

Moving metadata: batch ingesting from Sirsi WorkFlows to the DSpace workspace

Ling He, York University Libraries

Keywords: metadata crosswalk, metadata reuse, software reuse, SWORD v2, DSpace ingestion, DSpace workspace, Sirsi WorkFlows

As libraries continue to use repositories to serve a variety of scholarly research dissemination and preservation purposes, interoperability, sharing, and reuse of data and metadata becomes more and more important. At York University, we found the need to develop a system to export data from our library catalogue system (Sirsi WorkFlows) for ingest into our DSpace repository (YorkSpace).

York University Libraries are an active member of the Sheet Music Consortium. Digitization efforts are focused on the extensive music collection of the late pianist John Arpin (1936-2007). Arpin, perhaps best known as a Ragtime musician, built an extensive collection of sheet music (approximately 150,000 items) throughout his career as a performer, composer and music director. The collection includes examples of Canadian, Broadway, American Standard, Pop music, Jazz and Ragtime from the late nineteenth century to the present day.

The digital project team spans multiple departments and consists of a part time cataloging librarian, and several part time student digital project assistants. Our original workflow was inefficient as both a catalogue record and DSpace record were manually created at different stages of the description and digitization process. At the point of digitization, a DSpace record was created, where limited descriptive metadata was pasted in from the recently created catalogue record. This introduced the potential for error as well as added extra steps to the workflow. The revised workflow frees the digitization staff from having to re-enter catalogue metadata into the repository.

This presentation describes the author's efforts to reuse the metadata catalogued in cataloguing system to DSpace. Currently, there is no efficient tool to batch ingest metadata into the DSpace "workspace" that enables users to edit submissions before they are published. A workaround is to push the items into submissions workflow, however, extra steps are required to approve the tasks and accept or reject the item. This solution makes use of one of the SWORD v2 characteristics, where SWORD uses the In-Progress HTTP header to denote whether the item should be published upon ingest. This enables the item to be deposited into DSpace workspace and queued for final editing.

The author reused and customized two open source programs to fulfill the automation process. The MARC4j program was reused to covert MARC format to MARCXML format. The Library of Congress MARCXML to DC Stylesheet was reused and modified to convert MARCXML format to DSpace qualified DC format. The PHP SWORD v2 client library was reused and a command line tool and a web application were developed based on this library to generate atom packages and batch ingest the packages into the DSpace workspace using the SWORD v2 protocol.