Building Collections in IRs from External Data Sources

Data Selection → Data Transformation → Data Curation → Data Ingest → Data Preservation

Sai Deng, University of Central Florida
Susan Matveyeva, Wichita State University
Introduction

* Data Acquisition and Content Recruitment for Institutional Repository (IR)
  * The reality: Lack of faculty self-submission; limited resources deposited...
  * Factors affecting faculty’s contribution to digital repositories
    * Responses of 1700 scientific researchers to an international survey of digital repositories: Variable quality of materials; insecurity over IR’s long-term viability; researchers prefer subject repository... (Nicholas et al., 2012)
    * Perceptions of faculty members from 17 Carnegie doctorate-granting universities: long term concerns, copyright concerns (Kim, 2011)

* Data Archives development in Open Access Repositories: Among over 2000 OpenDOAR repositories, 80 claim to contain datasets: 31 had no datasets found, 13 had few datasets, 7 were not accessible, only 29 contain datasets. Among those 29 repositories, 15 are subject repositories and 7 are IRs. (Luzi et al., 2012)

* How libraries have been dealing with this reality?
  * Strengthening digital preservation and copyright management (long time effort);
  * Mediated deposit by librarians has become a norm: librarians collect, curate and deposit data from Graduate Schools, faculty, schools and departments...
Data Acquisition and Content Recruitment for IR

* Promoting Data Archiving, Curation and Preservation Services to be Part of the Research Lifecycle

  * Purdue University’s Institutional Data Repository service: Partner with the campus research office, facilitate data curation and cyberinfrastructure, data reference, data literacy... (Witt, 2012)

* Research Lifecycle @ UCF and its Supporting Services Model
  (Under discussion, a cross campus effort initiated by the Information Services & Scholarly Communication Unit in the library)

* Getting Data from External Sources
  * What are those external sources?
  * Which types of data? Datasets?
  * How?
Research Lifecycle at UCF (Work in Progress)

University of Central Florida Libraries

Impact Measures & Prestige

Global Scholarly Community

21st Century Digital Scholarship Cycle

Planning Cycle

Research Planning

Grant Planning

Background Research

Grant Search

Proposal Development

Funding

Communication

Data/Output

Experiment/Project

Grant Management

Conclusions

Draft Work

Peer Review

Publication Cycle

Publication/Presentation

Comments & Revisions

Preserve

Final Reports

Compliance

Disseminate

Research

Research Computing

Institutional Repository

FCTL

Libraries

ORC

Research Computing

Not yet supported
Getting Data from External Sources

* Building Partnerships between IRs and Data Services
  * IASSIST 2012 panel: Institutional Repositories and Data
    * IRs “on the front line of curating a growing variety of data sources;”
    * Partnerships with data producers, other IRs and digital repositories to improve local data curation.

* Creating Collections in an IR or Digital Repository from External Data Sources
  * External data sources: Arxive, PubMed, IEEExplore, Web of Science, open web...
* **Repository Choices: IR, Subject Repository…**

* Wrestling between IR and Domain Based Digital Repository
  * Domain-based Repository vs. IR, IR advantages: Institutional identity, university data management service
  * Researchers prefer subject repository (Nicholas et al., 2012)

* Data Curation and Digital Repositories
  * Monash University’s two types of repositories: Collaboration Repository, Publication/Preservation Repository.
  * Data curation continuum, boundary between the different repositories (Treloar et al., 2007)
A Real Case: Building Collections in DSpace from PubMed

* How to build collections in IR or digital repository: Data harvesting? Data pre-populating to an IR or digital repository? Query external data provider and export results to the repository?

* Some Observations of Data Services/Activities Related to PubMed and DSpace
  * SWORD (Simple Web-service Offering Repository Deposit) ([http://swordapp.org/](http://swordapp.org/))
  * Populating Metadata in Submission with Data from PubMed ([https://wiki.duraspace.org/display/DSPACE/PopulateMetadataFromPubMed](https://wiki.duraspace.org/display/DSPACE/PopulateMetadataFromPubMed))

* Commercial Services
  * BioMed Commercial Service ([http://www.biomedcentral.com/libraries/aad](http://www.biomedcentral.com/libraries/aad))
  * @mire Customization (Dryad case: [https://atmire.com/website/?q=references/dryad](https://atmire.com/website/?q=references/dryad))
* PubMed-DSpace project at WSU Libraries.

* **SOAR**: Shocker Open Access Repository has been in production since 2007.

* **Comprehensive Collection Development Strategy for SOAR**
  * During the first several years, the comprehensive collection development **strategy** was explored;

  * IR Librarian promoted a new **service** to faculty and university administration and accepted numerous types of materials;

  * A variety of **projects** were completed: videos, presentations, science-museum like projects (e.g. collection of pottery of the Museum of Anthropology, virtual herbarium of the Department of Biological Sciences);
Comprehensive Collection Development Strategy for SOAR (Cont.)

* SOAR also included serial publications of the university, such as e-journals and proceedings;

* ETDs program was developed as a collaboration of the Graduate School and the Library;

* Two types of faculty collections were created: departmental and individual
  * Individual faculty collections included e-books, articles, conference papers and presentations;
  * Departmental collections included mainly articles and conference papers.
Comprehensive Collection Development Strategy for SOAR (Cont.)

* Some of these collections were **uploaded** to the repository **title by title**;

* The others, such as Virtual Herbarium were added to the system as **bulk**;

* However, whether manually entered or bulk uploaded, these collections use **internal sources of data**:
  * Full text for individual collections was provided by faculty authors;
  * Full text for departmental collections included materials found at the University’s departmental websites.

* In the last few years, our collection development strategy has changed from comprehensive to **limited**.

* Serial publications (ETDs, proceedings, e-journal) continued, but no new major projects were invited.
* Limited Collection Development Strategy for SOAR
* Bibliographic data is imported from external sources;

* Data is enhanced to ensure consistency of the repository collections and needs of the end users;

* **Full text** is accepted if permitted (practically, limited to Open Access under Creative Commons license);

* **Access** to full text is provided via links in preference order;

* The emphasis is made on **bulk import of faculty articles and conference papers**.
  * Faculty did not provide these materials;
  * All materials are searched for by librarians on the Web and in different databases.
* **How do we use data from external sources?**

  * As *source of information about the work* written and published by the University’s author (this information is not available for us within the University);

  * Record leads us to full text of the work on the publisher’s websites or a hard copy of the journal (if available);

  * We use the work itself as a *primary source of information of a metadata record*;

  * After we export records from external sources, we verify information and modify the records according to our metadata template;

  * We *acknowledge sources of information* by including *rights.holder* field and the record’s ID number in its original database.
PubMed-DSpace Project

Method Exploration: Options?

* Export PubMed XML file (search results) to spreadsheet directly (require heavy manual editing);

* Export Medline txt file to spreadsheet (requires data transpose, columns numbers vary, problematic);

* Transform PubMed XML file to DCXML file using XSLT (source and target XML schemas needed);

* Transform PubMed XML file to DCXML file using VB script and XPath expressions.

PubMed vs. PubMed Central

* All PubMed articles vs. Full text articles only
PubMed-DSpace Project Data Flow

Librarian

1.0
Search PubMed by Institution Affiliation

T1
PubMed XML

2.0
Transform w/ VB and XPath expression

T2
DCXML

3.0
Export to Excel

T3
Excel file

4.0
Edit/Enrich/Enhance Data in Excel:
Data Accuracy and Consistency Analysis;
Name Authority Check;
Check fields against publication template;
Additional Fields added;
Link enrichment;
Peer Review Article Status Check;
Copyright Check;
Divide data into subsets.

T4
Curated Datasets

5.0
Transform to SIP Packages w/ Java Program

T5
SIP Packages for Departmental Collections

6.0
Export to DSpace

Users

External Entity

Process

Data Flow

Data Store/File
Source Data: Search PubMed (http://www.ncbi.nlm.nih.gov/pubmed) by institution affiliation, save the result as XML file;

Data Analysis and Mapping: Refer to MedLine/PubMed Data Element (Field) Descriptions; map PubMed fields to DC elements;

Transfer PubMed XML File to DC XML File

VBScript run in Microsoft Visual Web Developer. Use XPath expressions.

For example: Get value for DC element “identifierIssn” from node "./MedlineCitation/Article/Journal/ISSN” in the retrieved PubMed xml file:

`Check if there’s a ISSN
If node.SelectNodes("./MedlineCitation/Article/Journal/ISSN").Count > 0 Then
    writer.WriteStartElement("identifierIssn")
    writer.WriteValue(node.Item("MedlineCitation").Item("Article").Item("Journal").Item("ISSN").InnerText)
    writer.WriteEndElement()
End If
...

Complexity in Data Extraction and Transformation

**identifierCitation**: need to combine Journal title, Volume, Issue and Year, e.g., Journal title. 2011 Oct; 39(4):320-32.

**dateIssued**: need to combine Year, Month and Day under "PubDate", e.g., 2011-10-01. Used “yyyy-mm-dd” format. Formalized data for these situations: only year and month available; only year and season available. However changed back to their original formats at the final project stage.

**SubjectMesh**: PubMed has DescriptorName and QualifierName. Need to consider different situations: descriptor with one qualifier, descriptor with multiple qualifiers...

**contributorAuthor**: list all authors’ names under AuthorList...
<table>
<thead>
<tr>
<th>DC Field</th>
<th>PubMed Source Field</th>
<th>Node in PubMed export XML file</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>identifier</td>
<td>PMID</td>
<td>./MedlineCitation/PMID</td>
<td></td>
</tr>
<tr>
<td>identifierIssn</td>
<td>ISSN</td>
<td>./MedlineCitation/Article/Journal/ISSN</td>
<td>Check if there's a ISSN in node “./MedlineCitation/Article/Journal/ISSN” of the PubMed XML file, if yes, output its mapped DC element “identifierIssn” with the ISSN value.</td>
</tr>
<tr>
<td>identifierCitation</td>
<td>Title</td>
<td>./MedlineCitation/Article/Journal/Title</td>
<td>To form identifierCitation need to combine Journal title, Volume, Issue and Year e.g., Journal title. 2011 Oct; 39(4):320-32.</td>
</tr>
<tr>
<td></td>
<td>PubDate</td>
<td>./MedlineCitation/Article/Journal/Issue/PubDate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Volume</td>
<td>./MedlineCitation/Article/Journal/Issue/Volume</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Issue</td>
<td>./MedlineCitation/Article/Journal/Issue</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MedlinePgn</td>
<td>./MedlineCitation/Article/Pagination/MedlinePgn</td>
<td></td>
</tr>
<tr>
<td>dateIssued</td>
<td>Year</td>
<td>./MedlineCitation/Article/Journal/Issue/PubDate/Year</td>
<td>To form dateIssued, combined Year and Month under “PubDate”, e.g., “2011-10.”</td>
</tr>
<tr>
<td></td>
<td>Month</td>
<td>./MedlineCitation/Article/Journal/Issue/PubDate/Month</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Day</td>
<td>./MedlineCitation/Article/Journal/Issue/PubDate/Day</td>
<td>Check if there's a day, otherwise use default day “01.” “yyyy-mm-dd” format.</td>
</tr>
<tr>
<td>dateIssued</td>
<td>Year</td>
<td>./MedlineCitation/Article/Journal/Issue/PubDate/Year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Season</td>
<td>./MedlineCitation/Article/Journal/Issue/PubDate/Season</td>
<td>If only season available, replace “spring” with “03-01,” summer with “06-01” and winter with “12-01” (However changed back to its original format at the project final stage).</td>
</tr>
<tr>
<td>relationIspartofseries1</td>
<td>Title</td>
<td>./MedlineCitation/Article/Journal/Title</td>
<td></td>
</tr>
<tr>
<td>relationIspartofseries2</td>
<td>ISOAbbreviation</td>
<td>./MedlineCitation/Article/Journal/ISOAbbreviation</td>
<td></td>
</tr>
<tr>
<td>DC Field</td>
<td>PubMed Source Field</td>
<td>Node in PubMed export XML file</td>
<td>Note</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>title</td>
<td>ArticleTitle</td>
<td>./MedlineCitation/Article/ArticleTitle</td>
<td>Check if there’s a “ArticleTitle” in node ./MedlineCitation/Article/ArticleTitle of the PubMed XML file, if yes, output its mapped DC element “title” with the “ArticleTitle” value.</td>
</tr>
<tr>
<td>formatExtent</td>
<td>MedlinePgn</td>
<td>./MedlineCitation/Article/Pagination/MedlinePgn</td>
<td></td>
</tr>
<tr>
<td>RightsHolder</td>
<td>CopyrightInformation</td>
<td>./MedlineCitation/Article/Abstract/CopyrightInformation</td>
<td></td>
</tr>
<tr>
<td>descriptionSponsorship</td>
<td>Agency</td>
<td>./MedlineCitation/Article/GrantList/Grant/Agency</td>
<td></td>
</tr>
<tr>
<td>title2</td>
<td>VernacularTitle</td>
<td>./MedlineCitation/Article/VernacularTitle</td>
<td></td>
</tr>
<tr>
<td>subjectMesh</td>
<td>MeshHeadingList</td>
<td>./MedlineCitation/MeshHeadingList</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DescriptorName</td>
<td>./MedlineCitation/MeshHeadingList/MeshHeading</td>
<td>Check if there’s a DescriptorName. Need to consider different situations: a descriptor with one qualifier, with multiple qualifiers...</td>
</tr>
<tr>
<td></td>
<td>QualifierName</td>
<td>./MedlineCitation/MeshHeadingList/MeshHeading</td>
<td></td>
</tr>
<tr>
<td>contributorAuthor[1,2,3 ...]</td>
<td>AuthorList</td>
<td>./MedlineCitation/Article/AuthorList</td>
<td>Check if there’s a AuthorList and then list all authors’ names.</td>
</tr>
<tr>
<td>languagealso</td>
<td>Language</td>
<td>./MedlineCitation/Article/Language</td>
<td></td>
</tr>
<tr>
<td>type</td>
<td>PublicationType</td>
<td>./MedlineCitation/Article/PublicationTypeList/PublicationType</td>
<td></td>
</tr>
<tr>
<td>coverageSpacial</td>
<td>Country</td>
<td>./MedlineCitation/MedlineJournalInfo/Country</td>
<td></td>
</tr>
<tr>
<td>identifierIssn2</td>
<td>ISSNLinking</td>
<td>./MedlineCitation/MedlineJournalInfo/ISSNLinking</td>
<td></td>
</tr>
<tr>
<td>identifier2</td>
<td>NlmUniqueID</td>
<td>./MedlineCitation/MedlineJournalInfo/NlmUniqueID</td>
<td></td>
</tr>
<tr>
<td>identifier3</td>
<td>ArticleId (doi)</td>
<td>./PubmedData/ArticleIdList/ArticleId[@IdType='doi']</td>
<td></td>
</tr>
<tr>
<td>identifier4</td>
<td>ArticleId (pii)</td>
<td>./PubmedData/ArticleIdList/ArticleId[@IdType='pii']</td>
<td></td>
</tr>
<tr>
<td>identifier5</td>
<td>GrantID</td>
<td>./MedlineCitation/Article/GrantList/Grant/GrantID</td>
<td></td>
</tr>
</tbody>
</table>
Creating PubMed Collection in DSpace

* Refer to Faculty and Research Publication Template in DSpace (WSU SOAR);
  * dc.contributor.author
  * dc.date.issued
  * dc.identifier (doi)
  * dc.identifier.citation
  * dc.identifier.issn
  * dc.identifier.uri
  * dc.description
  * dc.description.abstract
  * dc.format.extent
  * dc.language.iso
  * dc.publisher
  * dc.relation.ispartofseries
  * dc.source
  * dc.title
  * dc.type
  * dc.coverage.spatial
  * dc.description.version (peer-reviewed status)
  * dc.rights.holder

* After the DCXML file is exported to Excel, additional fields need to be added to the spreadsheet from the WSU research publication template;

* Keep extra fields from PubMed export for this Collection?!
  * descriptionSponsorship
  * subjectMesh
  * MedlineTA (NLM journal title abbreviation)
* Creating PubMed Collection in DSpace

* Edit affiliation field, make department names consistent and turn them into departmental collection names;

* Name authority check: Check OCLC authority file, local Voyager authority file;

* Peer-review article check: Check journal’s peer review status; Search Sherpa/Romeo, or, Ulrichsweb (serials directory);

* Additional fields added: contributor (dept. name), publisher, identifier.uri, description;

* Enriched data in existing fields: rightsHolder;

* Copyright compliance: Only provide links to the article, not resave and host it;

* Data accuracy and consistency analysis;

* Divide records into several departmental collections.
* Excel Data Editing Steps

* Edit affiliation field and find affiliation if not available;
* Check “contributorAuthor” against OCLC and local Voyager authority file;
* Add peer review status for “descriptionVersion” by searching Sherpa/Romeo or Ulrichsweb;
* Delete the value in “identifierIssn2” if it’s the same as the value in “identifierIssn;”
* Change “relationIspartofseries” (journal name) to proper case (use PROPER());
* Add “contributor” (corporate contributor, university department);
* Sort by “identifierIssn,” add “publisher” (check Ulrich’s Periodical and Sherpa);
* Search “identifierUri” by title, enrich doi links and other links;
* Check “rightsHolder”; 
* Add “description” for links;
* Check special characters in the spreadsheet;
* Data accuracy and consistency check...
**Link Information Enhancement**

* Search for additional links and add link descriptions for publication records;

* Access to full text is provided via links in the following preference order:
  * Direct link to the title: DOI, e.g., Click on the DOI link below to access the article. (Description)

  * Link to the electronic journal / proceedings record in the library catalog if the library subscribes this journal; link to the journal website for publications not licensed by the library.
    * For example: The full text of this article is not available in SOAR. WSU users can access the article via commercial databases licensed by University Libraries: ... The stable URL of this article is: ...

  * If title not online, provide a note (no digital copy of this title available, or hard copy available in the library catalog)
    * For example: The full text of this article is not available in SOAR. Check the journal record ... for the paper version of the article in the library. Or,
  
  * The full text of this article is not available in SOAR.
An Add-on to Facilitate DSpace Batch Import Procedure

Part of the Google Summer of Code Project 2008;
This program was written by Blooma Mohan John at the Nanyang Technological University;
Written in Java. It transforms data prepared in Microsoft Excel or OpenOffice spreadsheet to DSpace batch import format, the XML-based Submission Information Packages (SIPs);
The spreadsheet requires inclusion of metadata as well as the locations of the digital resource.
Program wiki: https://wiki.duraspace.org/display/GSOC/Google+Summer+of+Code+2008+Batch+Import

Customization of the Add-on to Generate SIP Packages for Individual Collections from Spreadsheet

For customization of this program, refer to:
Deng, S. Optimizing Workflow through Metadata Repurposing and Batch Processing. Available at: http://www.tandfonline.com/doi/abs/10.1080/19386389.2010.524862#preview
* Data Batch Transfer to DSpace

* Export individual batch to its corresponding location (DSpace handle)

  * Transfer data to DSpace server using psftp, for example:
    
    open soar.wichita.edu
    cd /data/dspace/upload
    put PubMed_PSY.zip

* Log into the server and run the DSpace ItemImport command, for example:

    cd /usr/local/src/dspace/upload
    unzip PubMed_PSY.zip
    cd /data/dspace/bin
    ./dsrun org.dspace.app.itemimport.ItemImport -a -e firstnamelastname@institution.edu -c 123456789/3377 -s /usr/local/src/dspace/upload/PubMed__PSY -m mapfilePubMed_PSY
Examining attitudes of specialist physicians regarding physician assistant referrals

Referral of a patient from a primary care provider to a referral specialist physician for consultation is an important link in the continuum of patient care. Although the relationship between the physician assistant (PA) and supervising physician has been studied extensively, the effectiveness of the relationship between the primary care PA and the referral specialist physician is unclear. The purpose of this study was to examine the attitudes of the referral specialist physician toward direct referral of patients from primary care PAs. A random sample of 5,000 specialist physicians in five specialty areas across the United States were surveyed to determine their willingness to accept referrals from PAs, their attitudes about the appropriateness of PA referrals, and reasons for dissatisfaction, if any. Data analysis revealed that specialist physicians generally are willing to accept patient referrals from primary care PAs and that specialist physicians generally are satisfied with the appropriateness and timeliness of referrals from PAs. Whether or not the specialist physician employed (supervised) a PA in his or her own practice.
# Exports Sample Record in DSpace (Cont.)

In the examination of attitudes of specialist physicians regarding physician assistant referrals, whether or not the specialist physician employed (supervised) a PA in his or her own practice made a significant difference in level of satisfaction with PA referral. To determine the significance of these results, more research is needed to determine specialist physician satisfaction with referrals from primary care physicians compared with PAs.

<table>
<thead>
<tr>
<th>dc.format.extent</th>
<th>270-4</th>
<th>en_US</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.language.iso</td>
<td>eng</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.publisher</td>
<td>Association of Schools of Allied Health Professions</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.publisher</td>
<td>Journal of Allied Health</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.publisher</td>
<td>Volume 32, Number 4, Winter 2003</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.source</td>
<td>NLM</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Attitude of Health Personnel</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Health Services Research</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Humans</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Interprofessional Relations</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Medicine</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Physician Assistants</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Physicians/psychology</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Primary Health Care</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Referral and Consultation</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>Specialization</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.subject.mesh</td>
<td>United States</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.title</td>
<td>Examining attitudes of specialist physicians regarding physician assistant referrals</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.type</td>
<td>Article</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.coverage.spatial</td>
<td>United States</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.description.version</td>
<td>peer reviewed</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.rights.holder</td>
<td>Copyright © Association of Schools of Allied Health Professions</td>
<td>en_US</td>
</tr>
</tbody>
</table>

## Files in this item

<table>
<thead>
<tr>
<th>Files</th>
<th>Size</th>
<th>Format</th>
<th>View/Download</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>There are no files associated with this item.</td>
</tr>
</tbody>
</table>
Batch Editing in DSpace

In this example, the dates in yyyy-mm-dd format were changed back to their original format (as requested);

If number of changes exceeds maximum allowed, limit changes or alter bulkedit.gui-item-limit in dspace.cfg.
More Options for PubMed-DC Transformation

* PubMed to Excel Python Script (by Nitin Arora):
  * [http://blog.humaneguitarist.org/projects/pubmed2xl/](http://blog.humaneguitarist.org/projects/pubmed2xl/)
  * Export PubMed citations to spreadsheet;
  * XSLT stylesheet can be customized to output desirable data;
  * Does not validate data against PubMed and DC xml schemas.

* XSLT and XML Processor
  * Transform PubMed XML to DCXML with XSLT in an XML processor;
  * Use PubMed and DCXML schemas for element validation;
  * Question: DCXML in which format? SIP packages? How to add files/bitstreams to DSpace?
* More Cases: IEEE Project

* IEEE Data Source: WSU faculty publications from The Institute of Electrical and Electronics Engineers (IEEE) Xplore.

* This data was exported to individual research publications for Computer Science (CS), Electrical Engineering and Computer Science (EECS), Mathematics and Statistics (MAT), Mechanical Engineering (ME) and Physics (PHY).
IEEE-DSpace Project Data Flow

1.0 Search IEEE by Author Affiliation and Export Results

T1 Excel Comma Separated Value File

T2 Curated Datasets

T3 SIP Packages for Research Publication Collections

2.0 Edit/Enrich/Enhance Data in Excel:
Map fields to DC and Check fields against publication template; Additional DC Fields added; Distribute titles to appropriate collection sets; Data Accuracy and Consistency Analysis; Peer Review Article Status Check; Copyright Check.

Users

Librarian

Data Flow
More Cases: WOS (Web of Science) - Retrospective Project

*The project goal is to export to the repository’s appropriate collection retrospective records derived from the WOS database for all available years;

*Because librarians do not have an internal source of data on faculty publications, the only way to collect this data is to turn to external sources;

*WOS is the largest commercial source of data on published articles & conference proceedings in all subjects from 1988. It contains over 41 million records collected from over 12,000 journals.
WOS-DSpace Project Data Flow

Librarian

1.0
Search WOS by Organization and Export Results

T1
EndNote Citation

T3 Curated Datasets

3.0
Transform to SIPs w/ Java Program

T3
Curated Datasets

T4
SIP Packages for Faculty Publication Collections

4.0
Export to DSpace

Users

2.0
Export EndNote Txt Files to Excel

Excel Data

Data Store/File

Edit/Enrich/Enhance Data in Excel:
- Sort Excel file by Authors;
- Map fields to DC and Check fields against publication template;
- Additional Fields added;
- Data Accuracy and Consistency Analysis;
- Name Authority Check;
- Link enrichment;
- Peer Review Article Status Check;
- Copyright Check;
- Divide data to subsets;
- Check for duplicates.

Process → Data Flow

External Entity
Related Issues

Issues to Think About

* Is it appropriate for an IR to be a back up resort to institutional research generated datasets in subject repositories and other digital repositories?

* Open access sources and U.S. federal government sources are okay to use. We need to read commercial companies papers to see what they permit us to use.

* Can librarians provide data curation services? Will science data curation services be better handled by institutions or scientific communities? (The Research Data Access & Preservation Summit, Wickett et al., 2012)

* Do librarians have sufficient training to describe and preserve scientific data? (Newton et al., 2011)
Thinking About Data Curation: DCC Curation Lifecycle Model

- DCC: Digital Curation Centre, UK
- Full Lifecycle Actions: Description and Representation Information, Preservation Planning, Community Watch and Participation.
- Sequential Actions: Conceptualize, Create or Receive, Appraise and Select, Ingest, Preservation Action, Store, Access, Use and Reuse, Transform.
- Occasional Actions: Dispose, Reappraise, Migrate

From http://www.dcc.ac.uk/resources/curation-lifecycle-model
It involves activities in “Appraise & Select”, “Ingest,” “Preservation Action” and “Store.”

Building collections from external data sources put emphasis on: “Access, use, reuse” and “Transform.”

Reuse: repurposing datasets from external data sources to IR or digital repository. Transform: preserve subsets.

Value added data services are part of the curation process. This dynamic model can be entered at any point?

These projects promote data sharing between the subject repository, data providers and institutional repository?!
To get data from other sources is an alternative way to develop collections for different disciplines since author self-deposit has not become a common practice for IRs. It can be a good way to find information for retrospective projects.

This is an effort in line with the current metadata cataloging trend of moving from item by item cataloging to batch processing of datasets, repurposing of metadata between different systems and communities, and providing value-added data curation services to students and faculty in an IR.

For small to medium sized institutions that do not have the budget for commercial data services, it is possible to get data from external databases or data sources through some procedures without infringing copyright policies.

If an institution prefers data enhancement to the external source, it may work better to export the data to a software program allowing local batch editing such as Excel than pre-populating it to the IR directly.
Thoughts/Conclusion

* The need for external data source in faculty publications arises from its unavailability on the institutional level.

* The situation may change with the improvement of research data management at universities.

* University administration needs better statistical data for reporting and assessment purposes; more universities organize faculty research data using online systems and some systems provide seamless flow of data to institutional repositories.
* Possible Topics to Discuss

* Data acquisition and content recruitment for IRs
* Institutional Repository vs. Subject Repository
* Creating research publication collection in IRs and digital repositories
* Data batch transfer from external sources to IRs
* Data curation for IR and digital repository
* Data curation services to support research lifecycle
* Data management plan (Project level, institutional level)
* Transform of Medline in PubMed to DC in DSpace
* Workflow management
* Value-added metadata services
* Metadata fields selection and enrichment
* Copyright compliance
* Data acquisition from the open web...


Contact

Sai Deng
Associate Librarian and Cataloger/ Metadata Librarian, University of Central Florida Libraries
Email: sai.deng@ucf.edu
(Previously worked as a Metadata Cataloger at Wichita State University Libraries)

Susan Matveyeva
Associate Professor & Catalog & Institutional Repository Librarian, Wichita State University Libraries
Email: susan.matveyeva@wichita.edu

Thank you!