Introduction to the WSU Research Exchange service

WSU Research Exchange (URL http://research.wsulibs.wsu.edu/) is an institutional repository service that is designed to capture, preserve, and disseminate the intellectual output of researchers at Washington State University. The Research Exchange repository is based upon the DSpace open source repository software developed by the MIT Libraries and Hewlett-Packard Laboratories. Research Exchange has strong support for digital preservation through such features as durable identifiers and a Bitstream Format Registry that maps a file format to a declared level of support. One additional area of investigation, described in this session, is building support for resource harvesting by trusted repositories to support the preservation of Research Exchange content.

Existing use of Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH): Exposing digital collections

Since 2002, the Washington State University Libraries has used the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH) to expose information on its digital collections to repositories around the world. OAI-PMH was released as a specification in 2001; it is a simple but powerful protocol first designed to support coarse resource discovery for electronic pre-prints of journal articles. Its use broadened over time to support the discovery of digital content more generally. Examples of repositories that have harvested digital object metadata hosted on WSU Libraries’ servers include: UIUC’s Digital Gateway to Cultural Heritage Materials, the Western Waters Digital Library (hosted at the University of Utah), and the University of Michigan’s OAStor service. This metadata harvesting has dramatically increased the usage of digital collections hosted by the Libraries. OAI-PMH harvesting can also support the exposure of a department’s or faculty member’s research materials to a subject repository, which has the effect of making their research more prominent to colleagues working within their discipline.

Identified problem with OAI-PMH: Resource harvesting

At its core, OAI-PMH harvesting is based upon the retrieval of metadata, not upon resource retrieval. The most common metadata format used by DSpace repositories is Dublin Core, a simple metadata format consisting of 15 elements. XML schemas have been created for both simple Dublin Core and Qualified Dublin Core (the latter supporting refinements to elements).

In a foundational December 2004 article, Herbert Van de Sompel and his colleagues described two major problems with using OAI-PMH harvesting and Dublin Core metadata to support digital preservation:

- The Dublin Core Identifier element’s meaning may be ambiguous. Particularly, considering the DSpace example of a single repository item with several associated bitstreams/files, there is no clear way for the repository to communicate all relevant identifiers using Dublin Core. Additionally, the Dublin Core Identifier element could refer to an item’s file that needs to be harvested and preserved, or it could refer to an intermediary or proxy page which provides access to the file.
- The meaning of the datestamp contained in OAI-PMH records reflects the updating of the Dublin Core metadata, and not of the digital objects that need to be harvested by the trusted repository.

Van de Sompel and his colleagues conclude that “unqualified Dublin Core does not possess sufficient rigorous semantics to unambiguously express the information essential for resource harvesting” and also question the ability of Qualified Dublin Core to support effective resource harvesting due to the OAI-PMH datestamp issue described above. For these reasons, the use of a metadata format more expressive than Dublin Core, such as DIDL (Digital Item Declaration Language) or METS (Metadata Encoding and
Transmission Standard) is an important development path. As noted by Jeroen Bekaert and his colleagues in a 2003 *D-Lib Magazine* article, “digital libraries have reached a point where acceptable architectures must accommodate objects that aggregate datastreams of a wide variety of media-types, and must allow for the association of secondary data – including metadata supporting discovery, digital preservation and rights management – with those datastreams.”

**Enhanced metadata standards: MPEG21-DIDL and METS**

To resolve the two fundamental problems described above, enhanced metadata formats can be employed within an institutional repository system to effectively support digital preservation. MPEG-21 DIDL and METS are metadata standards that enable standardized, interoperable harvesting and mirroring of data, as well as ingesting objects, into repositories.

The MPEG-21 DIDL standard provides for consistent modification of all OAI-PMH elements, thus eliminating the need to modify separate OAI-PMH elements to facilitate the harvesting and exchange of accurate content. In addition, this standard enables trusted digital repositories to synchronize automatically with originating data repositories and preserve data transferred from one repository to another with minimal point-to-point interference. MPEG-21 DIDL is a standard native to DSpace repository software.

METS is a metadata schema that functions in ways similar to the Dublin Core standard to describe, administer, and organize metadata about objects residing in digital libraries. METS, however, provides an XML document format for encoding metadata necessary for both management of digital library objects within a repository, and exchange of such objects between repositories. In addition, METS provides for the exchange of objects between repositories and repository users. The Library of Congress maintains the METS standard.

**Summary**

Digital preservation support in the WSU Research Exchange service is being approached at several levels. The DSpace software employed for Research Exchange provides permanent identifiers using the CNRI handle system and has a Bitstream Format Registry that maps each file format to a declared level of support. Additionally, the Libraries is investigating the use of emerging, third-party digital preservation services, which are designed to provide preservation support and expertise to institutions hosting repositories.

The research and testing work described in this presentation is adding an important layer of support to the WSU Research Exchange service. Dublin Core metadata has been an effective tool in describing digital objects and in supporting the dissemination of Washington State University digital object information to a number of research repositories. This metadata, however, is not robust enough to support digital preservation through content mirroring with trusted repositories. For that reason, enhanced metadata formats, such as MPEG-21 DIDL and METS, should be employed in digital repositories.

The concept of mirroring to support the digital preservation task has been employed in a number of successful repository efforts including LOCCKS and repositories maintained by the nonprofit corporation Internet Archive. Supporting an enhanced metadata format within the WSU Research Exchange system enables the copying of content by a trusted, partner repository and provides important support in the digital preservation process.