the common ground for the safekeeping and sharing of digital knowledge assets. But it will be a challenge. David Blair in his review article, "Knowledge Management: Hype, Hope, or Help?" makes an important observation about the critical role that a "culture of sharing" plays in a successful

knowledge management program. According to Blair, "the kind of culture of an organization can be an important factor influencing whether it encourages the sharing of knowledge between employees or inhibits it. Organizations that have a culture of being very competitive internally, with employees competing against each other for customer and projects, will have a great deal of trouble convincing expert employees to pass their expertise/knowledge on to those who are less knowledgeable."[19] The culture of academic institutions is complex, with cooperative as well as competitive elements. Faculty and students will have, and will choose from, a range of options in managing their digital assets: keeping them close to home on individual computers or departmental servers, depositing them in institutional repositories, or turning to disciplinary repositories or publishers in their fields of interest.

CONCEPTIONAL MODEL AND STANDARDS FOR A DIGITAL REPOSITORY

The Open Archival Information System (OAIS) Reference Model

For an overview of the requirements, components, and functions of a digital repository, it is useful to review the *Reference Model for an Open Archival Information System (OAIS),* which provides a "conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long run."[20] The OAIS reference model, which is a 148-page document published in January of 2002, became an ISO standard in 2003 (ISO 14721: 2003).[21] The model can really apply to any type of archive -- physical or digital, institutional or disciplinary – and delineates the general environment, the characteristics of information, and the basic functional arrangements and responsibilities of an archive or repository.

At its most abstract level as seen in figure 1, the environment around an archive or repository has three basic players: producer, management, and consumer.

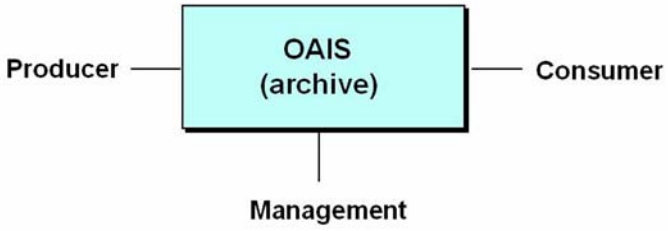

Figure 1: OAIS Environment

The assets or "information objects" that go into an archive or repository can be deconstructed into different types of information, such as content information or metadata

information, which can take the form of preservation description information, packaging information, and descriptive information, as illustrated in figure 2.

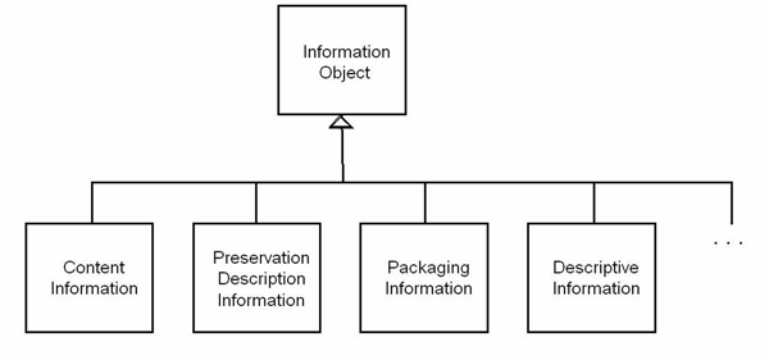

Figure 2: Information Object in OAIS Model

An archive or repository has six basic functional activities or responsibilities, as illustrated in figure 3: ingest, archival storage, data management, access, administration, and preservation planning.

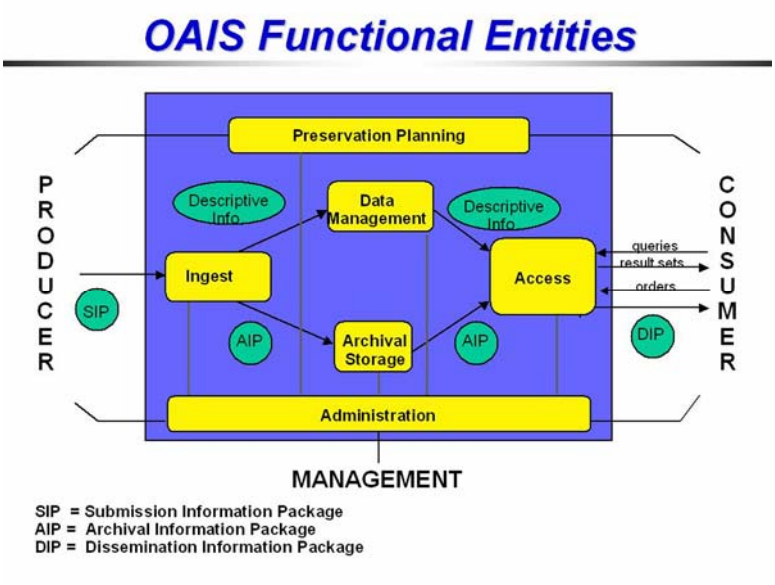

Figure 3: OAIS Functional Model

In the OAIS functional model, the producer prepares a submission information package (SIP), which has content and metadata information. The repository ingests the SIP and generates an archival information package (AIP), which complies with the archive's data formatting and documentation standards, and extracts descriptive information from the AIP for inclusion in the Data Management function. Archival Storage provides services and functions for the storage, maintenance and retrieval of AIPs; while Data Management maintains descriptive information that identifies archive holdings and administrative data used to manage the repository. Access is the function that allows consumers to learn what is in the repository and request and receive a Dissemination Information Package (DIP)

from the repository. Administration and Preservation Planning are high-level responsibilities of the management of the repository, ensuring overall operation of the repository system and ongoing preservation of content accessibility for the consumer even if the original information format or computing environment becomes obsolete.

The need to manage digital assets for long-term use is one of the driving forces behind the establishment of institutional repositories, and the OAIS reference model provides a strategy for accomplishing this challenging responsibility in the highly changeable computer technology environment. The Preservation Planning function in the OAIS reference model is based on information structures and on digital migration strategies. First and foremost, metadata information about a digital asset that is essential for preservation must be captured and stored in a standard manner. The Archival Information Package (AIP) in the OAIS reference model contains both content information and preservation description information, as illustrated in figure 4.

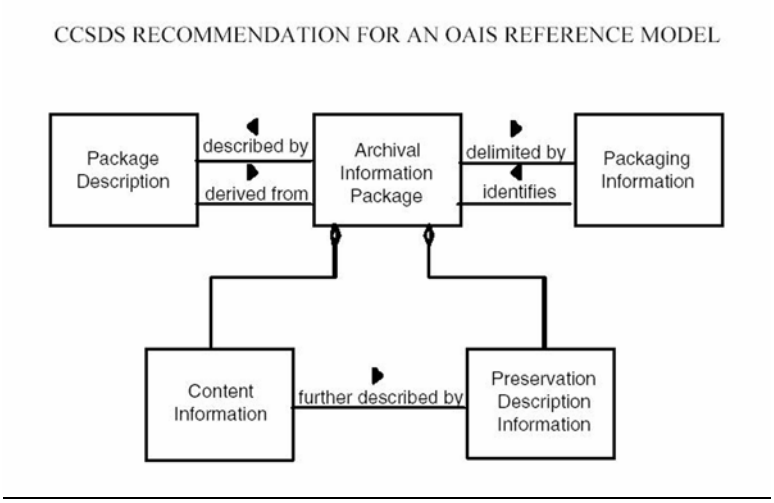

Figure 4: OAIS Archival Information Package

Preservation Description Information (PDI) can be broken down into its components of reference information, provenance information, context information, and fixity information, as illustrated in figure 5.

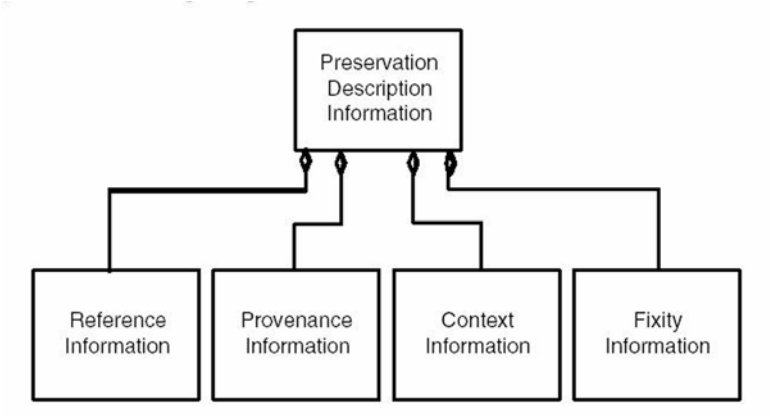

Figure 5: OAIS Preservation Description Information

With this preservation information available to the management of a repository, a digital technology migration strategy can be employed when factors such as media decay or software evolution dictate a need for change. Digital migration can take the form of refreshment, replication, repackaging, or transformation, all of which are defined in the OAIS reference model. While the OAIS reference model structure and strategy for digital preservation provides a reassuring guideline or standard, it will take fortitude and resources from the management of the repository to carry out this preservation responsibility over time.

The Open Access Movement and the Open Archive Initiative

The word "open" has become quite popular in the computing and academic communities. There is the Open Archival Information System (OAIS) Reference Model discussed above; there are the Open Access Movement and Open Access Initiative (OAI) discussed below; and there are several important other "opens" such as the Open Knowledge Initiative (OKI)[22] in course management software and the Open Source Initiative (OSI)[23] in computer code and programming. All these "opens" can be confusing, for they are not all the same thing in purpose or scope. Some are broad, general movements or social positions, while others are more specific technical standards or models. However, they all use "open" to connote an approach that fosters the free exchange of information, whether that be in the development of software codes or standards, or in the actual sharing of content in digital repositories. The culture of the Internet and the World Wide Web, where one finds common, public protocols underlying a system that provides mostly free access to digital information and services (for example, the Internet Archive's motto is "universal access to all human knowledge"[24]), has an "open" philosophical base to it, and this "open" approach is clearing influencing the development of institutional repositories

in their software systems as well as in their approaches to sharing metadata and content information.

The general Open Access movement and the more specific technical protocol activity called The Open Archive Initiative (OAI) – not to be confused with OAIS, which is a reference model for the design of a digital repository -- grew out of the scientific and library communities' experience with disciplinary e-print archives or repositories.[25] The most prominent example of such a disciplinary e-print repository is arXiv, which was begun in 1991 by Paul Ginsparg at the Los Alamos National Laboratory and is now hosted at Cornell University.[26] Today arXiv, as a disciplinary, open, digital repository, manages 230,000 papers with abstracts and citations in fields of physics, mathematics, non-linear science, and computer science. The success of arXiv and a few other disciplinary e-print archives as an innovative collection and dissemination approach to digital scholarly communications has generated a growing international interest in the "open access" model, where scholarly publications housed in repositories are made freely available to the public over the Internet. Both the Budapest Open Access Initiative[27] and the Public Library of Science[28] are efforts by scientists to make their literature more openly accessible to the public through the use of national, institutional, or disciplinary repositories that share their content freely over the Internet.

The Open Archives Initiative (OAI) is an information and library science group supported by the Digital Library Federation, the Coalition for Networked Information, and the National Science Foundation to develop and promote interoperability standards that "facilitate the efficient dissemination of content" in digital repositories. The group, led by Carl Lagoze from Cornell University and Herbert Van de Sompel from the Los Alamos National Laboratory, has formulated and is promoting an Open Archives Metadata Harvesting Protocol (OAI-PMH) that provides guidance on a common format for metadata in digital repositories.[29] The Protocol guides repository data providers on how to structure and format their metadata in a manner that allows service providers to harvest the metadata for centralized search and discovery services. The OAI-PMH protocol facilitates open, union access to at least the descriptive metadata in multiple repositories, exposing pointers to the content in these repositories. Depending on the rights management policies and procedures of the individual repositories, content can then be shared or restricted.

Steven Hitchcock from Southampton University has prepared a report entitled *Metalist of Open Access E-Print Archives: the Genesis of Institutional Archives and Independent Services*.[30] Hitchcock's primary interested in this report is to identify and describe open access archives or repositories, particularly those that contain digital full-text papers that have been self-archived by their authors, but he admits that it is difficult to quantify the growth and number of digital archives or repositories even of this type. There are growing numbers of institutional archives or repositories as well as subject-based, disciplinary archives. Material in open access disciplinary repositories tend to be pre-prints or e-prints of scholarly articles and technical reports, while institutional repositories are more eclectic with assets drawn from the institution's diverse teaching and research output. Hitchcock questions how likely it will be for the public to search an

individual institutional repository's holdings, and he believes the harvesting and central search services that the OAI protocol can enable will more effectively expose the content of institutional repositories. In describing institutional archives or repositories, Hitchcock refers to lists of institutional repositories maintained by SPARC and by Signal Hill, and he points to the University of California's eScholarship Repository, Caltech's Collection of Open Digital Archives (CODA), and the U.S. Department of Energy's Information Bridge as representative examples of institutional repositories.

BASIC COMPONENTS OF AN INSTITUTIONAL REPOSITORY

Technology Platform

The Open Archival Information System (OAIS) Reference Model and the Open Archives Metadata Harvesting Protocol (OAI-PMH) discussed above, along with the Metadata Encoding and Transmission Standard (METS) and other metadata schema, basically offer abstract models, guidance, and standards on the technical architectural issues in building a digital assets management system and in insuring its interoperability with other systems. In practice, there are now a number of digital asset management (DAM) systems on the market and in operation in corporations, government agencies, and higher education institutions. These digital asset management systems fall into two basic categories: commercial, turnkey systems such as those offered by Documentum, Artesia, IBM, and the Berkeley Electronic Press; and non-profit, open-source systems such as Dspace developed by MIT and Hewlett-Packard, ePrints from the University of Southampton, and FEDORA now under development by the University of Virginia and Cornell University.

The digital asset management (DAM) systems industry is still in its infancy, but there are optimistic forecasts for its growth – according to one forecast, the industry will grow from a $200 million market level in 2003 to a $3.5 billion level by 2009 -- as private sector organizations such as broadcast agencies, multimedia publishers, and advertising houses as well as public sector institutions in government, health care, and education embrace DAM systems.[31] A growing number of higher education institutions and agencies are adopting, or at least experimenting with, digital asset management systems that can provide a technology platform for their institutional repository needs. The University of California system is using bepress from the Berkeley Electronic Press for its eScholarhip repository program. The California Institute of Technology and more than fifty other educational agencies are employing ePrints software in their digital repository programs. Washington State University is using CONTENTdm to manage its growing collection of digital maps, historical photographs, texts, and videos. Stanford University is deploying Artesia Technologies' TEAMS system in its repository program; while OhioLINK, the statewide consortium of academic libraries in Ohio, has built its Digital Media Center repository for multimedia digital assets on the Documentum platform. MIT in partnership with Hewlett-Packard has developed and released its open-source,

institutional repository system called Dspace, which stands for "Durable Digital Depository," and is now testing a federated implementation of this platform with six other universities: Columbia, Cornell, Ohio State, and the Universities of Rochester, Toronto, and Washington. The University of Michigan is currently reviewing the field of DAM systems options for an enterprise-wide solution and has prepared a request for proposals from DAM system vendors.[32]

In this expanding and competitive market for digital asset management systems, institutions will have to choose among a variety of proprietary and open-source options. Many of these technical platform options are still in development, and all of them will have to continually expand and upgrade their capabilities to remain competitive. As one might expect, each digital asset management system (just as in the integrated library system platform arena) has its advocates and critics, its strengths and shortcomings in handling various types of digital content, and its range of costs. While open source options such as Dspace and ePrints may have free or low entry prices, ongoing local development and support requirements for this type of system must be factored in when comparing longer-term costs to those of commercial, turnkey system options. Interoperability and extensibility -- a system's ability to interface with other systems as well as its ability to expand and add new components -- are also important factors in evaluating the merits of a digital asset management system. The OAIS reference model and interoperability standards and protocols provide helpful guides to evaluating the functionality, interoperability, and extensibility strengths and weaknesses of various DAM systems.

Whether or not institutions adopt MIT and Hewlett-Packard's Dspace for their digital asset management system software, this private university and computer company deserve credit for openly sharing the development of their institutional repository program. Not only is the software of this system open source and freely downloadable, but the policies, procedures, and business plans for MIT's institutional repository are all carefully documented and open to public scrutiny at the Dspace web site. Anyone interested in institutional repositories can gain insight into all aspects of such a program by studying the Dspace documentation and literature. Its technical architecture, for example, which is built on three basic layers of application, business logic, and storage, as illustrated in figure 6, is explained in open detail at their web site and in published articles.[33]

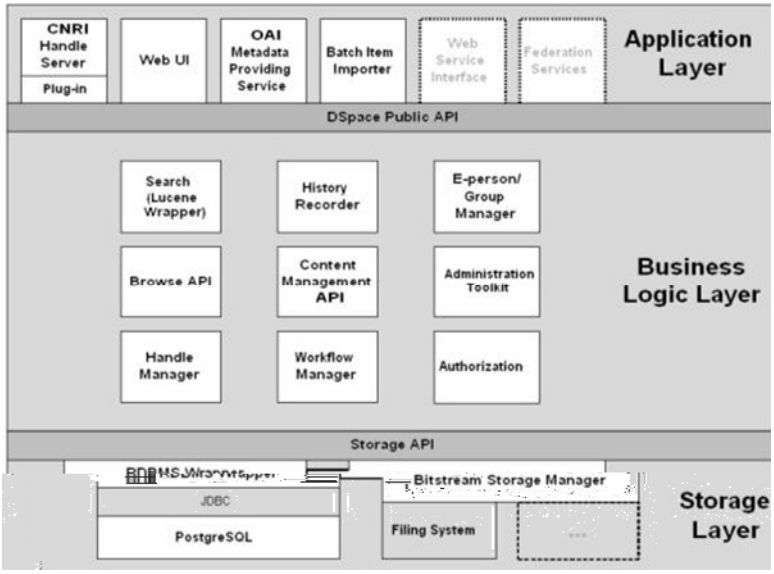

Figure 6: DSpace Technical Architecture

Service Model

The technology platform is, of course, an essential component – and its capability a driving force in the establishment -- of an institutional repository, but it may prove over time the least expensive and least complicated component. As DAM system technology matures and as digital storage cost decline, service activities and organization that surround and support an institutional repository may turn out over the long-term to be the more expensive and challenging aspects of such a program. A service model for an institutional repository will have to include some or all of the following activities:

- assistance with digital asset creation and submission,
- metadata preparation, or training and guidance in metadata preparation,
- intellectual property rights management,
- preservation management,
- assistance with content access and use,
- marketing.

An institutional repository should be an integral part of a larger knowledge management or information services program of an institution. The repository itself simply stores and provides capabilities for preserving and sharing digital assets. To be successful, individuals in the institution must understand the purpose and benefits of the repository, willingly submit digital assets to the repository, and finally, make full use of the assets in the repository in their work. Using the terminology of the OAIS reference model, management must make the institutional repository program understandable to producers and consumers through marketing and training, and easy to use through efficient, streamlined, and highly automated or personalized services.

At the Ohio State University, for example, the Knowledge Bank project places its institutional repository in the larger context of a multifaceted knowledge management program.[34] The university library's traditional focus on collecting, storing, and preserving published scholarly material is related and extended to new responsibilities for handling unpublished digital assets such as working papers, research databases, and multimedia course material. Administrative and academic computing's responsibilities for data warehousing, teaching technology, and course management systems also are related to the institutional repository through the Knowledge Bank project. And other knowledge management activities such as the development of expertise directories and information policies for rights and privacy are viewed as related parts of an overall knowledge management program. Figure 7, from the Ohio State University's Knowledge Bank project, provides a listing of the components of this knowledge management program and establishes a context for an institutional repository as an integral – not isolated --function within an institution's information service environment.

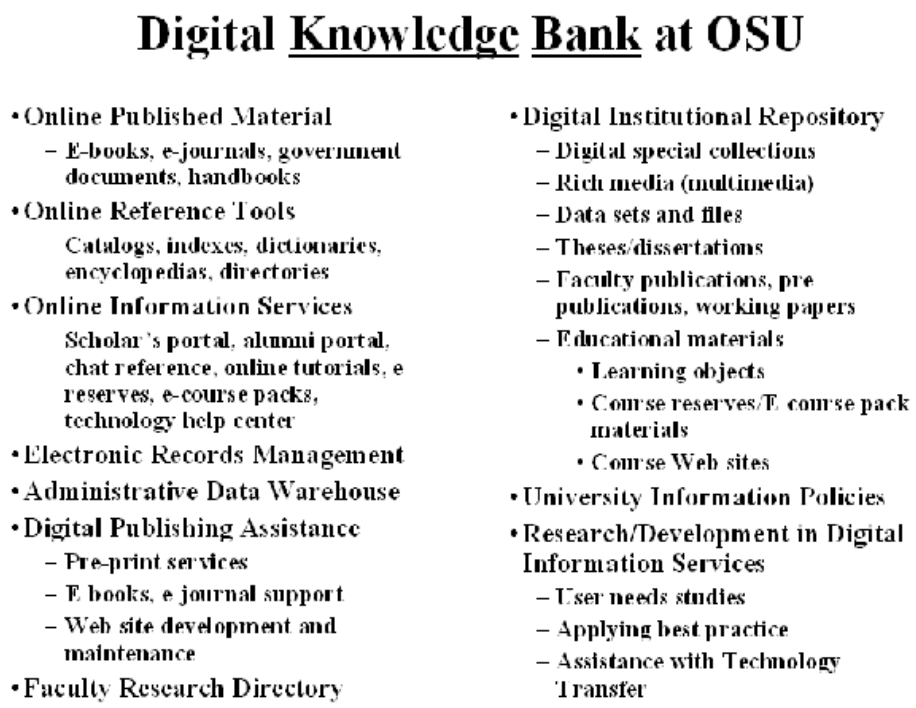

Figure 7: OSU Knowledge Bank Project Components

In designing the service model for an institutional repository, management (again using the terminology of the OAIS reference model) will have to decide how much assistance it wants and can afford to provide producers and consumers. Up-front services to producers

might include multimedia production and design assistance, digitization, and metadata training and preparation. Or management could decide that self-archiving, the use of templates, and automated services that place preparation and submission responsibilities almost entirely in the hands of the producer are more scaleable and economical ways to design front-end services. Metadata preparation and conformance to metadata standards play a critical role in a digital repository, for they underpin the search and discovery, interoperability, as well as the preservation capabilities of a digital repository. Whether the producer or management takes responsibility for metadata preparation and compliance, this front-end service function demands careful attention in any institutional repository program.

On the back-end, management will have to decide how much access availability and service assistance it will provide consumers of an institutional repository. What search and discovery tools will management provide for consumers and will there be training in the use of these tools? Will metadata and content in the repository be exposed and accessible to general or specialized search services? Will there be reference assistance to help consumers make effective use of the knowledge assets in the institutional repository? Critical to access services for the consumer are the preservation and rights management responsibilities of management in an institutional repository program. Management must carry out proper backup, disaster preparedness, and timely migration strategies in the institutional repository in order to ensure long-term access to assets by consumers. And finally, consumers should have access only to assets they are authorized to see and use. In a full service institutional repository, rights management will extend the gamut of access control, from private assets only available to the producer, to assets restricted to certain groups or types of consumers, to complete open access to the public domain.

In establishing an institutional repository in an academic setting, the greatest initial service challenge is likely to be getting faculty or student engagement and participation. In a corporate setting, management may require that all employees deposit their appropriate digital assets in the organization's repository, but in an academic setting where power and control are diffuse and where faculty and student are treated as free agents at least when it comes to creative work, convincing producers to contribute their work to an academic institutional repository becomes a significant marketing challenge. One approach to this marketing challenge – one used in the Ohio State University Knowledge Bank project -- is for management to conduct an inventory of current digital information projects on campus and then to discuss with identified producers their needs and the capabilities of the repository program to store, preserve, and share their assets. Such identification of early adopters of digital technology, assessing their needs, and involving them in the design of repository services and policies can only strengthen an institutional repository's viability. Another service and marketing strategy to reach producers and consumers is to identify, design services, and market to "communities of practice," a concept taken from the knowledge management field. Most knowledge work is done neither in isolation nor in big, impersonal institutional settings but rather in smaller, more focused, communities of practice where producers and consumers share some common interest in a subject or activity. In both the eScholarhip project in the

University of California system and in Dspace at MIT, communities of practice –that is, academic departments or disciplinary centers – are providing an organizing focus and structure for their institutional repository marketing, service, and policy development activities.

Collection and Information Policy

What goes into an institutional repository, and can it come out once it goes in? Will the repository accept all kinds and manner of digital assets created by members of an institution, or will there be selectivity based on a collection policy that guides digital format, subject, and content submission? In a traditional library, of course, selectivity is a grounding principle. Only materials with predicted immediate demand or with some degree of lasting value to the members of the institution are selected for acquisition and the ensuing costly tasks of cataloging and long-term inventory maintenance and preservation. Except for their special collections, libraries, by and large, deal with published material in a limited range of formats. As a result, libraries have been collecting an important segment but not a full range of knowledge assets for their organizations. If one looks at the universe of library holdings represented in the OCLC WordCat database, for example, and characterizes these holding by broad type, as Lorcan Dempsey, the Vice President for Research at OCLC, has done in the data map of metadata in Figure 8, it becomes clear that new, unpublished, digital resources have fallen outside the scope of most traditional library programs.

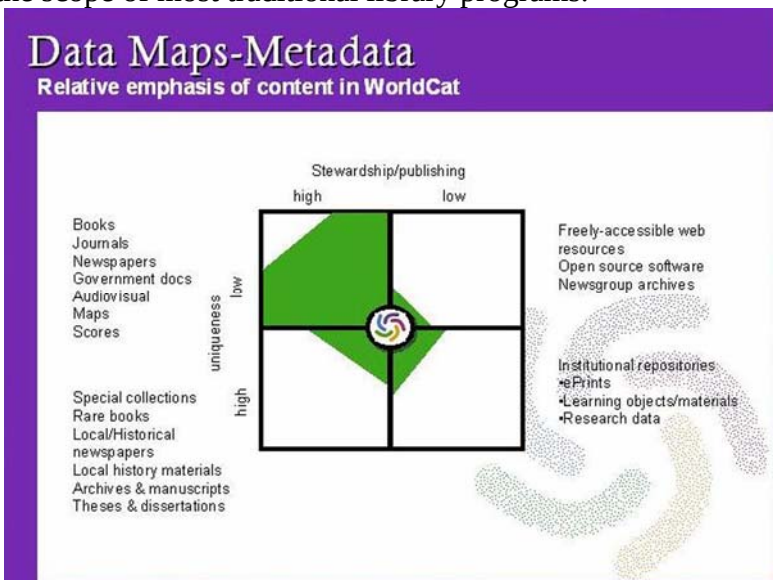

Figure 8: OCLC Data Maps -- Metadata

This traditional collection focus and selectivity are not necessarily bad, for they have allowed librarians to concentrate their resources on what might be the most valuable or vetted and standard forms of explicit knowledge assets. The expanded collection scope of

an institutional repository that includes unpublished digital assets in many new formats will require more resources as well as more economical, less labor-intensive procedures for cataloging, storing, and preserving digital assets. But even with an expanded collection scope, an institutional repository will still need some degree of selectivity. Choices will still have to be made -- and who will make these choices? – about which particular digital assets should be archived in a repository. Are they worth the cost of stewardship or are they of such an ephemeral nature they do not warrant deposit in an institutional repository? DAM system technologies are being designed to handle an increasing variety of digital formats, from text, to images, to sound, to moving images, to multimedia. However, there are still limitations in many of these systems as to the range of digital formats they can handle or handle well, and there is still a lack of established standards for every new digital format that makes metadata preparation and preservation planning difficult. The technology platform capabilities and established metadata standards, therefore, may dictate some of the format selection criteria in an institutional repository.

Intellectual property rights policy is likely to loom large in many institutional repository programs. In the digital information environment, the more powerful capabilities to copy, reuse, repurpose, as well as restrict digital assets are intensifying intellectual property rights issues. Most universities have established intellection property policies for technical information, but many have no such explicit policies for creative works, such as scholarly e-prints or courseware. In general, academic institutions have a tradition of not exerting ownership rights over the creative works of faculty or students unless that work was done with the support of “significant “or “substantial” institutional resources. As an institutional repository begins to collect and store a faculty member’s unpublished working papers or course material or a student’s e-portfolio, all the parties involved need to know the institution’s ownership policies and the repository’s rights management protections for this material. Several universities, such as Cornell and Brigham Young University, do have carefully prepared and comprehensive intellectual property and copyright polices, and these can serve as models for other institutions.[35] At MIT in the Dspace program, contributors to their institutional repository complete a non-exclusive distribution license that state the rights and responsibilities for intellectual property for both the producer and management.

Administration and Cost

The management and administration of an institutional repository could be taken up by a variety of entities in an organization: in most cases responsibility will likely fall to an information technology (IT) unit, to a library, or to a combination of these units. The traditional mission of a library to collect, preserve, and share books, journals, and other published materials for the common good of its institution’s members could be extended to cover the same responsibilities for a wider variety of unpublished digital assets. However, this extension of collecting, cataloging, servicing, and preservation responsibilities can appear daunting in the face of new and unclear boundaries,

technologies, and additional costs. Librarians do know the cost and their own limitations in managing the universe of published information. How can they afford to take on yet a larger universe of published as well as unpublished knowledge assets? Early adopters of institutional repository programs obviously believe in their value, and they are finding ways to redirect activities, create new partnerships, and invest new resources to make them happen.

It is not possible to describe the exact costs of an institutional repository, because they are new and because they can vary so much in size and scope. At the high cost end, CNN is working with IBM and Sony Electronics to create the CNN Global Content and Storage System. This is a 5 to 7 year, $20-million project to digitally archive 120,000 hours of CNN footage from the last 21 years as well as 15,000 to 20,000 hours of footage arriving annually at this news network.[36] On a smaller scale and with much more modest investments, some academic institutions are establishing repository programs without such significant costs. The California Institute of Technology, for example, is creating its Caltech Collection of Open Digital Archives (CODA) repository and keeping costs low by integrating repository and library services as much as possible and making use of free software such as ETD-db from Virginia Tech and Eprints from the University of Southampton.[37] The Hewlett-Packard Corporation and MIT have spent approximately $2 million on the development of Dspace, but now that this open-source repository system is past the research and development phase, MIT estimates that it will take $285,000 to operate annually for the scope of their program. A business plan with a complete breakdown of costs for staff, equipment, and supplies for MIT's operation of Dspace is available at the project's web site. Beginning with fiscal year 2003, the Ohio State University is investing $265,000 in new funds annually for the implementation and operation its Knowledge Bank repository program. For library programs the size of MIT or Ohio State, these projected costs for operating an institutional repository program are rather modest, amounting to just one to two percent of their total budgets.

But because they are new, the true costs of operating an institutional repository program are still tentative. How big an institutional repository becomes in an academic setting depends partly on the repository's collection policy and partly on the willingness of faculty and students to contribute digit assets to it. And what are the long-term costs of maintenance, particularly the costs for active preservation, of digital assets into the future? On the other hand, what will it cost an institution or society not to provide stewardship of its important digital knowledge assets? If Peter Drucker is right that the modern organization's value is based primarily on its collective knowledge, then investing in an institutional repository is a wise decision today for an institution. Librarians, of course, as a profession have to worry about access to knowledge on a broader scale and for a longer period of time. Institutional repositories offer hope that local resources will be committed to the development and application of metadata standards, open access strategies and connections, and careful preservation management to ensure that important digital knowledge assets will be safeguarded and accessible for future generations.

REFERENCES

[1] Dspace Federation, http://www.dspace.org/ (accessed August 2003)

[2] Crow, Raym. *The Case for Institutional Repositories: A SPARC Position Paper*. The Scholarly Publishing and Academic Resources Coalition: Washington, D.C., 2002.

[3]Drucker, Peter. The Coming of the new organization." In *Harvard Business Review on Knowledge Management;* Harvard Business School Press: Cambridge, MA., 1998. (Reprint of the article from the *Harvard Business Review*, January-February, 1988)

[4] Blair, David C. Knowledge management: hype, hope, or help? Journal of the American Society for Information Science and Technology **2002**, *53* (12), 1019-1028.

[5] Davenport, Thomas H.; Prusak, Laurence. *Working Knowledge: How Organizations Manage What They Know*; Harvard Business School Press: Boston, 1998; 144-148.

[6] Davenport, 5.

[7] Lynch, Clifford A. Institutional repositories: essential infrastructure for scholarship in the digital age. ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC **2003**, *226*, (February), 3.

[8] Preserving our digital heritage. RLG News **2003**, *56* (Spring), 10.

[9] Open Archives Initiative. http://www.openarchives.org/ (accessed August 2003).

[10] METS, http://www.loc.gov/standards/mets/ (accessed August 2003).

[11] Shareable Courseware Object Reference Model (SCORM), http://www.adlnet.org/ (accessed August 2003).

[12] Publishing Requirements for Industry Standards Metadata (PRISM), http://www.prismstandard.org/ (accessed August 2003).

[13] Dspace, http://www.dspace.org. Eprints, http://www.eprints.org. Fedora, http://www.fedora.info/. Bepress, http://www.bepress.com. Documentum, http://www.documentum.com, CONTENTdm, http://contentdm.com/. IBM DB2 Content Management, http://www-3.ibm.com/software/data/cm/. Artesia Technologies, http://www.artesiatech.com (accessed August 2003).

[14] ArXiv; http://www.arXiv.org (accessed August 2003).

[15] McSherry, Corynne. *Who Owns Academic Work? Battling for Control of Intellectual Property*; Harvard University Press: Cambridge, MA., 2001.

[16] Smith, Abby. *New-Model Scholarship: How Will It Survive?* Council on Library and Information Resources: Washington, DC, 2003, 3.

[17] Smith, iv.

[18] Crow, 2.

[19] Blair, 1023.

[20] Lavoie, Brian. Meeting the challenges of digital preservation: the OAIS reference model. OCLC Newsletter **2000**, *243* (January/February), 26 -30.

[21]Consultative Committee for Space Data Systems. *Reference Model for an Open Archive Information System (OAIS)*; National Aeronautics and Space Administration, Washington, DC, 2002. Sawyer, Don; Reich, Lou. ISO Reference Model For an Open Information System (OAIS); Tutorial Presentation, Library of Congress, June 13, 2003. http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html (accessed August 2003).

[22] Open Knowledge Initiative, http://web.mit.edu/oki/ (accessed August 2003).

[23] Open Source Initiative, http://www.opensource.org/ (accessed August 2003).

[24] Internet Archive, http://www.archive.org/ (accessed August 2003).

[25]Open Archives Initiative, http://www.openarchives.org/ (accessed August 2003).

[26] ArXiv; http://www.arXiv.org (accessed August 2003).

[27] Budapest Open Access Initiative, http://www.soros.org/openaccess/index.shtml (accessed August 2003).

[28] Public Library of Science, http://www.publiclibraryofscience.org/ (accessed August 2003).

[29] The OAI-PMH Protocol, http://www.openarchives.org/ (accessed August 2003). Lagoze, Carl; Sompel, Herbert Van de . The making of the open archives initiative protocol for metadata harvesting**.** Library Hi Tech **2003**, *21* (2), 118-128.

[30] Hitchcock, Steve. Metalist of open access e-print archives: the genesis of institutional archives and independent services; ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC **2003**, *227* (April), 4-11. http://opcit.eprints.org/archive-metalist12.html (accessed August 2003).

[31] Lynn, Chris. Users, vendors concur on maturing asset-management field; The Seybold Report: Analyzing Publishing Technologies **2003**, *3* (2), 17-21. Grimes, Brad. Digital asset management 101; TechNews **1998**, *4* (6). http://www.naa.org/technews/tn981112/editorial.html (accessed August 2003).

[32] University of Michigan, Digital Asset Management Initiative, http://sitemaker.umich.edu/dams (accessed August 2003).

[33]Smith, McKenzie. Dspace: an open source dynamic digital repository; D-Lib Magazine **2003**, *9* (1). http://www.dlib.org/dlib/january03/smith/01smith.html (accessed August 2003).

[34] The OSU Knowledge Bank Planning Committee. *A Proposal for Development of the OSU Knowledge Bank;* The Ohio State University Libraries: Columbus, Ohio, 2002. http://www.lib.ohio-state.edu/Lib_Info/scholarcom/Kbproposal.html; Rogers, Sally A. Developing an institutional knowledge bank at Ohio State University: from concept to action plan. portal: Libraries and the Academy **2003**, *3* (1), 125-136.

[35] Cornell University Copyright Policy, http:www.research.cornell.edu/CRF/Policies/Copyright.html. Brigham Young University Intellectual Property Policy, http://ipsinfo.byu.edu/ippolicyt.htm. (accessed August 2003).

[36] Gordon, Christine. IBM, Sony provide digital archiving to CNN; Internetnews.com, April 23, 2001. http:www.internetnews.com/bus-news/article.php/749871 (accessed September 2003).

[37] Caltech Collection of Open Digital Archives (CODA), http://library.caltech.edu/digital/default.htm (accessed August 2003).