Chapter 2

DEFINING AN INSTITUTIONAL REPOSITORY

Despite of the amount of press that the topic of institutional repositories has received in the past few years, providing a concrete, precise definition of an institutional repository (IR) is difficult to do.

Clifford Lynch (2003), executive director of the Coalition for Networked Information, describes an IR as "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members."

Other than replacing "a university" with "an organization," this definition reflects how IRs are discussed in this report. Lynch's definition is particularly appropriate because services, and not technologies, are the emphasis.

Nevertheless, even with this definition in hand, the question remains: what exactly comprises this "set of services" to which Lynch refers? Unfortunately, this question has no single, correct answer because an institutional repository must be *institutionally defined.*

To be successful an IR must provide the set of services needed by its unique community of users, and these services will and should differ from institution to institution.

Core features

Despite this ambiguity, all IRs share some common features and functions, and it is this combination of features and functions that distinguishes IRs from other types of services.

The first of five common features, and perhaps most obvious, is that an IR contains digital content. The range of different types of digital content can be vast, including text, audio, video, images, learning objects, and datasets. The material may be born digital or of a physical medium that has been digitized, such as scanned images.

5 core features of institutional repositories

- Digital content
- Community-driven & focused
- Institutionally supported
- Durable & permanent
- Accessible content

Regardless of the details, the content in an IR is always in digital form. An IR is not a replacement for the physical collections of a library or archive, or the artifacts of a museum.

A second core feature of an IR is that it is community-driven and community-focused. The community of users not only determines what should be deposited into the IR, but they are individually responsible for making the deposits. The members of the community also are the authors and copyright owners of the content. As a result, the IR reflects or showcases the scholarship, research, and interests of an organization.

A third core feature is that the IR has institutional support. An institutional repository is not a simple or cheap undertaking. A successful IR requires

collaboration among divisions across an institution, which is accomplished most easily with top-down institutional support.

Moreover, an IR necessitates ongoing, long-term financial support to ensure, for instance, the integrity of the content through digital preservation. Without an institutional commitment to the project, the costs and obligation of an IR likely are too great for any one department or unit to bear on its own.

An assumption of durability and permanence is the fourth common feature of an IR. When a digital file is deposited into an institutional repository, the author expects that the document will remain there for the perceivable future.

The frustrating experience of receiving an "HTTP 404 Error- File not Found" message while trying to retrieve a Web document should not happen with materials in an IR.

The supposition is that the content in an IR is persistent and permanent. Whether digital preservation techniques can make this supposition a reality remains to be seen. But the assumption of permanence is there, and to many content owners, this feature is the most appealing of an IR.

Finally, an IR is not a black archive; the content in an IR is not hidden from the entire world. With some exceptions, the content of an IR can be accessed by more than just the content's owner because the material within an IR is meant to be shared.

Access can vary greatly from just a few people within a department to the entire world, but the cost and scale of an IR discourages its use as a person's private, digital dumping ground.

An IR is not a replacement for someone's hard drive, but rather a community-shared alternative to it. This description implies the material deposited into the repository is scholarly in nature and ready for public dissemination, such as preprints, conference papers, and other types of scholarship.

Core functions

All IRs have core functions. The six core functions are material submission, metadata application, access control, discovery support, distribution, and preservation. Although an institutional repository may provide for many more functions, these six are essential, and a system lacking any of these cannot adequately support an IR.

6 core functions of an institutional repository

- Material submission
- Metadata application
- Access control
- Discovery support
- Distribution
- Preservation

An IR system must include some means by which an author or proxy can submit content to the system. Many times, material submission is accomplished through a Web-based form, which includes a file upload feature. Usually the document submission process is simplified so that anyone, with little or no training, can make a submission.

An IR system also may have the option of appointing several editors with tasks such as assuring quality of the content, judging the appropriateness of the document's inclusion to a particular collection, and enhancing metadata. The material submission process might include added features such as automatic document conversion (Word to PDF) or e-mail alert services. Regardless of the means or the accompanying bundle of additional features, material submission is a core functionality of all IR systems.

Each document within an IR requires some level of metadata. Usually a set of basic identification metadata, such as title and author, is required as part of the submission process. Abstracts, keywords, and other descriptive metadata fields also are common, although usually optional.

The system itself adds administrative metadata, including date and time of deposit and identity of depositor. The system also may provide for the role of a metadata expert, such as a library cataloger, who can enrich the metadata by adding subject terms and name authority control. Just as with the library catalog, the richer the metadata, the more accessible the collections become.

Access control, also referred to as digital rights management (DRM), is another core function of the IR. An IR system must have controlled access to the content. This access may be accomplished by integrating an organization's authentication or identity management system with the IR.

Other IR systems rely on logins and passwords distributed by the system's administrators. An institution can control access by IP ranges or by limiting rights to just those computers mapped to the IR servers. These access controls ensure that only appropriate people obtain an IR's content. Even if all the content of an IR is to have worldwide access, the system must still ensure that only authorized people can add, delete, approve, and edit content.

All IRs must have a discovery mechanism by which users can ascertain its content. Most commonly, this mechanism is a search engine, although the search engine's sophistication can vary from just a handful of searchable metadata fields to full-text searching of the documents themselves.

Lucene, http://jakarta.apache.org/lucene/docs/index.html

IR systems commonly rely on a third-party search engine, such as Lucene from the Apache Software Foundation. A system also can support discovery through browsing, which provides an overview of the type, breadth, and relationship of the content contained within the repository.

Closely intertwined with access control and discovery mechanism is an IR's distribution function. Once an authorized user locates the desired content, the IR system must then have a mechanism by which a copy of the digital file can be provided or displayed to the user.

Depending on the type of file, the system may require that users first download the document onto their computer, and then open it using software on the computer, such as Microsoft Excel. Some file types can be displayed directly through the Internet browser using plug-ins, such as Adobe Reader.

Preservation is the sixth core function of an IR. Although no librarians have a definitive answer for the preservation of digital documents, an IR is based on the assumption that the documents will be retrievable in the short and long term.

To assist with short-term preservation, IR systems support some means by which its content and metadata can be backed up. For long-term preservation, the system may include ways to identify and isolate files by type to assist in their migration. Or the system can facilitate the conversion of less preservable formats on submission, such as Microsoft Word to PDF or HTML.

CNRI Handles, www.handle.net

DOI system, www.doi.org/index.html

The incorporation of a persistent document identification system, such as CNRI Handles or DOI system, is another piece of an IR's preservation function.

Context of scholarly communication

Institutional repositories reside within the greater context of digital scholarly communication. In this context the full promise and value of an IR can be seen, as well as the potential controversies.

Since the early 1990s, scholars have contributed content to digital collections. At present numerous subject-based repositories exist, such as arXiv. ArXiv is an electronic preprint service for physics begun in 1991 by Paul Ginsparg, a scientist formerly at Los Alamos National Laboratories (LANL).

arXiv, http://arxiv.org

Often digital collections, such as arXiv, are tied closely to a single person or group of people, and so too is the fate of the collections. For instance, when Ginsparg joined the Cornell University faculty in 2001, arXiv made the move from Los Alamos to Ithaca, New York, as well. Hypothetically, Ginsparg could decide tomorrow that he is no longer interested in continuing to expend the time, money, and effort to run arXiv and unplug the servers.

Although subject-based repositories, such as arXiv, frequently depend on the efforts of single person, the responsibility of an IR falls to the entire organization. Institutional repositories provide a centralized framework in which faculty, researchers, scholars, and others can build their digital collections.

These collections may be subject-based like arXiv, or may be more institutionally focused, such as a university's electronic dissertations. IRs provide these digital collections with an infrastructure and permanence that can sustain changes, including a person's retirement or change of interests.

By themselves, institutional repositories are akin to islands of information across the landscape of the Web. Fortunately, discovering what is contained in the IRs does not require a visit to each.

Because of the work of the Open Archives Initiative (OAI), information about the content of institutional repositories can be pulled together into one place. The OAI Protocol for Metadata Harvesting (OAI-PMH) provides "an application-independent interoperability framework based on *metadata harvesting.*"

Open Archives Initiative, www.openarchives.org

(Source: quoted from www.openarchives.org/OAI/openarchivesprotocol.html (emphasis theirs))

An OAI metadata harvester is a service that can obtain the metadata for the individual items within any OAI-compliant digital repository. The metadata harvesters extract only metadata and not the actual documents. Once harvested, the metadata can then be searched, while providing a pointer back to the actual item in its native repository.

An excellent example of an OAI harvester is OAIster, a project of the University of Michigan's Digital Library Production Services, originally funded by an Andrew W. Mellon grant. As of April 2004, OAIster was harvesting the metadata of more than 3 million documents from more than 270 repositories across the world.

OAIster, http://oaister.umdl.umich.edu/o/oaister

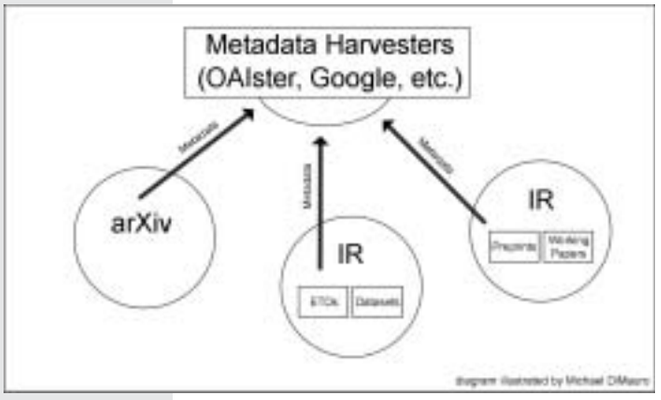
Although OAIster is a subject-neutral harvester, subject-specific harvesting also is possible, including the Digital Gateway to Cultural Heritage Materials at the University of Illinois at Urbana-Champaign.

Digital Gateway to Cultural Heritage Materials, http://oai.grainger.uiuc.edu

Google also is interested in the harvesting of IRs. Recognizing that most IRs contain high-quality, scholarly material, Google has partnered with 17 universities that have IRs in a pilot project. The goal of this project is to create a scholar's search portal, similar to the advanced search service that Google already has for government information.

To summarize, the emerging digital scholarly communication paradigm is one in which digital documents, such as electronic theses, preprints, digital images, and conference papers, are gathered into digital collections.

Some of these collections are housed and cared for at their home institutions within institutional repositories, and others exist as independent, subject-based repositories. Connecting all these disparate collections are OAI metadata harvesters and other search services, which facilitate seamless searching across all the repositories.



Digital Scholarly Communication Paradigm.

Most commonly, the scholarly material shared in this manner is gray literature. IRs become most controversial, however, when used as a way to disseminate scholarly communication as an alternative to formal publication. Doing so may have both positive and negative implications for the author.

On the one hand, the article may not carry as much weight on the author's CV as would an article published in a peer-reviewed, established journal. On the other hand, because the article was freely accessible to the world from the IR and not limited just to those colleagues with access to the journal's subscription, it may be cited many more times, and therefore have a larger impact in the field of scholarship.

IRs are a relatively new entrant into the scholarly communications paradigm. For some people, the IRs provide an infrastructure to store and distribute their previously underused gray literature. And for others, IRs are a mechanism by which to bypass traditional publishers and work toward a possible solution to the current scholarly communications crisis.

Gray literature is informal scholarly communication, such as dissertations, preprints, and technical reports that is not published through commercial publishers.