

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

Paper presented at the Fedora User Group session, July 12th, 2013

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

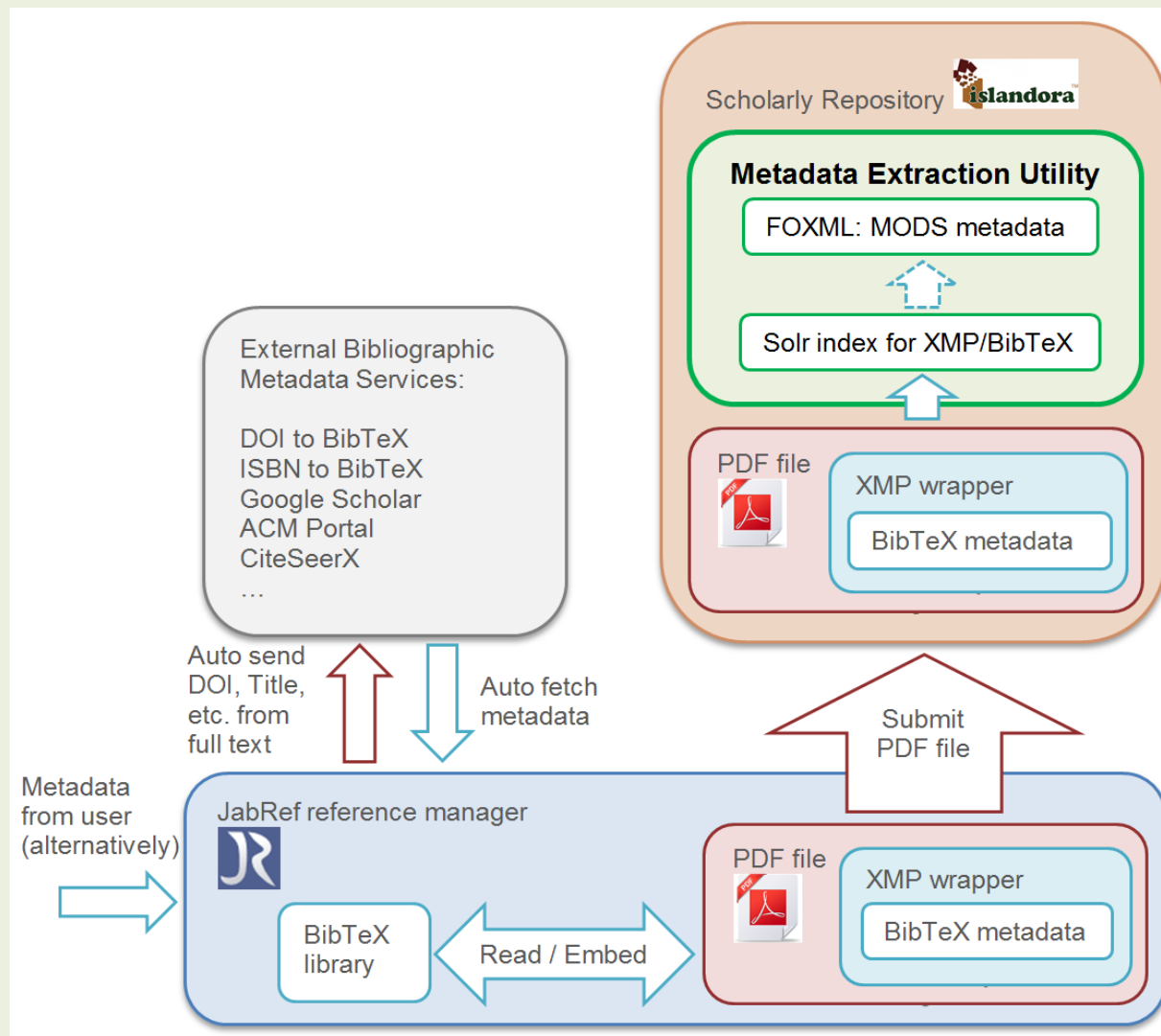
The utility is designed for use with objects comprising:

- 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.

Technical details (2/4)

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)

Technical details ^(3/4)

- 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 become 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.

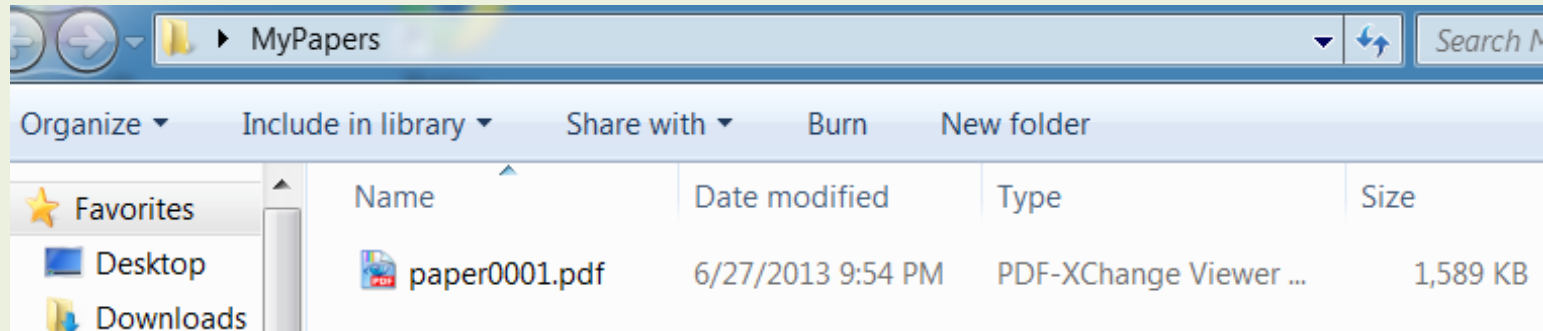
Technical details (4/4)

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

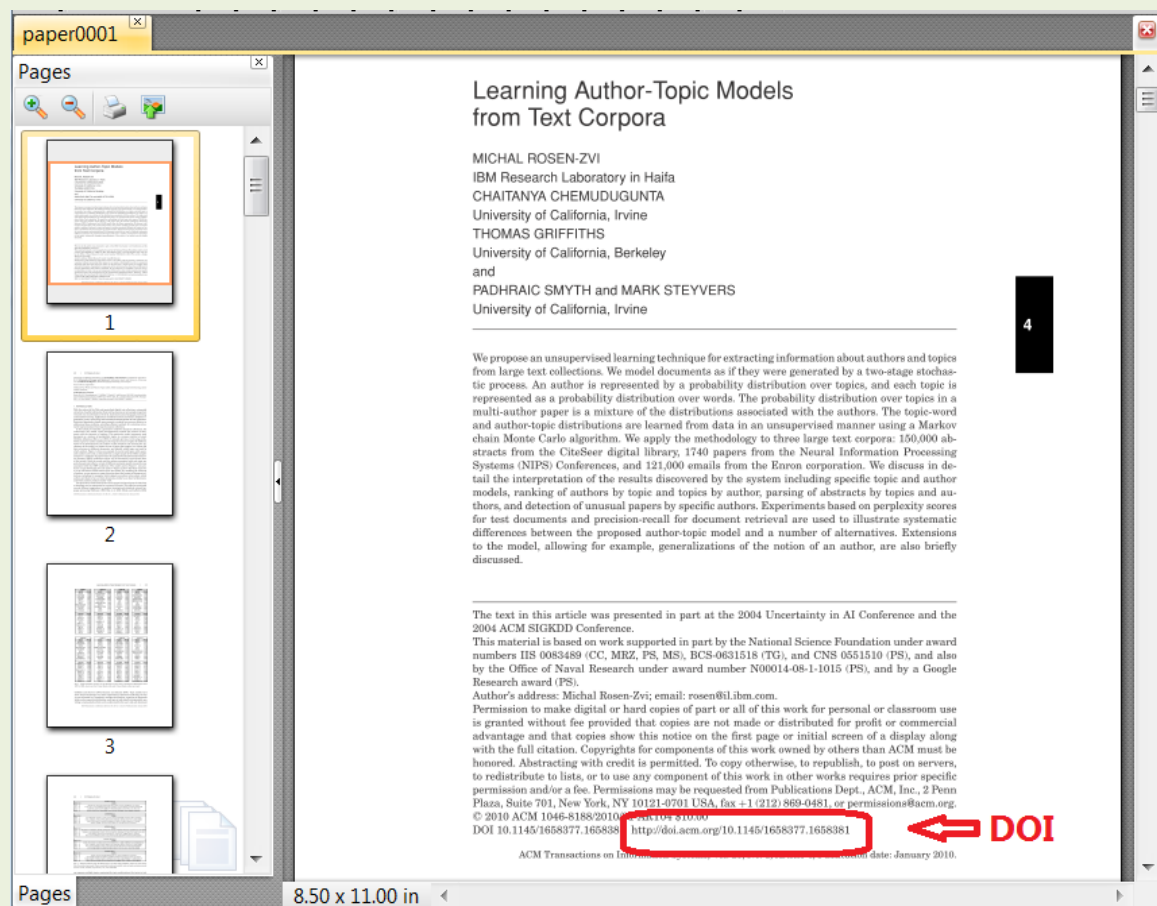
Workflow walkthrough (1/12)

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

File Info

File Name: paper0001.pdf
Location: G:_RESEARCH\Conferences\2013_OR\...\Sample_Paper\ 
Created: 1/19/2010, 5:45:31 AM
Modified: 1/22/2010, 8:39:34 AM
File Size: 1,626,458 bytes (1.55 MB)

Document Info

Title: Learning author-topic models from text corpora
Author: lop
Subject:
Keywords: Gibbs sampling, Topic models, author models, perplexity, unsupervised learning

PDF Producer: Acrobat Distiller 7.0 for Macintosh
Application: Textures®: LaserWriter 8 8.7.3
PDF Version: 1.3
Pages Count: 38
PDF Viewer: <Unknown>

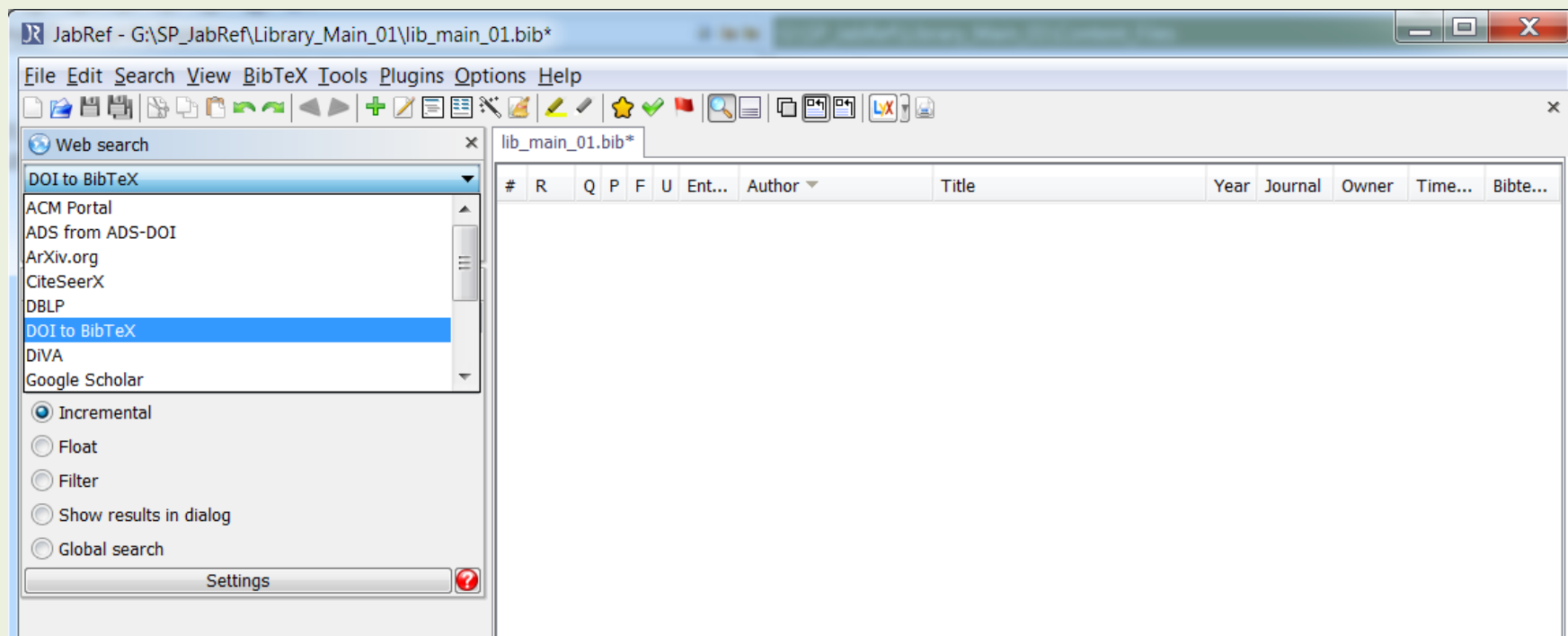
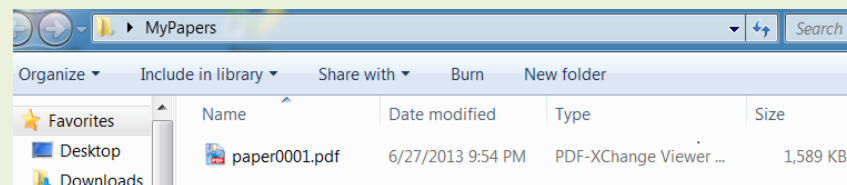
Additional Metadata...

PDF embedded descriptive metadata is often missing, incorrect, or incomplete.

<http://ns.adobe.com/pdf/1.3/>
<http://ns.adobe.com/xap/1.0/>
<http://purl.org/dc/elements/1.1/>
 dc:format: application/pdf
 dc:creator (seq)
 [1]: lop
 dc: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/>

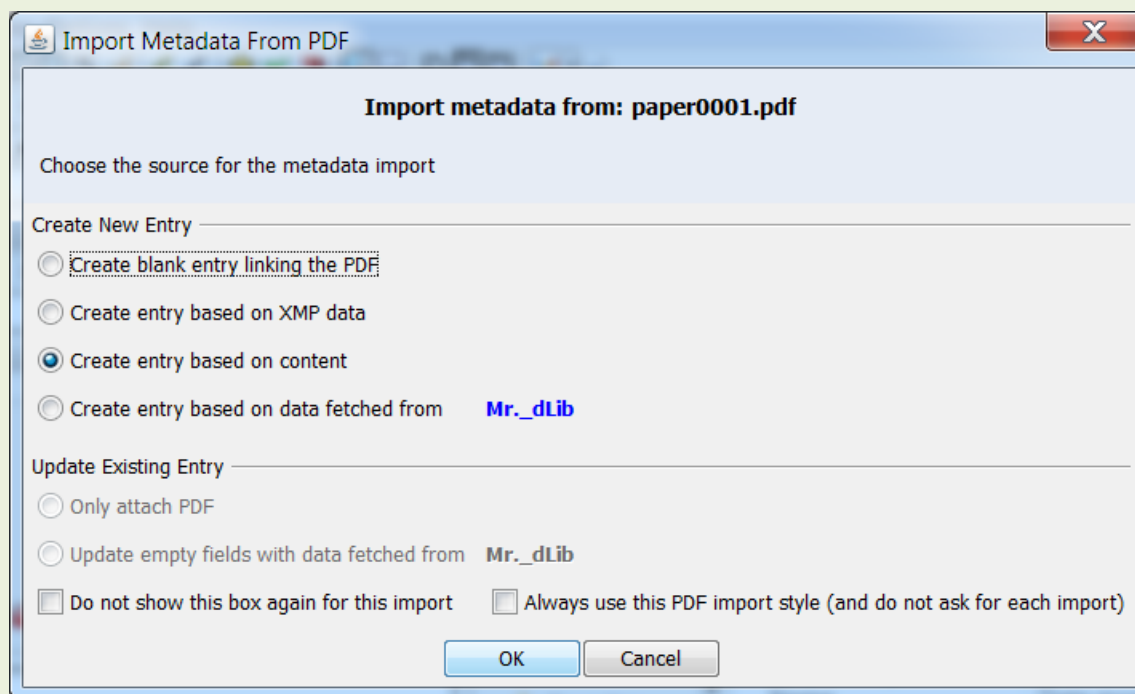
Workflow walkthrough (4/12)

Drag and drop the file into JabRef



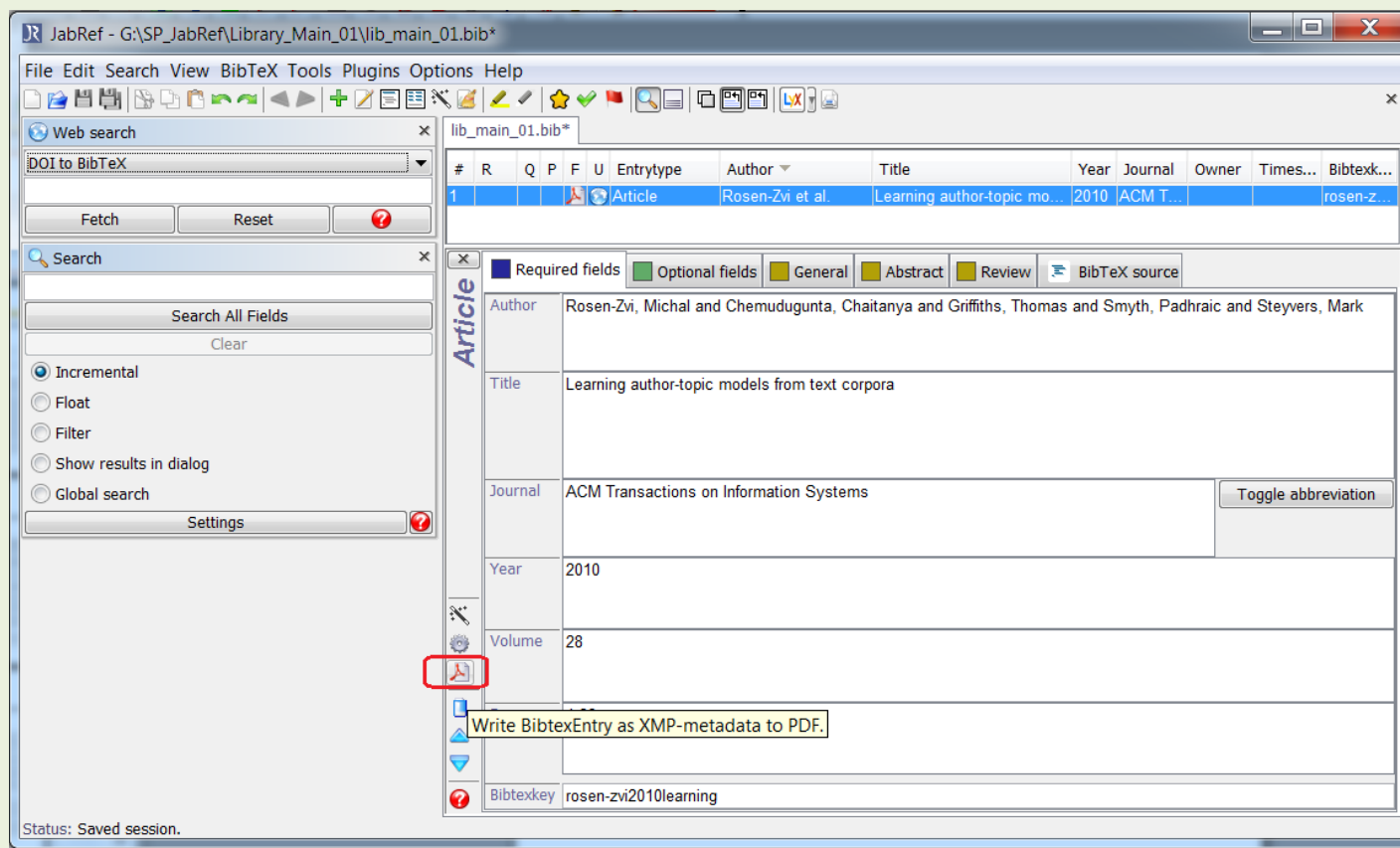
Workflow walkthrough (5/12)

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

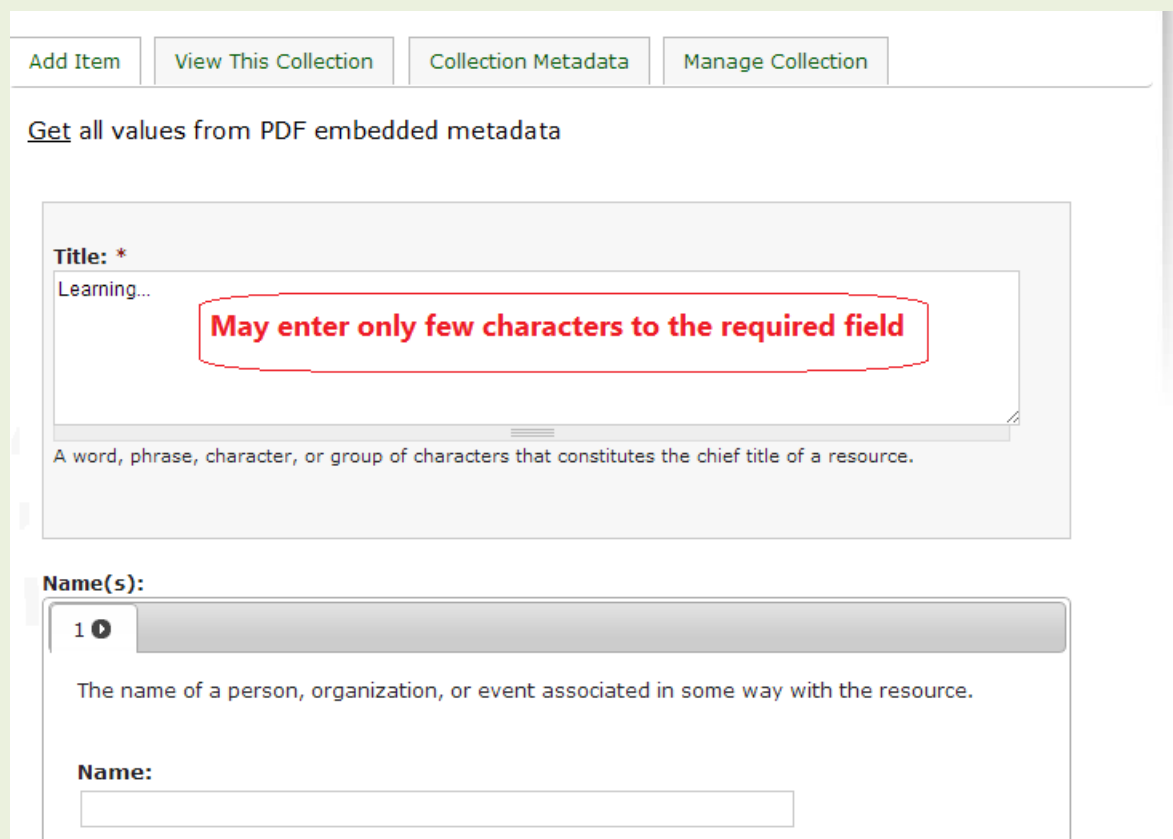
```
⊞ http://ns.adobe.com/pdf/1.3/
⊞ http://ns.adobe.com/xap/1.0/
⊞ http://purl.org/dc/elements/1.1/
  dc:format: application/pdf
  ⊞ dc:creator (seq)
    [1]: lop
  ⊞ dc:title (seq)
    [1]: Learning author-topic models from text corpora
⊞ http://ns.adobe.com/xap/1.0/mm/
⊞ http://ns.adobe.com/photoshop/1.0/
⊞ 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/
⊞ http://jabref.sourceforge.net/bibtexXMP/
  ⊞ bibtex:author (seq)
    [1]: Michal Rosen-Zvi
    [2]: Chaitanya Chemudugunta
    [3]: Thomas Griffiths
    [4]: Padhraic Smyth
    [5]: Mark Steyvers
  bibtex:bibtexkey: rosen-zvi2010learning
  bibtex:doi: 10.1145/1658377.1658381
  bibtex:file: :rosen-zvi2010learning - Learning author-topic models from text
  bibtex:journal: ACM Transactions on Information Systems
  bibtex:month: Jan
  bibtex:number: 1
  bibtex:pages: 1-38
  bibtex:publisher: Association for Computing Machinery
  bibtex:title: Learning author-topic models from text corpora
  bibtex:url: http://dx.doi.org/10.1145/1658377.1658381
  bibtex:volume: 28
  bibtex:year: 2010
  bibtex:entrytype: 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.



The screenshot shows a web interface for a repository. At the top, there are four buttons: "Add Item", "View This Collection", "Collection Metadata", and "Manage Collection". Below these buttons, the text "Get all values from PDF embedded metadata" is displayed. The main form area contains two sections. The first section is titled "Title: *" and has a text input field with the placeholder text "Learning...". A red rounded rectangle is drawn around the input field, containing the text "May enter only few characters to the required field". Below the input field, there is a description: "A word, phrase, character, or group of characters that constitutes the chief title of a resource." The second section is titled "Name(s):" and has a text input field with the placeholder text "1". Below the input field, there is a description: "The name of a person, organization, or event associated in some way with the resource." At the bottom of the form, there is a section titled "Name:" with a text input field.

[Add Item](#) [View This Collection](#) [Collection Metadata](#) [Manage Collection](#)

Get all values from PDF embedded metadata

Title: *
Learning...

May enter only few characters to the required field

A word, phrase, character, or group of characters that constitutes the chief title of a resource.

Name(s):

1

The name of a person, organization, or event associated in some way with the resource.

Name:

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
...
```



Structural Metadata

Is Member of this Collection [Sample Articles](#)

Submitted item view

[View or Download File](#)

User Supplied Item Metadata (display from MODS)

Title	Learning...
Name(s)	
Journal	
Date Issued	
Volume	
Issue	
Pages	
Abstract	

Auto Generated Item Metadata

View

File (OBJ) Embedded Metadata (display fom Solr index)

Title	Learning author-topic models from text corpora
Name(s)	Rosen-Zvi, Michal Chemudugunta, Chaitanya Griffiths, Thomas Smyth, Padhraic Stevvers, Mark
Journal	ACM Transactions on Information Systems
Date Issued	2010
Volume	28
Issue	1
Pages	1-38
Producer	Acrobat Distiller 7.0 for Macintosh
Keywords/Tags	Gibbs sampling, Topic models, author models, perplexity, unsupervised learning
Content Type	application/pdf
Number of Pages	38
Creation Date	2010-01-19, 05:45:31am CST

Workflow walkthrough (11/12)

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

The screenshot shows a web interface for editing metadata. At the top, there are four buttons: "Add Item", "View This Collection", "Collection Metadata", and "Manage Collection". Below these buttons, the text "Get all values from PDF embedded metadata" is displayed, with the word "Get" enclosed in a red square. The main form area contains two sections. The first section is titled "Title: *" and contains a text input field with the placeholder text "Learning...". Below the input field is a description: "A word, phrase, character, or group of characters that constitutes the chief title of a resource." A red arrow points from the right side of the form to this section. The second section is titled "Name(s):" and contains a list box with one item, "1", and a description: "The name of a person, organization, or event associated in some way with the resource." Below this is a "Name:" label and an empty text input field. A red arrow points from the right side of the form to this section. To the right of the form, there is a red text annotation: "Populates with values extracted from PDF and stored in Solr index".

Get all values from PDF embedded metadata

Title: *
Learning...

A word, phrase, character, or group of characters that constitutes the chief title of a resource.

Name(s):
1

The name of a person, organization, or event associated in some way with the resource.

Name:

Populates with values extracted from PDF and stored in Solr index

Workflow walkthrough (12/12)

Metadata has now been copied into the MODS datastream.

View Metadata and Manage Files | Edit Metadata

Structural Metadata

Is Member of this Collection [Sample Articles](#)

Submitted item view

[View or Download File](#)

User Supplied Item Metadata (display from MODS)

Title	Learning...
Name(s)	
Journal	
Date Issued	
Volume	
Issue	
Pages	
Abstract	

Auto Generated Item Metadata

View

File (OBJ) Embedded Metadata (display from Solr index)

Title	Learning author-topic models from text corpora
Name(s)	Rosen-Zvi, Michal Chemudugunta, Chaitanya Griffiths, Thomas Smyth, Padhraic Stevvers, Mark
Journal	ACM Transactions on Information Systems
Date Issued	2010
Volume	28
Issue	1
Pages	1-38
Producer	Acrobat Distiller 7.0 for Macintosh
Keywords/Tags	Gibbs sampling, Topic models, author models, perplexity, unsupervised learning
Content Type	application/pdf
Number of Pages	38
Creation Date	2010-01-19, 05:45:31am CST

View Metadata and Manage Files | Edit Metadata

Structural Metadata

Is Member of this Collection [Sample Articles](#)

[View or Download File](#)

User Supplied Item Metadata

Title	Learning author-topic models from text corpora
Name(s)	Rosen-Zvi, Michal Chemudugunta, Chaitanya Griffiths, Thomas Smyth, Padhraic Stevvers, Mark
Journal	ACM Transactions on Information Systems
Date Issued	2010
Volume	28
Issue	1
Pages	1-38
Abstract	

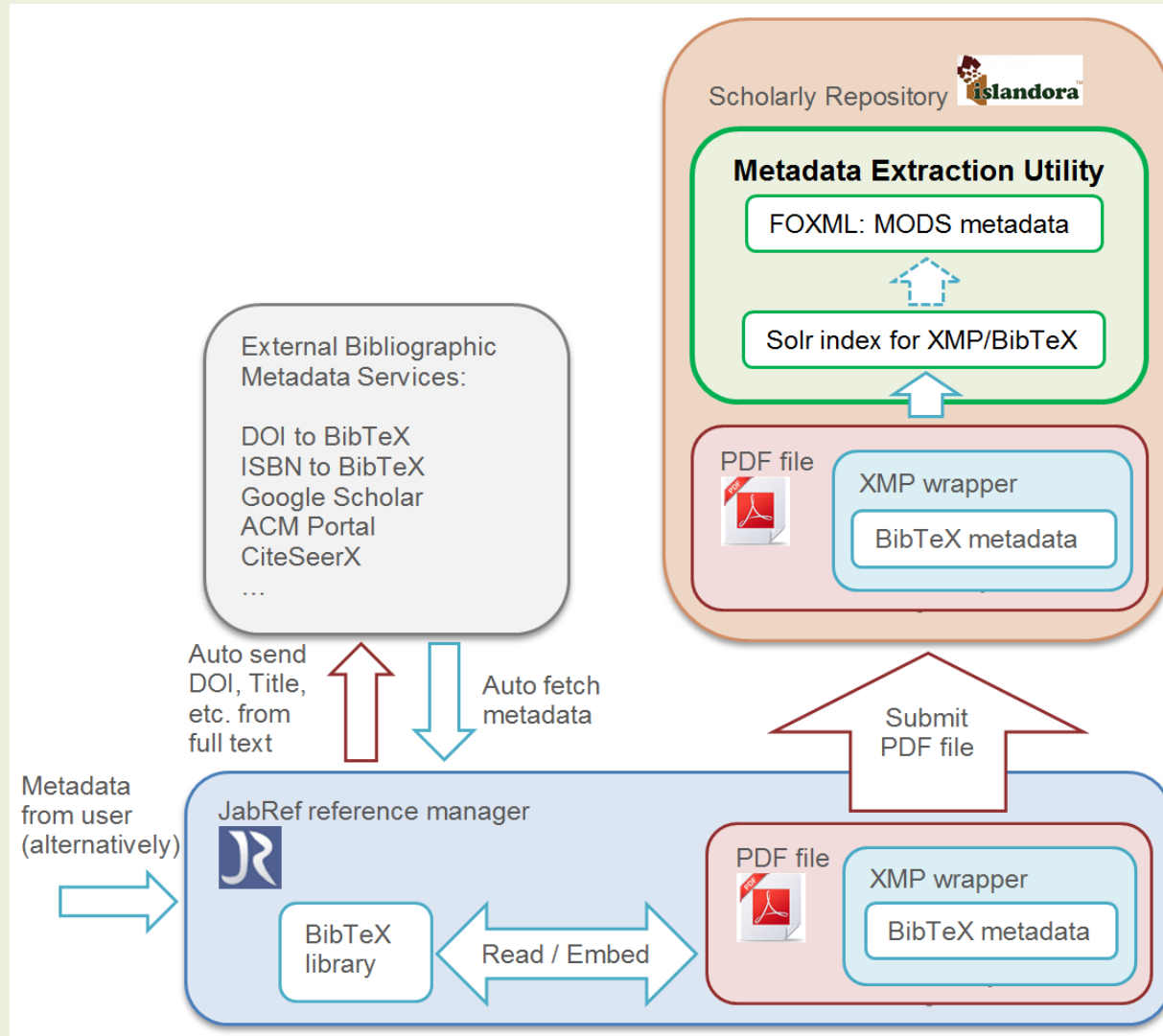
Auto Generated Item Metadata

View

File (OBJ) Embedded Metadata

Title	Learning author-topic models from text corpora
Name(s)	Rosen-Zvi, Michal Chemudugunta, Chaitanya Griffiths, Thomas Smyth, Padhraic Stevvers, Mark
Journal	ACM Transactions on Information Systems
Date Issued	2010
Volume	28
Issue	1
Pages	1-38
Producer	Acrobat Distiller 7.0 for Macintosh
Keywords/Tags	Gibbs sampling, Topic models, author models, perplexity, unsupervised learning
Content Type	application/pdf
Number of Pages	38
Creation Date	2010-01-19, 05:45:31am CST

Proposed utility and workflow revisited



Bibliography

- 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](http://research.upei.ca/files/research/v9%20Senate%2022Apr08.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>