Expanding Metadata Reuse with an Islandora Metadata Extraction Utility

Serhiy Polyakov, William E. Moen
University of North Texas

Abstract

This paper describes the prototype of a metadata extraction utility implemented in the Islandora-based repository. The paper also proposes a particular workflow that takes advantage of the extraction utility. The workflow allows for the managing of scholarly objects (e.g., journal articles and other materials) at the various stages of their lifecycle. In addition to the extraction utility, the paper recommends standards and available open source metadata management tools that are used in the workflow. The recommended existing metadata management tools allow the end users (e.g., scholars, authors) to create or fetch descriptive metadata using external services and to embed metadata into the scholarly objects. The embedded metadata then travels with the scholarly objects. When an object is submitted to the repository, metadata is automatically extracted and added to the auxiliary index and may be added to the repository metadata datastreams with the implemented utility, ensuring effective reuse of the metadata.

Background

Institutional databases and repositories for scholarly publications have the potential to be efficient means for dissemination of research output. Development of these repositories is often among the goals of the research strategic plans and the open access policies adopted in universities and research institutions (University of North Texas Faculty Senate, 2011; University of Prince Edward Island Senate, 2008, 2012).

Scholars are encouraged to routinely provide a copy of the materials that represent their scholarly output to be deposited in open access repositories. Such materials can be articles published in scholarly journals and conference proceedings, research data, conference presentations, and posters. The materials may be in the form of an author's final post-peer review manuscripts, preprints, or (where permitted by the publisher) publisher's PDFs (University of North Texas Faculty Senate, 2011).

Problem

The process of submitting scholarly objects to repositories can include providing the content file(s), assigning metadata, and depositing the object. The parties involved in these processes may include scholars, authors, repository managers, or catalogers. In the submission workflow, the source of the metadata is normally the objects' content or secondary sources such as bibliographic databases. The submission workflow that is normally used in scholarly repositories may be enhanced in a way that will reduce cost and/or time required for depositing by increasing the reuse of metadata.

It would be beneficial if scholarly objects that represent research output were always accompanied by metadata in a form that is easy to manage by the end users and automatically readable by the repositories or other systems such as reference management software.

Objects and Limitations

A single file in PDF format is the most common form for storing and disseminating the content of a scholarly object. For example, a content file of a journal article may be stored in a repository, in a commercial full-text database, or in a folder on a hard drive; it may be linked to the author's web page; or it may be disseminated as an email attachment. The proposed extraction utility is designed for use with these types of objects comprised of a single PDF content file. Multiple files in various formats may be handled if added to a single PDF portfolio file.

However, in other instances scholarly objects may comprise multiple files in PDF or other formats, or may be represented only by citation information (metadata) when content files are not available for immediate access and dissemination (e.g., because of various technical or licensing reasons). Enhancements suggested in this paper are only partially applicable to these types of objects. At the same time, these types of objects may be

managed using traditional workflow. For example, these objects can be deposited into the repository and metadata can be added using the submission form.

Proposed Solution

As part of the proposed workflow, we recommend a set of existing tools that aid in the management and reuse of metadata in different stages of the scholarly objects' lifecycle. In addition to the existing tools, we propose a metadata extraction utility implemented in an Islandora-based repository. The capabilities of this utility in conjunction with other elements such as reference management software and associated standards and services provide all of the necessary components for the workflow. The proposed workflow includes the steps of metadata creation or fetching from the external sources, metadata editing, embedding of metadata into the PDF files, metadata extraction by the repository on submission, and semi-automatic addition to the Fedora Commons metadata datastreams (see Figure 1).

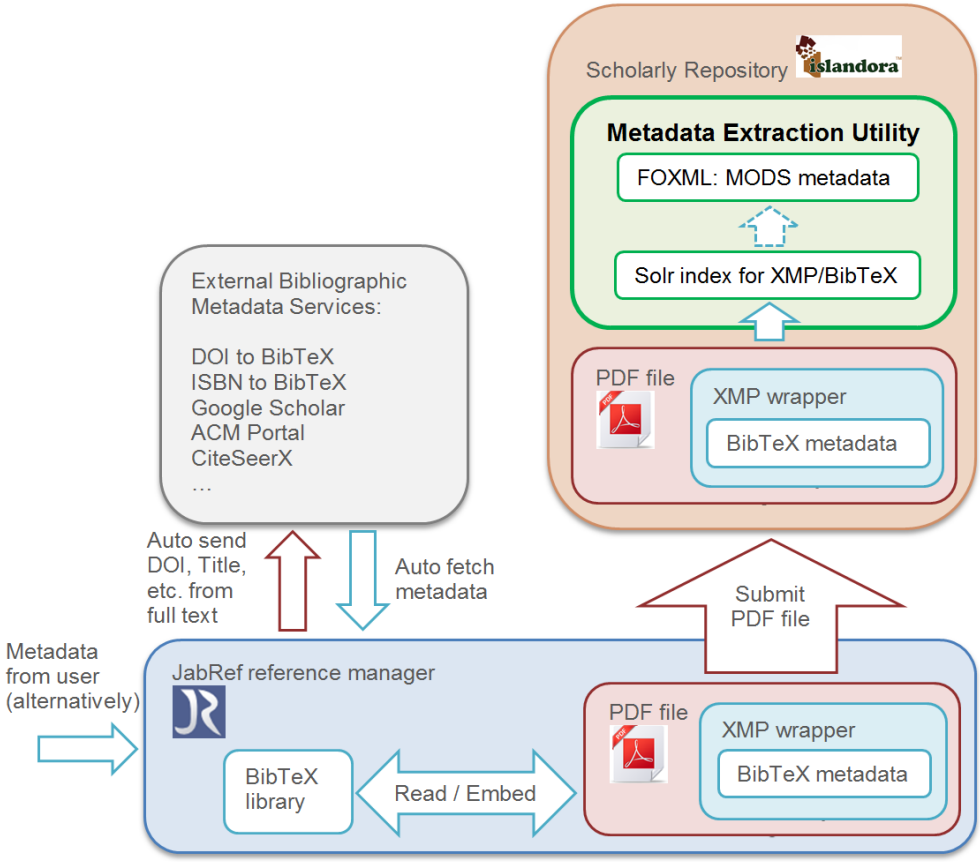

Figure 1. The workflow of fetching, embedding, extracting, and reuse of the descriptive metadata in a scholarly repository. Note: components in solid lines are either developed or recommended and tested; those in dashed lines are under development.

Technical Details

A number of standards, for example, Metadata Encoding and Transmission Standard (METS), DSpace's Simple Archive Format, and Fedora Object XML (FOXML), are used in the repositories and may support tasks of describing the metadata and associated content files. Metadata and content described with the use of these standards are stored separately, and when objects are exported from the repositories using these standards, metadata and content are stored in separate files. These standards are not recommended as appropriate for the end users in their routine management of the scholarly objects because of the absence of the software supporting the standards on the end user side and because of the complexity of the standards.

BibTeX is another format that is commonly used for the storing and exchanging of metadata between bibliographic databases, repositories, reference management software, and word processors. This format is less complex and also stores metadata in separate files called libraries. Libraries may accompany the content files and include pointers to the content files.

Scholars commonly use reference management software applications to manage their scholarly objects' content files and metadata. Most of these applications either use BibTeX as a native format or support import/export using this format.

JabRef is the only reference management software that has the capabilities of embedding and reading BibTeX metadata using the Extensible Metadata Platform (XMP) standard (http://jabref.sourceforge.net) as a wrapper that stores BibTeX metadata. XMP was originally developed by Adobe Systems Inc. and become an ISO standard (International Organization for Standardization, 2012). Content in XMP travels with the file and can be embedded in many common file formats including PDF (PDFlib, 2013). Additionally, JabRef software includes powerful features that allow the fetching of metadata from external services using the content of the PDF file of the object. Some of the supported services are DOI to BibTeX, ISBN to BibTeX, Google Scholar, ACM Portal, and CiteSeerX. For example, JabRef's DOI to BibTeX fetcher uses a service located at http://dx.doi.org.

The first step of the workflow is obtaining the PDF file of the scholarly object, which at that point may simply reside on the hard drive of end user's computer. When the PDF file of a scholarly object is dropped (through drag and drop via the user interface) in JabRef, the software automatically reads elements of its content such as the title or the DOI on the first page of the document and retrieves values of the metadata elements from the external services. Retrieved descriptive metadata values are copied to the BibTeX library file. If fetching fails or provides incorrect values, the values can be manually added or edited by the end user though the JabRef user interface. After the values are created, a user may select the JabRef option to write the BibTeX entry as XMP metadata to PDF. Similarly, if the PDF already contains XMP metadata, the values may be extracted when the PDF document is added to the JabRef library. The existing metadata in a PDF can also be enhanced with additional external or locally edited metadata.

In the last step of the workflow, we have implemented a prototype of the extraction utility for a repository built on the Fedora Commons, Drupal, and Islandora platforms. This service supports automatic extraction of the BibTeX metadata embedded in the PDF files using XMP standard. At this time, the component of the utility that extracts BibTeX metadata from the PDF files submitted to the repository index is being developed and tested. We are working on the part that will copy metadata from the index to the MODS schema. The utility is based on the previous work on developing a generic extraction utility that enhances indexing (Polyakov, 2012). The core components of the utility are Fedora Generic Search, Apache Tika, Apache Solr, and a set of associated stylesheets and scripts.

Conclusion

The paper proposes a metadata extraction utility that together with other existing tools and standards makes it possible to construct an efficient workflow for metadata management. The goal of the workflow is to bring metadata creation closer to the end users, ensuring that descriptive metadata is an integral part of the scholarly objects and automatically reusable in the scholarly repositories. Future work includes testing and refining other components of the workflow such as components that copy values from the index to the MODS schema. We also would like to recommend to the developers of the reference manager software that they add metadata embedding tools to their products.

References

International Organization for Standardization. (2012). ISO 16684-1:2012: Graphic technology—Extensible metadata platform (XMP) specification—Part 1: Data model, serialization and core properties. Retrieved from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57421

PDFlib. (2013). XMP metadata. Retrieved from http://www.pdflib.com/knowledge-base/xmp-metadata

Polyakov, S. (2012, May). *Enhancing a digital repository with objects' embedded metadata*. Poster session presented at the Texas Conference on Digital Libraries (TCDL 2012), Austin, TX. Retrieved from https://conferences.tdl.org/TCDL/TCDL2012/paper/view/540

University of North Texas Faculty Senate. (2011). *Policy on open access to scholarly works*. Retrieved from http://openaccess.unt.edu/sites/default/files/03-11/OpenAccessPolicy_UNTFacultySenateApproved_9Mar2011_.pdf

University of Prince Edward Island Senate. (2008). *Strategic research plan 2008-2018*. Retrieved from http://research.upei.ca/files/research/v9 Senate 22Apr08.pdf

University of Prince Edward Island Senate. (2012). *Policy: Open access and dissemination of research output*. Retrieved from https://cab.upei.ca/sites/default/files/attachments/OpenAccessandDisseminationofResearchOutput.pdf