Libraries in the Lead:
The Institutional Repository Phenomenon

MacKenzie Smith
Associate Director of Technology
Massachusetts Institute of Technology Libraries
kenzie@mit.edu

Abstract:
As scholarship, instruction, publishing and communication become increasingly networked and digital, how libraries respond? Can libraries help scholars communicate in a networked era? What is the library's role in an age of Web publishing and Google? Is preserving digital collections still part of their mission? Institutional repositories begin to address these questions and allow libraries to show initiative and leadership in a scholarly world being transformed by technology. The Massachusetts Institute of Technology in the USA with its DSpace institutional repository program has witnessed how much has changed, and continues to change, as libraries step up to these challenges.

Over the past year, as the full impact of digital scholarship, instruction, publishing, and communication begins to become clearer, the library world has begun to see the possibility of a fundamental change in its role as the steward of the scholarly record.

Research libraries and institutional archives are, among other things, cultural memory organizations with a mission of collecting, managing, and preserving the scholarly record, while simultaneously supporting the active research and teaching programs within their institutions. In the past these two activities were quite separate in the information life-cycle: resources were selected and acquired, described and made available, used in teaching and research – usually more often initially and less so over time – and only much later (often decades) conserved or preserved for ongoing usability, if they had lasting value. The research and teaching products of the library's constituents – their research articles, notebooks, data, lecture notes, and so on – were only acquired if formally published, or in some cases by the institutional archives.

As scholarship, instruction, publishing and communication go digital, this traditional information life cycle is changing. For example, the products of formal publishing are not always available for accessioning into local collections – they are instead access by licensed directly on a publishers website, where they may or may not exist in the remote future. As another example, inaccessibility of research and teaching resources in digital form due to format obsolescence often occurs long before their initial value has begun to diminish.

How must libraries respond to this? Can libraries help scholars communicate in the digital era? Do libraries have a future in an age of digital publishing and Web indexes like Google? Is preserving their collections still part of their mission? If libraries don't, who will?

Institutional repositories are a first step towards addressing these questions, and allow libraries to show leadership and initiative as the scholarly world transforms itself into networked, digital everything. Institutional repositories are a new concept, but are usually defined as having the following four attributes:

- Institution-based
- Scholarly material in digital formats
- Cumulative and perpetualOpen and interoperable

From this definition we see that institutional repositories bear many characteristics of a traditional institutional archive, except that the content is always digital, and is usually aimed exclusively at research and teaching material rather than institutional records or special collections. It seems clear that institutional archives represent a new role for libraries that blends its traditional role with that of the archives, and it similarly affects the role of the institutional archive as the institutional output, especially the unpublished documents and other research data, becomes digital.

Faculty at the Massachusetts Institute of Technology are early adopters of advanced technology, and so have become both adept and dependent on technology in their research and teacher over the past few decades. In particular, in the past five years the trend towards online, digital publishing and sharing of digital research data has exploded on campus. As the MIT Libraries observed this phenomenon it was clear that immediate action was needed. Starting in 2000, the MIT Libraries, with funding and collaboration from Hewlett Packard, created an institutional repository system called DSpace. The vision for this project was of a federated repository to make available the collective intellectual resources of the world's leading research institutions, and the mission of the project was to create a scalable digital archive to preserve and communicates the intellectual output of MIT's faculty and researchers and to support adoption by and federation with other research institutions.

MIT was an early instigator of the institutional repository concept, is running the DSpace system as its own institutional repository, and has been actively collaborating with other research universities world wide to help define what these systems should do, and how they can help research universities deal with the technological transformation they're undergoing. DSpace was publicly launched in November of 2002 and the DSpace group at MIT has watched with excitement as the institutional repository phenomenon has begun to take root and flourish throughout the world. Our experiences over the past year show how much is changing.

To forward these goals, DSpace was launched in November of 2002 as a free, open source system which any person or institution anywhere could download and run locally. The DSpace open source license allows for local customization, including building commercial applications based on the distributed code. Governance structures are being developed to support contributions of new or improved features and functionality by adopters back into the open source codebase for the benefit of all the other DSpace adopters. The system can be used for it's originally intended purpose of serving as an institutional repository, but it could also work in a variety of other types of applications where a simple digital asset management system or persistent digital repository is needed.

Since the 2002 launch, the DSpace system has received much publicity worldwide and consequently much scrutiny. It's strengths and weaknesses have been publicly noted and debated on various mailing lists, and the decisions of the original development team about such things as which persistent identification system to use, what type of descriptive metadata to support, or how items should be submitted to the repository, are hotly contested. This is all to the good, and exactly the goal. The more brains that are thinking about these issues, and the more eyes that are pointed at the code, the better we'll understand what to do and how to do it – open source software and open communication is what will allow to make real progress together. Having a real system to look at, love it or hate it, focuses discussion and allows us to think more concretely about what we need.

One of the more interesting and difficult aspects of institutional repositories is the area of local policy development. Libraries and archives have highly developed policies and practices around what they acquire, how it is added to the collection, who will have access to it, if and when it can be superceded or withdrawn, and so on. But when collection development becomes the basis for collaboration with faculty to acquire and manage their actively used research and teaching material, these policies become a negotiation between faculty and librarians who both have a vested interest in the decisions and do not always agree. For example, faculty might wish to replace an earlier version of an article with a new version, while the library wants to keep all versions available as part of the scholarly record and to satisfy citations to the earlier version. While this loss of absolute control can be difficult for librarians and archivists, it also allows them to re-engage with faculty around changes in how they function in an online, digital world, and what the library and archive need to do to support them.

The preceding discussion has hopefully made clear that institutional repositories are a moving target, that DSpace is a very young system, and that both will evolve rapidly in the coming decade. Already in the past year a research agenda has emerged that is clearly needed to make DSpace, and other systems in this domain, more sophisticated. This agenda includes research in the processes of digital preservation for a variety of digital formats (faculty do not constrain themselves to “trusted” formats like TIFF and XML), how to create virtual collections over distributed institutional repositories, digital publishing models. Research in storage architectures that scale to petabytes or more is needed. On the non-technical side, much more research is needed on the economics and incentive models of institutional repository operations, and the legal and regulatory constraints to information sharing that may hinder academics from optimal communications.

2004 should prove to be yet another year of rapid progress in this field. The number of institutions creating institutional repositories grows monthly, from a few dozen now to over a hundred in the next year (the process of creating such a repository is not quick – it seems to take about a year for most institutions to work through all the issues involved). For DSpace, we are hosting a first user group meeting in March, and are actively developing both a governance organization to help sustain and manage the system, and a group of like-minded institutions that want to collaborate on added-value services built on top of their repositories. Among the top priorities are increasing faculty awareness and use of institutional repositories, and educating faculty and public users of digital scholarship about copyright issues and the benefits of open access to scholarship. Beyond that we need to begin to define and build new value-added services layered over the DSpace Federation (i.e. the collective set of participating DSpaces), and to begin to define long term digital preservation strategies for the material we are collecting. There is much left to do, and no end of need for others to get involved.