Expanding Metadata Reuse with an Islandora Metadata Extraction Utility

Serhiy Polyakov and William E. Moen
University of North Texas

International conference Open Repositories 2013
Charlottetown, Prince Edward Island, Canada

Outline

• Background
• Problem
• Types of objects and limitations
• Proposed solution
• Technical details
• The utility and workflow walkthrough
Background (1/2)

Islandora-based repository

Metadata reuse

Reference Manager Software, e.g.:
• Mendelay
• RefWorks
• Qiqqa (+ research manager and mind maps)
• JabRef
• Docear (academic literature suite)
• Zotero
• EndNote
Background (2/2)

Scholars use Reference Management Software for managing:
• their own research outputs
• publications/sources they use in research
• sets of articles for Metadata and Information Retrieval experiments (specific to our research)
• ...

At the same time:
• scholars are encouraged to routinely deposit their scholarly outputs into open access repositories
• in our research we also need to deposit larger sets of articles and use the repository for information retrieval experiments
Problem

• The workflow of submitting scholarly objects to repositories can include providing the content files, assigning metadata, and depositing the objects.

• It would be beneficial if scholarly objects that represent research outputs were always accompanied by embedded metadata in a form that is easy to manage by the end users (e.g., scholars, authors) and automatically readable by the repositories or other systems such as reference management software.
Types of objects and limitations

The utility is designed for use with objects comprising:

- a single file in PDF format (the most common form for storing and disseminating the content of a scholarly output)
- PDF portfolio file

PDF or PDF portfolio files are normally:

- stored in a folder on a hard drive of the researcher’s computer
- stored in a reference manager software
- stored on a web server and linked to the author’s web page
- disseminated as an email attachment
- stored in a repository
Proposed utility and workflow
Technical details (1/4)

Embedded metadata can be extracted for indexing in an Islandora-based repository. The components of a repository that are directly involved in this process are:

- Fedora Generic Search Service
- Apache Tika (content analysis toolkit)
- Apache Solr (search platform)

However, embedding and extraction have been previously used primarily for technical metadata.
How to embed descriptive metadata into PDF content files on a users’ (e.g., scholars, authors) side?

We tested a number of reference management software:

• Mendelay
• RefWorks
• Qiqqa (+ research manager / mind maps)
• JabRef
• Docear (academic literature suite)
JabRef is the only reference management software that has the capabilities of embedding and reading metadata into PDF files using BibTeX format and the Extensible Metadata Platform (XMP) standard.

XMP was originally developed by Adobe Systems Inc. and became an ISO standard.

BibTeX format stores metadata in separate files called libraries.

Most of the reference management software either use BibTeX as a native format or support import/export using this format.
Additionally, JabRef software includes powerful features that allow the fetching of metadata from the external services using the content of a PDF file:

- DOI to BibTeX (http://dx.doi.org)
- ISBN to BibTeX
- Google Scholar
- ACM Portal
- CiteSeerX
Workflow walkthrough (1/12)

Sample file of an article residing on a researcher's computer
Workflow walkthrough (2/12)

Content of the file shown in a PDF viewer
Workflow walkthrough (3/12)

File properties (basic embedded metadata) shown in a PDF viewer

PDF embedded descriptive metadata is often missing, incorrect, or incomplete.
Drag and drop the file into JabRef
Workflow walkthrough (5/12)

JabRef provides options for metadata generation (including automatic and manual).
Workflow walkthrough (6/12)

Metadata is fetched using DOI to BibTeX and embedded into the PDF file with the Write XMP button. Metadata can be also added manually.
Workflow walkthrough (7/12)

Rich descriptive metadata is now embedded into the PDF file.

Original file

- http://ns.adobe.com/pdf/1.3/
- http://ns.adobe.com/xap/1.0/
- http://purl.org/dc/elements/1.1/
  - dcf:format: application/pdf
  - dcf:creator (seq)
    [1]: lcp
  - dcf:title (seq)
    [1]: Learning author-topic models from text corpora
- http://ns.adobe.com/xap/1.0/mm/
- http://ns.adobe.com/photoshop/1.0/
- http://ns.adobe.com/png/1.0/
- http://ns.adobe.com/tiff/1.0/

After embedding

- http://ns.adobe.com/pdf/1.3/
- http://ns.adobe.com/xap/1.0/
- http://ns.adobe.com/xap/1.0/mm/
- http://purl.org/dc/elements/1.1/
  - bib:author (seq)
    [1]: Michal Rosen-Zvi
    [2]: Chaitanya Chemudugunta
    [3]: Thomas Griffiths
    [4]: Padhraic Smyth
    [5]: Mark Steyvers
  - bib:doi: 10.1145/1658377.1658381
  - bib:file: rosen-zvi2010learning - Learning author-topic models from text
  - bib:journal: ACM Transactions on Information Systems
  - bib:month: Jan
  - bib:year: 2010
  - bib:volume: 28
  - bib:pages: 1-38
  - bib:publisher: Association for Computing Machinery
  - bib:title: Learning author-topic models from text corpora
  - bib:url: http://dx.doi.org/10.1145/1658377.1658381
  - bib:enttype: Article
- http://ns.adobe.com/photoshop/1.0/
- http://ns.adobe.com/png/1.0/
- http://ns.adobe.com/tiff/1.0/
Workflow walkthrough (8/12)

Repository step 1. On the submission form, enter a few characters into the title field, attach the PDF file, and submit.
Workflow walkthrough (9/12)

Embedded descriptive metadata is extracted with Apache Tika on submission and sent to the pre-configured Solr index.

fedoragsearch.daily.log

DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/pages value=1-38
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/journal value=ACM Transactions on Information Systems
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/bibtexkey value=rosen-zvi2010learning
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/doi value=10.1145/1658377.1658381
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/month value=Jan
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/entrytype value=Article
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/volume value=28
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/url value=http://dx.doi.org/10.1145/1658377.1658381
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/number value=1
DEBUG 2013-07-02 00:32:06,307 (TransformerToText) METADATA name=bibtex/file value=:rosen-zvi2010learning - Learning author-topic models from text corpora.pdf:PDF
DEBUG 2013-07-02 0:32:06,307 (TransformerToText) METADATA name=bibtex/year value=2010

...
Submitted item view

<table>
<thead>
<tr>
<th>Structural Metadata</th>
<th>Sample Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Member of this Collection</td>
<td></td>
</tr>
</tbody>
</table>

User Supplied Item Metadata
(display from MODS)

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Learning...</td>
</tr>
<tr>
<td>Name(s)</td>
<td></td>
</tr>
<tr>
<td>Journal</td>
<td></td>
</tr>
<tr>
<td>Date Issued</td>
<td></td>
</tr>
<tr>
<td>Volume</td>
<td></td>
</tr>
<tr>
<td>Issue</td>
<td></td>
</tr>
<tr>
<td>Pages</td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td></td>
</tr>
</tbody>
</table>

Auto Generated Item Metadata
(display from Solr index)

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Learning author-topic models from text corpora</td>
</tr>
<tr>
<td>Name(s)</td>
<td>Rosen Zhu, Michal Chennuru, Chaitanya Griffiths, Thomas Smyth, Padhraic Steyvers, Mark</td>
</tr>
<tr>
<td>Journal</td>
<td>ACM Transactions on Information Systems</td>
</tr>
<tr>
<td>Date Issued</td>
<td>2010</td>
</tr>
<tr>
<td>Volume</td>
<td>28</td>
</tr>
<tr>
<td>Issue</td>
<td>1</td>
</tr>
<tr>
<td>Pages</td>
<td>1-38</td>
</tr>
<tr>
<td>Producer</td>
<td>Acrobat Distiller 7.0 for Macintosh</td>
</tr>
<tr>
<td>Keywords/Tags</td>
<td>Gibbs sampling, Topic models, author models, perplexity, unsupervised learning</td>
</tr>
<tr>
<td>Content Type</td>
<td>application/pdf</td>
</tr>
<tr>
<td>Number of Pages</td>
<td>38</td>
</tr>
<tr>
<td>Creation Date</td>
<td>2010-01-19, 05:45:31am CST</td>
</tr>
</tbody>
</table>
Workflow walkthrough (11/12)

Repository step 2. Edit the submitted item. Click "Get" and all values will be copied into the form fields.

Populates with values extracted from PDF and stored in Solr index.
Workflow walkthrough (12/12)

Metadata has now been copied into the MODS datastream.
Proposed utility and workflow revisited
Bibliography