DCPAPERS

Evolving Metadata Needs for an Institutional Repository: MIT's DSpace

Margret Branschofsky, Rebecca Lubas, MacKenzie Smith, Sarah Williams
MIT Libraries, Massachusetts Institute of Technology, USA
{margretb, rll, kenzie,sew}@mit.edu

Abstract

As the DSpace digital repository system develops, various metadata needs have emerged to accommodate the differing uses being made of the system. In the initial stages of the project a qualified form of the Dublin Core metadata set was developed for use within the system. Subsequently it became evident that metadata would also have to be imported from existing MARC sources for batch loads of scanned digital items. As the scope of DSpace content at MIT expanded to include the archiving of course web sites, the need for accommodating SCORM metadata became apparent. Interoperability requirements between like and unlike digital repositories has necessitated the creation of an OAI export format for metadata harvesters and a METS profile for exchange of metadata and disseminating information packages (DIPs) between DSpace federating institutions.
Keywords: *DSpace, SCORM, METS, Institutional Repository*

DSpace was originally conceived as an institutional repository system to capture, distribute and preserve the "born digital" products of research conducted at MIT. Although the "regular" path for metadata capture in DSpace is through the submission user interface, many items entering DSpace come with associated metadata already available. Part of maximizing the usefulness and efficiency of DSpace will be devising ways in which to utilize this pre-existing metadata.

Some of our early adopter communities have submitted batches of electronic items converted from print to electronic form, including a collection of pdf images of MIT Press out-of-print books and collections of scanned pdf images of print technical reports and working paper series from various labs, centers and schools on campus. MIT librarians had already cataloged most of the items in these collections, and therefore we had access to MARC format metadata records for them.

MARC records are created by professional library catalogers who follow a stringent set of rules, the Anglo-American Cataloging Rules. It would be out of the question to ask individual DSpace contributors to submit metadata of this complexity. However, since MARC records already existed for several collections we were ingesting, we decided it would be wise to use these high-quality metadata records. A batch importer was developed to import these MARC records and crosswalk them into the DSpace application of Dublin Core(DC), which is heavily based on the DC-Libraries group application profile proposal of 2001 (LAP).

The MIT metadata group developed the DSpace application of DC at the same time as its crosswalk from MARC to DC. Examining records created through MIT cataloging practice influenced some of the choices made to include particular qualifiers.

During this time the DCMI-Libraries Working Group was producing several drafts of the LAP. As a result of this simultaneous development, the DSpace application does not conform exactly to the latest version of the LAP.

One of the most controversial decisions made was the elimination of the use of the Creator element in favor of the use of Contributor. This decision was influenced by the first draft of the LAP, but also by the MARC practice of making a distinction between first author and succeeding ones. By mapping both 100 and 700 fields into Contributor, we were giving all authors equal billing. MIT's MARC to DSpace DC crosswalk can be viewed at ____.

MIT's OpenCourseWare (OCW) project, which makes available course materials from MIT classes to the world, will be placing its archives in DSpace. OCW courses and their associated resources will have metadata records using the Sharable Content Object Reference Model (SCORM) format, which employs a subset of the Institute of Electrical and Electronics Engineers' Learning Objects Metadata (IEEE LOM) for descriptive metadata.

The OCW objects will have high quality metadata. During the OCW publication process, faculty liaisons create preliminary metadata in

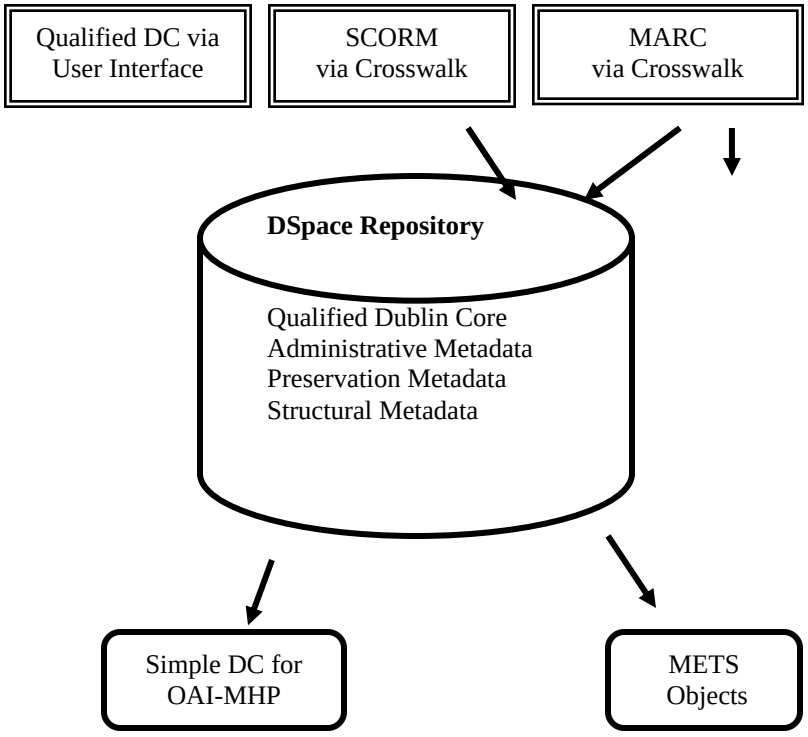
https://doi.org/10.23106/dcmi.952107531

consultation with the course creators. This metadata is reviewed and enhanced by MIT Libraries' Metadata Unit. The enhancement process includes quality control, creation of technical metadata, the addition of subject headings, and use of authoritative names for course creators and contributors.

The OCW items will represent a very large volume increase for DSpace. Current estimates show that there will be a total of 70,000 learning resource objects when OCW has a version of every MIT course online. This number will increase as new versions of courses are created. One way to make use of the existing metadata would be to crosswalk IEEE LOM to DSpace's qualified Dublin Core. This crosswalk effort is currently in progress to accommodate short term needs. Another way to deal with this significant amount of metadata would be to enable DSpace to accept metadata standards other than Dublin Core, as the SIMILE project is exploring over the longer term. (See http://web.mit.edu/simile/www/ for more information about SIMILIE.)

Interoperability between repositories is one of the aims of the DSpace project. To that end we are exporting simple, unqualified Dublin Core metadata records to allow harvesting of DSpace metadata through the Open Archives Initiative Metadata Harvesting Protocol (OAI-MHP).

It is expected that DSpace repositories will be operating at many institutions all over the world. One of the reasons for creating a DSpace federation is to facilitate the sharing of metadata and content across institutions. Examples of such sharing are the creation of virtual collections on specific topics consisting of metadata culled from each institution's repository, or the creation of overlay journals using metadata of items selected from various repositories. To facilitate these and other possible applications, MIT is developing a new METS profile for institutional repositories of heterogeneous content. We are developing an extension schema for our Dublin Core qualifiers and for technical metadata as implemented by multi-format repositories. When the profile is complete, DSpace will re-implement its export routines to support METS as an option, and will create an offline copy of the system's content with METS AIPs. We will also be investigating ingest of METS objects that conform to our profile, to support functions such as mirroring between institutions running the DSpace system.



https://doi.org/10.23106/dcmi.952107531