PLANNING FOR INSTITUTIONAL REPOSITORIES: LESSONS LEARNED

Ann Devenish
EPrints Archive Coordinator and WHOI Publishing Services
MBLWHOI Library
Clark 131 – MS26
Woods Hole Oceanographic Institution
Woods Hole, MA 02543 USA

ABSTRACT: The Woods Hole Open Access Server (WHOAS) is an institutional repository (IR) for Woods Hole scientific content. It is an open archive with web access to its content. It is accessible to metadata harvesting (OAI-PMH) and allows for the free dissemination of scholarly communication. It is expected that WHOAS will eventually host a variety of digital objects including technical reports, articles, books, data sets, images, etc. It is hosted by the MBLWHOI Library to serve the multiple science institutions within the community of Woods Hole, Massachusetts. This paper discusses the pilot project that launched WHOAS.

KEYWORDS: MBLWHOI Library, institutional repository, digital archive, DSpace, Dublin Core Metadata, Handle System, Universal Resource Identifier, URI, CrossRef, Digital Object Identifier, DOI, Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH, Woods Hole Open Access Server, WHOAS.

In December 2003, the MBLWHOI Library began a pilot project to move forward the creation of an institutional repository for the scientific community of Woods Hole. The project was the culmination of several years' work on the part of the library director to mount an e-print archive.

Before one can run it is necessary to walk, and so the decision to move forward with a pilot project was both good and necessary. The goals of the pilot project were to learn what we needed to know about running an IR: to identify an appropriate software platform to host the IR; to identify an appropriate metadata scheme to describe the digital objects within the IR; to determine appropriate formats for delivery of the digital object to end users; and to design appropriate workflows for intake, processing, and delivery of the digital objects. Joint library staff members were organized into a working group, which in turn divided into task forces, each task force assuming responsibility for one project goal.

The deposit of Digital Object Identifiers (DOI) became an added objective of the pilot project soon after the project was underway. Conversations with CrossRef reflected an interest on their part to encourage DOI deposit by IRs and a joint venture was entered into.

The subject of the pilot project was *WHOI Technical Reports,* published in 2001-2003; approximately 35 items. These items offered compelling selection criteria. Major portions of the documents already existed as computer files and the files were easily accessible. Authors did not have to agree to participate as these items are owned by the Woods Hole Oceanographic Institution (WHOI).

Platform selection

Acting on the recommendations of the Platform task force, the platform selected for the pilot project was DSpace, version 1.1. DSpace is an open source software program developed at the Massachusetts Institute of Technology (M.I.T.) and Hewlett-Packard, and initially released in November 2002. In recommending the platform, the task force identified key points in its favor. It is affiliated with M.I.T., an institution with organized ties to Woods Hole and a potential resource for support. There is an active user group that meets across several electronic discussion groups and conducts an annual meeting. In designing DSpace, M.I.T. provided for large datasets and digital objects in multiple formats. The platform is focused on long-term preservation and uses the CNRI Handle System as a Universal Resource Identifier (URI). DSpace uses an organizational structure of Communities and Collections that satisfies the requirements for full administrative capabilities, and where certain policy decisions can be made at the departmental level. It uses a qualified version of Dublin Core Metadata.
Task force members: Ellen Levy, Maggie Rioux, Amy Stout.

Metadata

The Metadata task force selected Dublin Core (DC) as the most appropriate metadata scheme. Key points in favor of its selection included that DC was a simple interoperable and open standard, expressible in MARC, HTML or XML. It is easy to understand, flexible and widely used. Sample records for technical reports and data sets indicated it provided all the access points needed for resource description. And, DSpace was emerging as the platform of choice.
Task force members: Marisa Hudspeth, Robin Hurst, Lisa Raymond.

Delivery

In recommending PDF as the delivery option, the Delivery task force noted that its specifications have been formally published, making it suitable for a long-term preservation format. It renders each page exactly as the creator intended. PDF viewers are free and easily downloadable. DC Metadata can be programmatically embedded inside of the PDF file. The format supports electronic signatures, watermarks, password protection, and encryption as security features to protect a record against unauthorized alteration or viewing. It is backward compatible and one of the formats chosen by other leading institutions and Digitization and Electronic Archiving projects, such as the U. K. Public

Record Office, the Australian Victorian Electronic Records Strategy, and MIT's DSpace. And, it was already in use by WHOI Graphics for some of the subject Technical Reports.

The task force also considered compressed PDF files as a second, faster delivery option. However, estimates from Atypon Systems, Inc. to compress the files and add CrossRef linkages proved to be too expensive an endeavor for the pilot project.
Task force members: Marisa Hudspeth, Ellen Levy.

Workflows

WHOI Technical Reports are collated and distributed by the WHOI Graphics as print documents. Authors may create their reports in a variety of software programs, including Word and LaTeX. WHOI Graphics, however, will only accept submissions in Word, PDF, or as print documents; authors convert their LaTeX documents to PDF files. As necessary, WHOI Graphics converted all 35 Technical Reports to PDF and delivered them to library staff for intake into WHOAS.

The intake workflow is an integral part of the DSpace platform. Users are given rights ranging from reading content, submitting content including metadata, reviewing submissions, to editing metadata and accepting submissions into the IR. Once accepted into the IR, the platform assigns the Handle and the object becomes available.

Library staff submitted content using the DSpace intake form. The intake form provided key descriptive fields, including: author, title, series/report numbers, item type, subject, abstract, and funding. The form was modified to include additional fields, such as latitude/longitude and cruise dates. Entering text into the fields mapped data to the appropriate metadata elements. Submitted content was reviewed and edits to the metadata made as required, e.g., omissions, typographical errors, etc. Acceptance of the object into WHOAS generated an internal item ID and Handle, and created additional qualified metadata, including: date.accessioned, description.provenance, format.extent, and format.mimetype.

The deposit of the DOI at CrossRef required an additional (non-DSpace) workflow. A program was written to search each item's Handle prefix/suffix in the WHOAS database and export selected metadata in an XML file. Each file was then uploaded to CrossRef and the DOI deposited.

Task force members: Ann Devenish, Robin Hurst, Marisa Hudspeth, Ellen Levy, Maggie Rioux, Amy Stout.

Outstanding concerns

The completion of the pilot project on 30 June 2004 has left discussions and concerns that will likely continue as WHOAS goes forward. A few are noted here.

Unnoticed until content was being deposited into WHOAS, author metadata is stored as "contributor.author" not "creator.author." After review of metadata records at numerous other DSpace installations, we have elected to continue with "contributor.author" for the present. The new release of DSpace (1.2) maps author to the unqualified "creator" element when exporting DC metadata via OAI-PMH to fit community practices.

An open source platform means there is no dedicated technical support from a vendor; we are "dependent on the kindness of strangers." The DSpace user community has dedicated and committed participants who are willing to share their knowledge and experiences. The new version of DSpace has been released and loaded. The process of learning and adapting the platform begins anew.

Mission "creep" was evident in the pilot project. In the decision to deposit DOIs, CrossRef was and remains a committed partner to the venture. The impact on CrossRef may be significant as the Handle prefix/suffix, e.g., 1912/63, is not consistent with the ANSI/NISO Z39.84 syntax; DOI prefixes are to begin with "10." CrossRef has created an internal work-around to permit the venture to go forward and allow successful resolution of the Handle prefix/suffix to the metadata and the digital object through the DOI system, e.g., http://dx.doi.org/1912/63.

Funding for WHOAS will be an ongoing concern for the library. The pilot project was carried out within the current library budget, but at the time the budget was established, the pilot project was not planned for. Library staff members were not released from their routine responsibilities in order to participate in the pilot project. The sponsoring institutions of the library, the Marine Biological Laboratory (MBL) and WHOI, operate on cost center models. The acquisition of the PDF files from WHOI Graphics resulted in monetary charge-backs to the library; the services of the MBL IT staff involved with the working group also resulted in charge-backs to the library. The pilot project used existing desktop personal computers to load DSpace and the Handle server, and to host the digital objects. However, a more robust infrastructure will need to be acquired if the IR is to grow significantly beyond its initial 40+ objects.

2004 was a tight budget year for the sponsoring institutions and 2005 is looking much the same.

Lessons learned

- A pilot project is a great learning tool
- Involve library and IT colleagues, early and often
- Be flexible
- Be patient
- Be persistent
- Be prepared to un-do and re-do
- Expect the unexpected
- Look forward

For more information

CrossRef: http://www.crossref.org/01company/index.html
DOI Handbook: http://www.doi.org/hb.html
DSpace: http://www.dspace.org/
Dublin Core Metadata Initiative: http://dublincore.org/
Woods Hole Open Access Server (WHOAS): https://darchive.mblwhoilibrary.org/

Acknowledgements

The completion of the pilot project and the successful launch of the Woods Hole Open Access Server could not have happened without the combined efforts of the working group, the senior administrators of the MBLWHOI Library, the MBL IT staff, and our joint venture partner, CrossRef.

Members of the working group:
MBLWHOI Library: Ann Devenish (Project Manager), Marisa Hudspeth, Robin Hurst, Ellen Levy, Lisa Raymond, Maggie Rioux, and Amy Stout.
MBLWHOI Library Administration: Colleen Hurter, Cathy Norton, and Eleanor Uhlinger.
MBL IT: Chris Dematos and Pam Fournier.
CrossRef: Amy Brand and Chuck Koscher.