Institutional Repositories

Sabbatical report

January 30 – July 9, 2005

by

Marilyn S. Billings

Summary

This sabbatical report is the culmination of several months of research, attendance at conferences or workshops related to the topic of Institutional Repositories (IRs), and site visits to institutions with representative IR solutions. The report begins with background work that laid a firm foundation for more focused research, then goes into specific benefits, core features and functionalities and possible content to be included in an IR. Since this is a relatively new service, there are challenges and issues facing those adopting IRs that need to be addressed. Many early adopters have provided guidance in this area plus have given examples of policies that need to be developed for a successful IR initiative. I will outline these areas and provide links to further content rather than repeat in this report what has already been done. The final sections of the report focus on lessons learned by early adopters of IRs and recommendations for next steps that the University could take to establish an institutional repository.

The links imbedded within the report and the appendices provide additional documentation and content to be used for this ongoing work. The enclosed CD contains the complete report with all supplemental material in PDF format.

Before proceeding, let's first define what is generally understood to be the definition of an institutional repository. According to Clifford Lynch, executive director of the Coalition for Networked Information, an institutional repository is "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members." Another popular definition is provided by Raym Crow in a 2002 SPARC position paper entitled "The Case for Institutional Repositories". He states "Institutional repositories … digital collections that preserve and provide access to the intellectual output of a single or multi-university community". IRs provide a compelling response to two strategic issues facing academic institutions. Such repositories

- provide a critical component in reforming the system of scholarly communication - a component that expands access to research, reasserts

control over scholarship by the academy, increases competition and reduces the monopoly power of journals, and brings economic relief and heightened relevance to the institutions and libraries that support them; and

- provide support to the university's digital scholarship; having the potential to serve as tangible indicators of a university's quality and to demonstrate the scientific, societal, and economic relevance of its research activities, thus increasing the institution's visibility, status, and public value.

In summary, IRs represent a rapidly growing movement in scholarly communication where the institution provides a web-based repository to collect, preserve, and provide access to the digital resources of the scholarly research process. Most of these services are envisioned to provide long-term access and preservation for these materials.

Background

In the fall of 2003, I submitted a sabbatical leave proposal to investigate the current environment of large online collections, or institutional repositories, to determine if such a project would be feasible for UMass Amherst to undertake. I began researching this topic in the fall of 2001; about a year after this concept began to appear in professional library literature and at library conferences. After further study, it became evident that more focused, dedicated time was needed to evaluate this new international initiative effectively. Since this was going to take substantially more time than I could devote during normal working hours, it became the clear candidate for my sabbatical research. While giving preliminary consideration to this topic for my sabbatical research, I had the opportunity to attend three important conferences to glean a better understanding of repositories. For additional information on these conferences, see Appendix A.

- American Library Association, Library Information Technology Association (LITA) pre-conference entitled "Practical Digital Libraries Overview" given by Dr. Edward A. Fox, Head of the Digital Library Research Laboratory at Virginia Technical Institute on October 12, 2001.
- Coalition for Networked Information (CNI) project briefing entitled "The Fedora Project: an Open Source Repository for the Management of Content and Services" given by Thornton Staples, Director of Research and Development at the University of Virginia Library and by Sandy Payette, Researcher at Cornell University, April 28, 2003.

- Northeast Regional Computing (NERCOMP) Library/IT Special Interest Group (SIG) entitled "DSpace@MIT: an Open Source Institutional Digital Repository presented by DSpace staff on April 30, 2003.

Following these conferences, I continued my research and began to do more in-depth study of a wide variety of related digital library topics such as digital preservation, interoperability and metadata standards, open source software and the open archives initiative. All the information proved critical to develop informed questions for the site visits component of the sabbatical research. The purpose of the site visits was multi-faceted and addressed the following topics.

- explore their rationale for undertaking such a project
- ask if they were to do it over again, would they do it, how would they do it differently, what did they wish they'd known or what resources did they wish they'd had available that would have made things go more smoothly along the development path
- identify which entity or entities at each institution was the driving force behind the project. Was that a good choice? Why or why not? Who should have been at the table that wasn't there, i.e. library, computing services, faculty, graduate school, etc?
- identify resources (both equipment and staff) that each site needed to implement an institutional repository,
- confirm with the sites the types of questions that need to be addressed and answered before undertaking such a project, for example issues of copyright, intellectual property
- review the results of my previous research, filling in gaps and addressing questions
- review business plans for sustainability and growth, including the staffing and other support needed, to continue the repository.

The questionnaire and results of the site visits are in Appendix F.

Prior to my sabbatical leave, I had the privilege of participating in two professional development opportunities concerning institutional repositories. In February 2004, the Deans and Directors of Libraries of the six New England Land-Grant Universities established a Task Force on Electronic Repositories to study the purpose and potential of IRs, to provide them with more information to see if potential existed for joint IR projects, to analyze options and provide recommendations for future action. Specifically, they ask the Task Force to undertake the following eight areas in the scope of their study: define the purpose and potential of a repository, identify major

software solutions, provide hardware and storage requirements, determine content issues, access restrictions, operational processes, potential models, and cooperation incentives. I was elected the chair of the Task Force that worked throughout the summer and fall of 2004. On behalf of the Task Force, I presented a summary of the white paper to the Deans and Directors of Libraries on December 17, 2004. While the Task Force was engaged in this work, I was asked to speak at NELINET's Institutional Repository Seminar on January 27, 2005 to report our findings and provide an overview of the topic to include the following points:

- give a brief introduction for what institutional repositories are and why institutions need them
- speak about how the New England Land-Grant group got together and discuss the Task Force's work and resulting white paper
- identify key policy issues that need to be resolved with institutional repositories.

The NECOP Task Force on Electronic Repositories final report and the NELINET presentation are in Appendix A.

Scholarly Communication Crisis

A major driving force behind the development of institutional repositories has been the dramatic shift in scholarly communication especially within the past three to five years. Scholarly communication is in the midst of a paradigm shift caused primarily by two factors, the increasing volume of information in digital form and spiraling publishing costs. Traditional print publications (books, journals) are being augmented by, in some cases replaced by, materials born digital. Spiraling costs of journals have necessitated major cuts in journal subscriptions, resulting in fewer journal purchases and reduced access to information. It has become clear that the changing economics of scholarly publishing are proving much too costly to sustain in the long run. We need a new model for managing our scholarly research output and disseminating that information to our users. Some of the issues we need to address when creating the new model include what do we collect, how do we address the needs of researchers, and how do we preserve and make accessible this body of material. This evolution of scholarly communication, using new technologies, digital information, and networked environments that are prevalent on college and university campuses today, has created a truly new environment for scholarly research which requires broad campus conversation. There are numerous resources available to assist

in that conversation, many of which are listed in Appendix C. There are additional resources for working with faculty and determining their needs and interest in Appendix E.

According to the results of my research, institutional repositories are seen as one of the key solutions to the scholarly communication crisis and to the advent of a variety of independent, silo-like web sites that store and provide access to digital materials. IRs go beyond the technology. They foster collaboration among librarians, archivists, information technology providers, the administration and others to provide institutional support and resources for this service. Their quick rise addresses many needs that have been identified: the escalating costs of scholarly materials, especially STM journals, the growing diversity of content in a variety of formats that needs to be grabbed and preserved before it disappears, escalating user expectations (immediacy, relevance, high level of management of content, resulting from experience with Google, Amazon), and faculty interest in new forms of scholarly communication for research and teaching (learning objects, e-Portfolios, and the like).

Although initially not part of my sabbatical research proposal, two areas that need to be considered simultaneously with the discussion of institutional repositories are learning objects and electronic portfolios since these initiatives may provide content for the IR. I found local workshops that focused on those topics to gain some expertise in each area:

"New England e-Portfolio Summit", June 2, 2004, Simmons College; NERCOMP SIG workshop

"Learning Objects", October 14, 2004, Smith College; NERCOMP SIG workshop

Two other events that provided additional resource material and network opportunities were:

"Institutional Repositories: Revealing Our Strengths", June 10, 2004; web-cast co-sponsored by SPARC and CARL. Co-sponsored locally by the University Libraries and OIT Academic Computing.

"Institutional Repositories: The Next Stage", November 18-19, 2004, Washington, DC. SPARC & SPARC Europe workshop.

Benefits of an IR

According to Kathleen Shearer in the SPARC web-cast mentioned above, a centralized repository provides common formats, uniform, well-defined structures,

provides other features such as searching or links from other sources to be applied consistently across the entire collection, data is more accessible and easier to use, there is greater integration with other related resources, and it is easier to develop additional tools for more functionality. Institutional Repositories have a unique niche in the organization to collect, organize, and provide access to a wide range of content that was previously inaccessible or scattered. They are now an important step in the evolution of scholarly communication. In the same web-cast, Daniel Greenstein from the California Digital Library described many value-added services that an IR can provide in supporting scholarship:

- greater links between content in IRs and scholarly publications
- better search and retrieval facilities
- enhance existing scholarly publishing (pre-prints, etc in digital form and interlink with paper)
- move scholarly communication forward
- harvesting of research
- broad range of scholarly materials
- enhanced professional visibility driven by broader dissemination and increased use

Core features and functionality

Let's now take a look at some of the features and functionality that need to be included in any discussion of an IR platform. Many of these items were repeated in numerous research articles and at workshops I attended. Common features of a repository include the following elements

- content is institutionally defined and scholarly in nature
- cumulative and perpetual
- open and interoperable
- content is in digital form in a wide variety of types (text, audio, video, images, data sets)
- community focus, where the community determines what is included in the repository. The community members are the authors and copyright owners of the deposited content.

- institutional support, requiring collaboration across an organization. The repository requires long-term financial support to ensure that the content is preserved and maintained.
- durable (persistent URL for material), permanent content that can be migrated over time. Digital preservation techniques need considerable more research to ensure lasting retrieval.
- access to content by a broad audience, a community-shared alternative to local storage of content, fosters serendipitous discovery across disciplines.

Desirable features

- Personalization of research pages by faculty
- Authoring needs such as version control, easy access to files

Core functions which are critical for a successful repository:

- material submission, some way for the author to deposit material, provide for editing to assure quality of content, conversion to archival format such as PDF
- metadata enhancement of content, such as author, title and descriptive information and administrative data such as date and time of submission
- access control, or digital rights management, to provide for controlled access to the repository content. Even if the entire community has access to the content, there needs to be a way to restrict the ability to add, delete, edit, and approve content.
- discovery support, usually a search engine that supports browsing and full-text searching of the content
- distribution and dissemination of content to enable display and download capabilities for further manipulation of content for ongoing research
- preservation, some mechanism(s) for the content to be preserved and retrievable over time, including a persistent documentation identification system
- batch loading capability so that, for example, articles can be loaded more easily from an open access publisher
- indexing of content based on OAI metadata harvesting protocol. The OAI website has a list of guidelines and protocol services.

Scholarly content

There has been an acknowledgement among early adopters of institutional repositories that recruitment of faculty content has been a challenge that needs to be addressed in order for IRs to be successful. Recent research conducted at the University of Rochester (Foster and Gibbons), has resulted in some concrete methodologies for others to follow to capture this new generation of scholarly work. Solutions include talking to faculty in their own language about grey literature, collaborating with the university press about monographs and faculty working papers or series, and batch loading of content from publishers adhering to open access principles. Some of this content is included in the following list:

- Working papers and other "grey" literature"
- Peer reviewed series (author or organization may want to create online version) and platforms for peer-reviewed journals (place where these can be created)
- Platforms for producing and distributing monographs
- Data archives (ICPSR and other large sets of research data)
- Classroom teaching materials and learning objects
- Digital archives
- Print repositories
- Student electronic portfolios
- Institution's annual reports
- Computer programs
- Data sets
- Photographs
- Art works
- Pre-prints
- Peer-reviewed articles
- Conference papers
- Electronic theses and dissertations

Policy Development

A successful IR cannot be developed without giving serious consideration to its overall structure and governance. One recommendation from early adopters is that an institutional repository needs a precise definition to be clear about what we're doing and what we're collecting, both within the institution and beyond. We need to

guarantee interoperability with other systems and repositories, provide persistence and integrity of content. This can be initiated by developing policies that, during pilot and implementation phase, should be kept in draft form to provide structure but also be flexible enough to change easily until it is working properly. Policies can then be finalized once the IR is in general release. Policies can be separated out into the following general areas:

1. Services:

What types of services are provided by the IR and its staff?

Is there a unit that will digitize materials?

What about assigning metadata, creating indexes and other support functions?

2. Content acquisition policies:

Who can deposit / submit materials (faculty only, students, librarians, etc)?

What types of materials can be deposited / submitted (pre-prints, post-prints, working papers, datasets, etc)

Who owns the content? What happens to the content if that person leaves?

What digital formats will be accepted (known) and preserved (supported)

What are the storage limits, if any (datasets especially can take up terabytes of storage)?

Can content be withdrawn? If it is cumulative and perpetual, material cannot be withdrawn except in rare cases.

3. Content management policies:

Quality control, who can approve (approval be department head, peer review, etc)

Who can update content, metadata, etc? Do we require only minimal metadata: author, title, keyword?

Which metadata scheme should be used? Follow metadata standards such as Dublin Core, METS, etc.

What about authority control? Do we allow subject area thesauri?

Does the software allow for separate communities / departments to use their own discipline-specific metadata schemes?

What about document version control? Do we keep only the latest version? Does the software do version tracking?

4. Preservation commitment:

How are materials going to be preserved over time?

Will the IR accept all types of formats? What do we agree to preserve and provide long-term access to (related to item 2 above)

What are current preservation standards that need to be followed? What developments and standards should be watched for future preservation needs and requirements?

5. Rights management

Although open, the IR may want to restrict access to some materials, parts of materials (chapters of books, etc) to within IP network, co-authors, etc for research, patent, or other reasons.

What about legal issues, copyright, intellectual property? How will that be handled?

Many early adopters have developed policies in all these areas and recommend that we borrow from other sites rather than re-invent the wheel. The SPARC web-cast participants offered many suggestions that are included in the site's resources. Dan Greenstein recommended that folks look at any of the California Digital Library services, particularly in the "about" section. The services mentioned in the IR Web cast are under the eScholarship program at: http://www.cdlib.org/programs/escholarship.html. Susan Gibbons did a search for IR policies on the web prior to conducting the LITA workshop "Establishing an Institutional Repository" in January, 2005. Please refer to the Bibliography and Appendix D for these and additional resources.

Staffing

Launching an institutional repository service does not in itself require a proscribed staffing plan. Much depends on the type of service one selects, whether commercial or open source, remotely hosted by the vendor or supported locally. Another consideration is how one plans to use the repository, start with a pilot, accept only a few format types at the start. However, once these decisions are made, a staffing plan needs to be developed. There are job descriptions from early adopter sites in Appendix B that outline the various types of skill-sets that are required to operate a sustainable IR. One could look among existing staff resources for these skills or define

new position(s) to take on this work. Based on the research and site visits, the general trend is to have this service managed within the library. One person is usually designated as the "Project Manager" with many of the other responsibilities distributed among existing staff as part of their usual responsibilities. Librarians have large-scale collection management experience, dealing with both the assessment and acquisition of scholarly content. Their skill in indexing and cataloging traditional library materials can be put to use creating metadata and indexing terms for the IR. Librarians have solid business practices when dealing with large collections and have the long-term commitment to preservation of scholarly materials that fits the mission of an IR. Some staff will need basic training in the new model of scholarly communication, open access, copyright, intellectual property and the like. Open Archives has developed the Open Archives Forum online tutorial: OAI for beginners to address some of these needs. Other resources are in Appendices C and D.

Software solutions

The major repository software systems marketplace falls into two main areas: open source products and commercially available solutions. I focused my sabbatical research and site visits on software that is in common use in the United States using either open source or commercial products. I have included cost estimates effective at the time of my research to provide examples of the financial commitment that needs to be provided by the institution for each type of solution.

Open Source software solutions: By definition, the open source software movement is an effort to "rationalize the software business" within higher education and to enable higher education to un-bundle software tools, not getting locked into one vendor. Open source software is freely available and OAI compliant. Further information comparing a variety of open source institutional repository software, unfortunately not including Fedora, can be found in the report written in Oct 2003 by the Open Society Institute entitled "A Guide to Institutional Repository Software". For those not familiar with the open source movement, freely available does not mean there is no cost involved. The software is free to download but there are hardware, configuration, maintenance, and sustainability costs that need to be taken into account.

DSpace

Developed jointly by MIT Libraries and Hewlett-Packard (HP), DSpace is freely available to research institutions worldwide as an open source system that can be customized and extended. DSpace is a groundbreaking digital institutional repository that captures, stores, indexes and redistributes the intellectual output of a university's research faculty in digital formats with an emphasis on preservation of digital materials. DSpace accepts all manner of digital formats. Some examples of items that DSpace can accommodate are: documents (e.g. articles, preprints, working papers, technical reports, conference papers); books; theses and dissertations; data sets; computer programs; visualizations, simulations, and other models; multimedia publications; and learning objects.

Sites using DSpace (DSpace Federation website) include: Columbia University, Cornell University, Harvard University, Massachusetts Institute of Technology, Ohio State University, University of Oregon, University of Rochester, University of Toronto, and University of Washington. Estimated costs from the MIT DSpace project: from $40,000 per year and up, depending on hardware, maintenance, local setup and ongoing sustainability costs.

EPrints

EPrints software has been created so that institutions can create OAI-compliant Archives quickly, easily and for free. EPrints servers are designed to help dissemination of research publications by sharing information (metadata) using the latest (OAI) standards.

Some examples of sites using EPrints software include arXiv, a large Open Archive in physics, mathematics and some related disciplines; Behavioural and Brain Sciences, an archive of preprints of target articles from the peer-reviewed journal of the same name; California Digital Library (3 Archives using eprints.org); CogPrints, a multidisciplinary Open Archive in the Cognitive Sciences (psychology, neuroscience, computer science, biology, linguistics, philosophy); Psycholoquy, an archive of refereed reprints, commentaries and responses from the online journal of the same name.

Fedora

Fedora was jointly developed by the University of Virginia and Cornell University. The project was funded by the Andrew W. Mellon Foundation to build an open-source digital object repository management system based on the Flexible Extensible Digital Object and Repository Architecture (Fedora). The new system demonstrates how distributed digital library architecture can be deployed using web-based technologies, including XML and Web services. Fedora is a general-purpose digital object repository system that can be used in whole or part to support a variety of use cases including: institutional repositories, digital libraries, content management, digital asset management, scholarly publishing, and digital preservation. In the newest version, the Fedora access interfaces now support date-time stamped requests, so that a client can "go back in time" and see a digital object as it looked in the past. Additionally, this release provides a migration utility for mass export and mass ingest of objects from either directories or other repositories.

Major features of Fedora software include: open source; web services; flexible digital objects with disseminators that associate a set of behaviors with a digital object; content versioning; XML ingest and export; digital object storage; access control and authentication; OAI metadata harvesting; migration and batch utilities. Upcoming versions of the software will add important functionality, such as Shibboleth-based authentication, fine-grained policy enforcement, workflow support, enhanced preservation features, and performance enhancements to support extremely large repositories.

Sites using Fedora include: Indiana University; New York University; Northwestern; Rutgers University; Tufts University; and University of Virginia. Cost estimates range in the same neighborhood as DSpace, depending on each site's unique needs.

Commercially available solutions: Many commercial vendors, including Integrated Library System vendors, are now offering repository solutions. When evaluating these products, it is critical to ensure that the vendor's product adheres to industry standards and is easily portable to another platform. Some of these vendors offer vendor-hosted solutions, meaning that the vendor provides hardware and technical support. This gives the institution an opportunity to pilot the IR without the heavy hardware and technical staffing commitment.

BioMed Central's Open Repository

Open Repository is a service from BioMed Central which builds and maintains repositories on behalf of institutions and organizations. This software allows organizations to quickly set up and run an industry-standard DSpace repository with minimal effort and investment in manpower and cost. At all times the customer retains ownership of the repository. Open Repository is built using the open source technology platform, DSpace. It allows institutions to migrate to their own in-house DSpace repository at a later date to take over its operation and maintenance. Major features of Open Repository software include set up, customization, service maintenance and ongoing support, document upload and formatting, populating the repository through data feeds such as BioMed Central and PLoS, search functionality, customer service and support, and the ability to showcase and exposure of the organization's intellectual output. To date, there are very few sites using OpenRepository. The University of Connecticut is currently conducting a three month trial.

Costs, as shown on their website, vary depending on the size of the site and amount of service from $9,000 - $27,400 with annual maintenance fee from $4,500 - $9,000.

CONTENTdm

CONTENTdm provides tools to organize, manage, publish and search digital collections on the Web. This flexible, multifunction software package handles multiple formats: documents, PDFs, images, video and audio files. CONTENTdm offers a scalable solution to grow with a repository's needs. It offers both a locally hosted and remotely hosted option and is intended to provide a solution for digital library collections, not to provide a complete institutional repository solution.

Sites that use CONTENTdm include: Colorado State University Libraries; Mount Holyoke College, University of Illinois at Urbana – Champagne, and University of Washington Libraries.

DigiTool

Ex Libris' DigiTool is a module offered by the new Integrated Library System that the Five Colleges are in the process of implementing. The Five College libraries are changing Integrated Library System vendors from Innovative Interfaces to Ex Libris. Each of these ILS vendors has a digital asset management module. However, if we are looking at our ILS vendor to provide this solution, I recommend that we look at

the Ex Libris product known as DigiTool. A new version was just released which provides vast improvements in functionality. One of my site visits mentioned that it has functionality similar to the Fedora product mentioned above.

Sites using DigiTool include: Brandeis, Boston Public Library for the Massachusetts Online Digital Library Initiative. The price quote we received for this product, including discounts, is $57,270 plus implementation services cost $15,000. Annual maintenance fees are $13,410.

UMI/ProQuest's Digital Commons@ (owning institution's internet address, e.g. umass.edu)

Digital Commons is a Digital Institutional Repository service provided by UMI ProQuest. It is OAI-compliant and provided as a partnership with the Berkeley Electronic Press (Bepress). Digital Commons is designed to be a remotely hosted solution; UMI manages the repository software and hardware.
Features of this product include: immediate population of content – theses and dissertations uploaded by ProQuest; full text searching and new features such as access paths to this work via Google's scholar search capability; integrated management tool; incorporates a wide variety of publication types, formats, and dynamically generated resources (articles, preprints, reports, sound, video, data sets, executables); and provides usage statistics. One of my site visits commented that they had a repository up and running in a day with this product.

Institutions using Digital Commons include: Amherst College, Boston College, California Digital Library, Florida State, New England Law Library Repository, and the University of Connecticut.
Cost: institutions pay at different rates for different levels of access and storage. The cost of a pilot project is $12,500 with limited storage capacity, the cost of unlimited access and storage is $34,500 per year.

Challenges / Issues for IRs:

Content: Populating the Institutional Repository

As stated earlier in this report, researchers need encouragement to put material into the IR. Frequently they are unaware of the benefits that an IR can

provide. Faculty respond best when those explaining the role of an IR speak directly to their need to be recognized, cited, and visible to their professional colleagues. Attributes such as the archival function of the repository, preserving a diverse mix of content, a place to collect and showcase their intellectual capital, easy access to articles and other content, faster dissemination of research results than paper, communication and outreach address their needs. Some other ideas for recruiting scholarly content came from my research and discussion with colleagues during site visits. One idea that came from the University of Rochester research cited earlier (Foster and Gibbons) is to gain a good understanding of the scholarly communication within a given discipline and how an IR might fit into the existing model. U of R created a grey literature sheet that lists the types of grey literature by discipline that one can use when talking with faculty, using their own language to discuss potential content for the IR. When the IR is presented as infrastructure for "non-traditional" material rather than published literature, then there is tremendous interest by faculty. They highly value the peer-review process so it can be best to recommend that the IR provides a complementary role to the existing publishing system. Other strategies include focusing on:

- retiring or recently retired faculty
- faculty with personal websites full of links to their works
- departments that have conference proceedings or working papers
- faculty who have signed the Budapest Open Access Initiative

According to Jim Pringle at Thomsen ISI, if one examines the Project RoMEO / Sherpa list of green publications, at least 56% of articles indexed by ISI are able to go into an institutional repository now. These are peer-reviewed journals and provide a quick way to add valuable content. And Kathleen Shearer stated that many publishers actually allow the right to deposit e-prints into IRs. Recent stats from the RoMEO list show 63% of publishers on the list formally allow some form of self-archiving (http://www.sherpa.ac.uk/romeo.php?stats=yes). Recent changes to the Elsevier copyright policy highlight this trend, as does the NIH policy requiring publicly funded research to be made available through a service such as an institutional repository.

A new methodology adopted by the University of Rochester after they completed a study of how faculty do their research is the addition of faculty researcher pages to their repository. Faculty can create their own space, highlighting their own research. Several repository software systems are now starting to adopt this feature into their own architecture. We need to do more study in this area to facilitate a cultural shift in the dissemination habits of researchers.

Funding / business models

How will the IR be funded? What business model will be used? How will the institution ensure long-term sustainability of this service?
My research indicates that these issues become the responsibility of institution. For example, at the University of California, IRs are part of the core information infrastructure that the university offers (through its libraries) – Dan Greenstein. I did find some specific examples of how some IRs are dealing with funding issues:

- Metadata enhancement services: charge back to content owner's department
- Annual storage costs based on sliding scale back to department
- One organization charges individual content owner for each deposit. Ancillary services are purchased through a local digitization vendor.
  Other institutions go completely the other way, completely subsidizing the IR stating that the increased visibility and use of the institution's scholarly works more than pays for the IR.

Legal issues

The top two topics that need to be addressed in this area are intellectual property rights and copyright. Early adopters have been working through these issues and developed policies that others can use. The sites mentioned earlier in the report have many resources to consult.

Marketing and publicity

Please see below in the "Lessons Learned" section for sage advice in this area.

Organization and administration:

Where will it be located? Who will manage the IR? What will be the relationship between the IR center and the departments, research communities of the institution? Earlier in this report, I referred to the results of my research that indicate that the library is generally the campus unit where this service is managed. However, these questions need to be addressed.

Preservation

This issue is a long way from being solved. There are currently at least four approaches to this problem: migration, technology preservation, emulation, and persistent object preservation. Libraries have some experience with preservation, especially in traditional library roles. Many libraries have been developing digital collections for years and will need to extend the knowledge gained developing these collections to supporting IR content and metadata preservation. One way to help with this issue in the short term is to try to limit future liabilities by restricting the number of formats that will be preserved (known and supported). The University of California is ensuring that information is provided in supported formats and metadata is sufficiently rich. MIT has developed a chart of known and supported format types that it will commit to preserving over time. Another way to put the IR in the best position for long-term preservation is to adhere to industry standards and best practices. IR managers will need to develop a coordinated strategy to ensure long-term preservation of the repository's content, possibly some combination of the strategies mentioned here. It is one area that will need ongoing attention.

Lessons learned from research and site visits:

As I have mentioned throughout this report, an important part of my sabbatical research included visits or extensive phone calls to sites that have already implemented repository software. The questionnaire plus individual site responses are in Appendix F. In addition, the contacts made at workshops and important lessons learned from early adopters have given me valuable guidance and approaches for next steps for the university. This section highlights the lessons learned from these resources which I will then turn into some next steps for the university to take in developing an institutional repository.

Communications

- Start talking about scholarly communication and the benefits of a repository early in the process
- Educating the faculty is of key importance. They need to understand the importance of a repository and adjust customs and culture
- Repository policies need to have clear explanations and examples. Use policies, etc from early adopters

Planning:

- Simultaneous with technical planning and installation, do business and operational planning
- Be flexible and responsive, especially when defining service model. Be prepared to change during implementation
- Prepare for additional storage capacity

Recruitment of early adopters

- There's no such thing as too much publicity
- Starting new communities is a long process, don't get frustrated
- Community policies need to be established at the top level, for example, chair of an academic department, head of a research institute, etc
- Use terminology that faculty members understand. Visibility of their content and digital preservation are critical issues for them. Don't get technical
- Personal contact is the best means for finding pilot projects. Talk with individual faculty and small groups
- When considering pilot projects, select those that provide a diversity of formats, storage requirements, and size
- Collection development coordinators are most helpful in recruiting early adopters
- Survey content providers / communities yearly for use, management, etc. Stay in touch with them, provide training and other support as needed

Top selling points

- Persistent identifiers for content; visibility of content
- One uniform presence for institutional research
- Community and author control
- Self-archiving clause for green publishers
- Digital preservation of approved formats
- Visibility of research leads to more recognition, contribution to scholarship in discipline

Staffing

- Educate library staff through presentations, frequent updates in library communications, updates at departmental and staff meetings to increase awareness and understanding
- Use cross-functional library staff teams when building repository's business plan
- Integrate repository in all aspects of library operations (collection development, acquisitions, cataloging, reserves, reference, information literacy, liaison program)
- Build the ability for staff to be able to manage through uncertainty by training in change management using a multiplicity of methods
- Support staff through professional development opportunities

Content submission

- Faculty have department or administrative staff submit content
- Set expectation that new communities will launch in bi-annual releases to motivate continuing development
- Provide frequent training sessions to content submitters on entering metadata
- Authority control is a large issue in metadata submission. Try using dropdown menus for author names, department names, etc
- In addition to faculty, train graduate students in submitting content

Publicity and marketing

- Ongoing communication throughout campus, timed to match academic calendar
- Continued contact with existing communities to provide support, training, move project along
- Use success stories, with quotes from early adopters, to market service to other faculty to recruit new content
- Work with Office of Grants and Contracts Administration to reach faculty that need to show funding agencies how their work will be distributed and preserved
- Invest time and money to get a community pilot started
- Address each discipline's needs using its own vocabulary, content examples, and grey literature

- Educate faculty in the topics of scholarly communication, digital preservation, etc and how an IR can help them
- Demonstrate to faculty the ease of submitting content, using the repository to find content
- Emphasize the persistent identifier for content, a key selling point for faculty
- Word-of-mouth publicity among faculty and other users of repository. Tap into this resource providing accurate information
- Success is not usually immediate

Impact on stakeholders

Libraries have a strategic and tactical role to play in moving beyond their traditional custodial role to active engagement in the scholarly communication enterprise. As mentioned previously, they have expertise in organizing and maintaining digital content, document preparation expertise (format control, archival standards, etc), expertise in metadata tagging, authority control, content management to increase access to and use of data, and work closely with faculty, administrators and other stakeholders in today's library environment. In the future, librarians will need to provide support for faculty as information content contributors and end users and can support faculty's open access digital publishing activities, given the appropriate training. Librarians will have an extraordinary visibility within the university as they take on these roles, and will become more critical to the university over time. Therein also lies one of our challenges, to maintain traditional library roles while developing the new.

Next steps:

After the extensive research on institutional repositories that I have done the past few months, I have noted several common themes that the University could use as initial steps in working toward an institutional repository.

1. Create a Scholarly Communications resource page. See Appendix C for a list of possible contents for the web page.
This page would be useful for librarians when talking with faculty and administrators about scholarly communication, copyright and intellectual property issues. Parts of the page could be used for publicity and marketing, others by those charged with

developing policies and other materials related to repositories. Most library web sites I visited have a section devoted to the scholarly communication topic, including the steps that the library has taken and the steps that faculty can take to address the changes in scholarly communication.
Content could be taken from numerous sources, for example

- presentations the University Libraries have given to the Faculty Senate on this topic, March 29, 2001 and to the Deans during fy '05
- the scholarly communication colloquium series that the University Libraries have co-sponsored with other research support organizations on campus
- ACRL's Scholarly Communication Toolkit
- ARL scholarly communication website
- Create Change website
- SPARC website

The web page should publicize steps the University Libraries and other campus entities have already taken to move toward a new scholarly communication model.

- Membership in BioMed Central
- Participation in ARL Create Change
- Sponsorship of the Scholarly Communication Colloquium series

2. Create an Institutional Repository Task Force that is representative of the institution to be charged with the governance and oversight of the remaining "next steps" recommendations in this report. Based on my research, I recommend that the University Libraries be charged to lead this effort with substantial institutional support for funding and university buy-in provided at the highest levels. It is recommended that one person be charged to coordinate this effort to maximize the success of this effort.

Membership in the Institutional Repository Task Force should include library administration, university administration, Office of Research, IT staff, faculty (possibly IT Task Force member), archives staff, others as appropriate.

3. Conduct a needs assessment throughout the University to determine the need for a service that an institutional repository could provide and to start building awareness of this effort. The survey should have as one of its goals the identification of potential pilot projects and faculty. Examples of survey instruments are in Appendix E.

Entities that should be interviewed during the needs assessment include

- Academic Departments and Deans
- Center for Teaching
- Commonwealth College
- Graduate School
- Office of Information Technology, especially the Academic Computing Department
- Office of Research
- Research Institutes and Centers

3. After the needs assessment is completed, the Task Force needs to define the services and elements that the University finds essential to meet existing and future needs, a requirements document. It could use the core features, functions, and services discussed previously in this report as a basis for this document, take additional recommendations from the results of the survey, and use all this data for the development of an RFB if required.
Decisions that need to be made:

- Decide to go with a commercially available vendor solution or with an open source product. Pros and cons of each solution type have been proved earlier in this report. Examples of potential vendors and open source solutions are also listed previously in this report.
- Depending on the above decision, determine source of ongoing technology infrastructure and support
- Determine the source of pilot and ongoing funding. Costs (depending on solution): hardware, software, associated maintenance contracts, promotion / publicity, and staffing.
- Staffing needs: manager / coordinator, systems administrator (depending on solution), cataloger, collection development and reference liaisons with faculty and other content providers, administrative support. Examples of position descriptions are included in Appendix B.

4. Simultaneous to the development of a requirements document or RFB, the Task Force needs to identify a Policy Advisory Group whose charge is to develop policies and guidelines that provide some of the infrastructure for the IR. Examples of policy statements are included in Appendix D.

From the research I've conducted, the minimum policies that need to be in place include

- Services that are available from the repository
- Who may contribute to the repository
- Appropriate content for the repository
- Management of that content, indexing and metadata creation
- Preservation and migration of content and metadata over time
- Rights management, copyright, and other legal aspects
- Decision-making for overall repository

Membership should include library staff, university administrators, faculty, legal counsel or copyright expert, archives staff, IT staff, others as needed.

5. Concurrent with item 4, assuming that the University Libraries is taking the lead in this effort, a Library IR Implementation Task Force needs to be formed and charged with developing concrete ways to integrate IR efforts with existing library services, to obtain content for the IR, provide associated metadata, and provide the services defined in item 3.

Building from the results of my research, content can be obtained from faculty, graduate students and others defined previously in item 2. Determine content quality / quantity issues to provide guidance for appropriate content. Strategies to obtain content include

- Approach faculty and research institutes and centers for their grey literature. See Appendix E for documents that define what grey literature is and for a list of grey literature by discipline as defined by research conducted by the University of Rochester library.
- Examine the SHERPA/RoMEO publishers copyright listings
- Talk with entities listed in item 1 above for additional content such as electronic theses, dissertations, final student projects, student e-portfolios, learning objects, and other materials as listed previously in this report

To provide the pre-defined services, the Library Task Force needs to define the appropriate standards that the institutional repository will be built upon, including

- interoperability (open standards to communicate with other systems and provide open access, XML, OAI, crosswalks to existing date such as MARC)

- metadata (descriptive, administrative, structural, and technical)
- type of metadata used (Dublin Core, METS, EAD, etc)
- digitization standards
- preservation standards

6. The University IR Coordinator and others involved in the IR initiative need to stay abreast of ongoing developments. Particular areas to pay attention to are recent developments in the U.S. and Europe in support of public access to research and ongoing preservation efforts.

The most important piece of advice I have received numerous times from many early adopters and site visit contacts is to just get started. It's the best way to learn. There are many resources and people available to assist with the project, provide counsel and advice. I strongly recommend that the university move forward with this service.

Post-sabbatical professional development opportunities:

I have organized a NERCOMP workshop on Repositories for the fall which will provide a tremendous learning opportunity for the New England region. I will do a presentation that will summarize the key findings of this sabbatical research and provide a broad framework of the issues for the other speakers. Other speakers, many of whom I met while doing site visits, will provide specific examples of repository solutions, both open source and commercial. Speakers from sites using DSpace, Fedora, CONTENTdm, Digital Commons and Ex Libris' DigiTool will all provide insights about their sites and offer suggestions to attendees. This NERCOMP workshop on Content Repositories will held at College of the Holy Cross on September 26, 2005.

I have been in touch with John Webb, Assistant Director for Digital Services / Collections and current editor of the Information Technology and Libraries. He is interested in an article reporting the results of my sabbatical, especially the lessons learned and site visits sections. ITAL is published quarterly so I hope to get a manuscript to him by mid-December 2005.