Migrating a DSpace Subject Repository to Islandora

Library Technology Conference

Macalester College

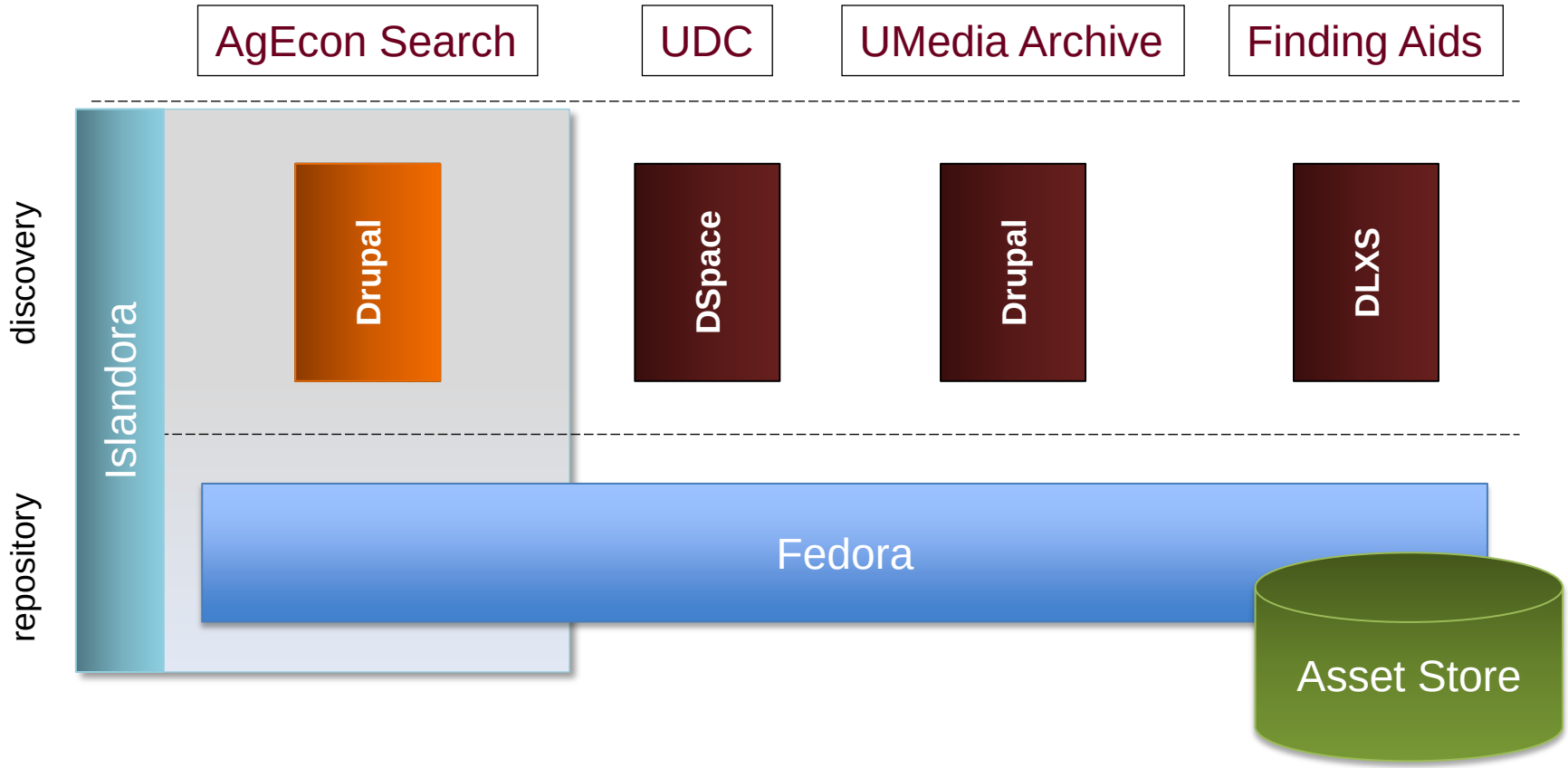
17 March 2011

Jason Roy & Jeff Silvis

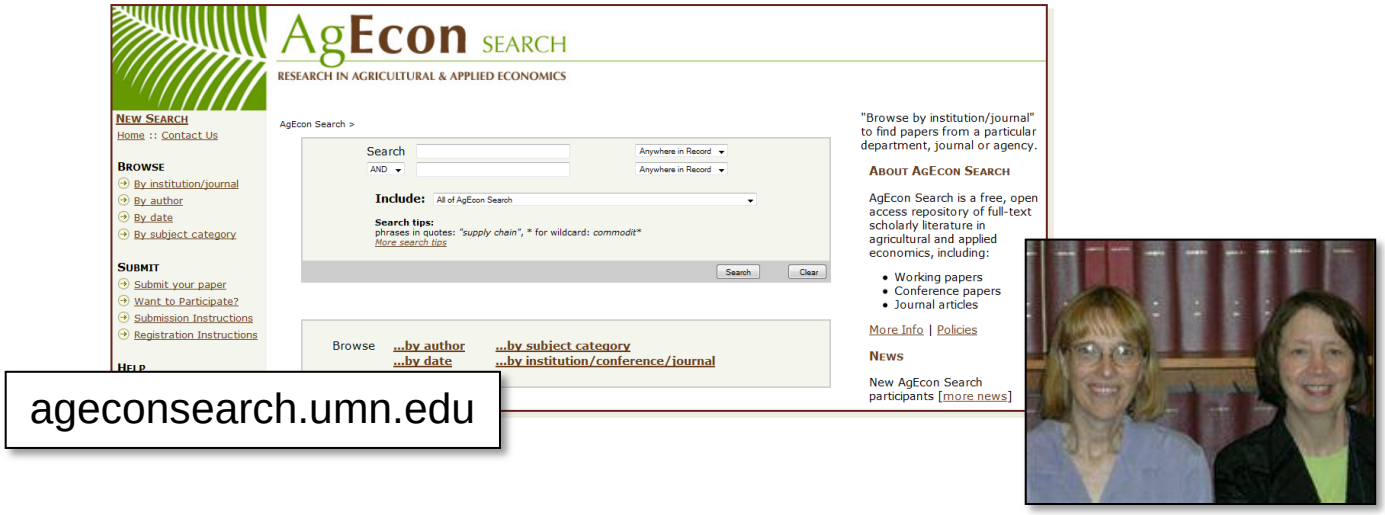
Digital Library Services

University of Minnesota Libraries

Digital Library Services



AgEcon Search



ageconsearch.umn.edu

- Leading international repository for Agricultural Economics.
- Currently over 42,000 records in the repository.
- Ranking Web of World Repositories, January 2011
  - 9th among all U.S. and Canadian repositories
  - 42nd worldwide
  - Based on a combination of size, visibility, file content, and scholarly depth.

Project Scope

Issues

- Critical stability issues, particularly during peak ingest periods.
- Need to extend data model to accommodate new data types.

Scope

- Deliver a stable and reliable system that replicates current functionality.
- Develop a more robust metadata schema to incorporate new material types.
- Position ourselves for future modifications and enhancements.

Out-of-Scope

- Enhancements or changes not related to either stability or functionality.

“The problems facing AgEcon Search if not addressed pose potentially critical consequences to the services’ reputation and its viability in the minds of its user base.”

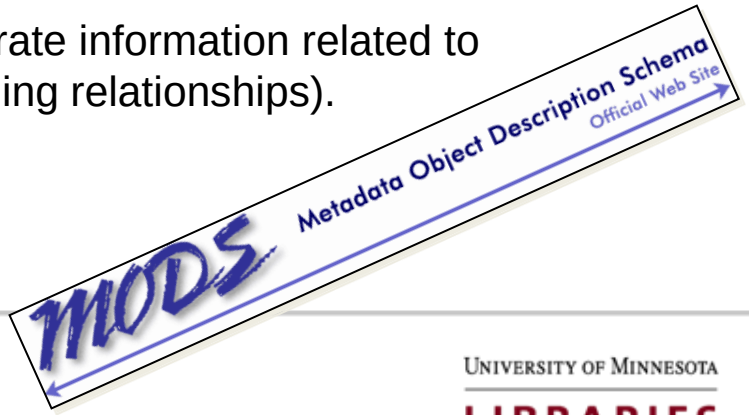
Metadata Maturation

Dspace provides support for a Dublin Core based data model. However, as AgEcon has matured its own metadata needs have grown. The new **Islandora** framework will encompass a much more expansive MODS format.

Our challenge has been to map and migrate all of the existing metadata into a new MODS-compliant structure. This mapping occurred against three distinct areas within the DSpace data set: item level, collection level, and people level (only two of which matter here!)

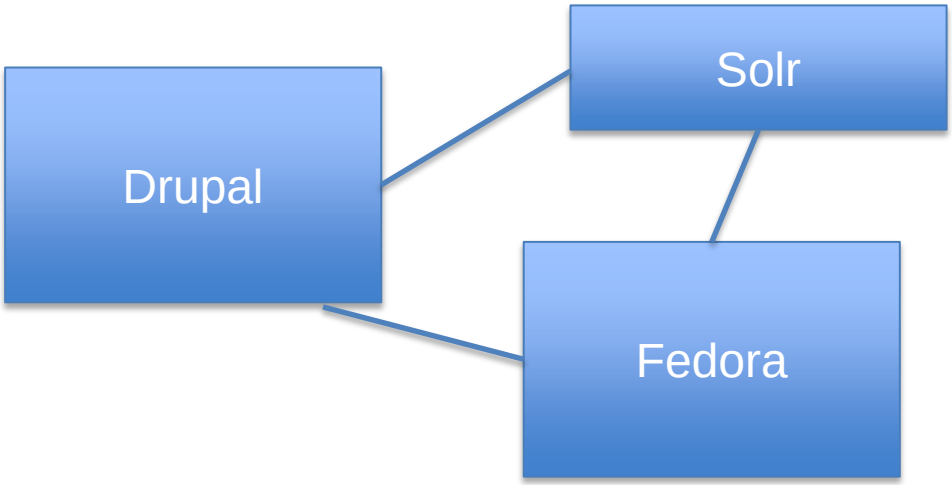
1. **Item level** – mapping of the record level information from Dublin Core to MODS. I've included a handout detailing our MODS XML framework.

2. **Collection level**– identifying how best to migrate information related to communities and collections (parent-child-sibling relationships).



Islandora

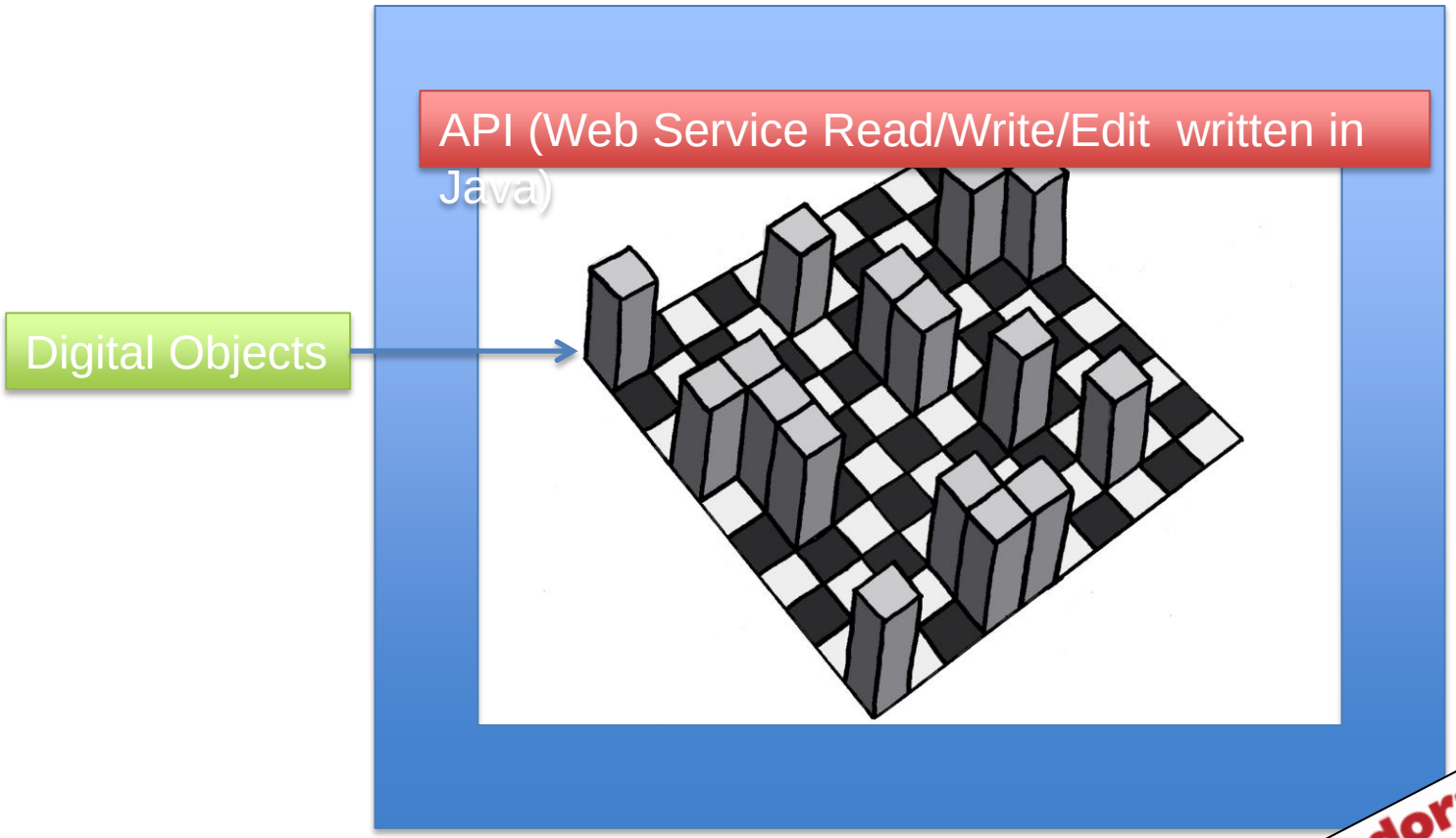
Developed by the *University of Prince Edward Island*



Three components

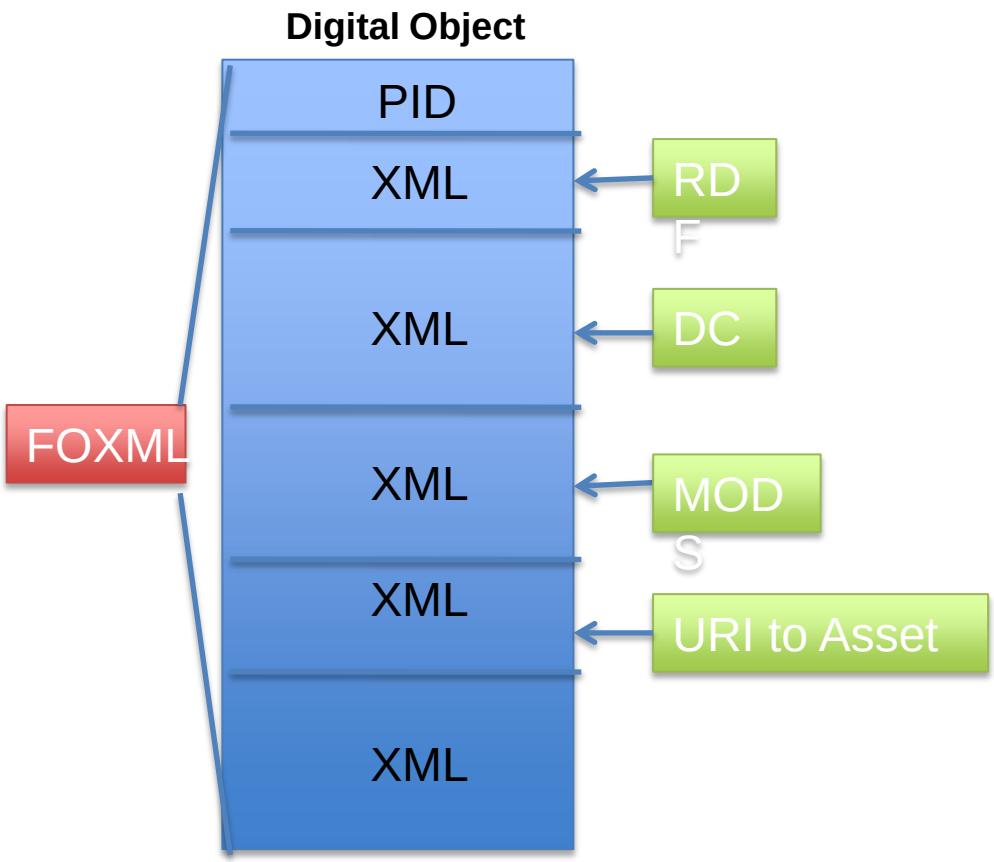
- Drupal front-end facing the web
- Solr – Search engine
- Fedora data store

Fedora from 10,000 feet



FedoraCommons™

Anatomy of a Digital Object



A digital object is **just plain text**. It consists of a series of xml documents that are wrapped together to form FOXML.

- Item metadata is stored the MODS datastream
- There is a datastream that points to our asset (a pdf, tiff ect)
- The rdf data stream connects digital objects together

Fedora Tour

- Ingest
- Web Services
- RDF to make interconnections
- External workflow

Fedora Tour

Ingest → Web Services → RDF to make interconnections → External workflow

- Use the API
- Use Fedora provided script. Write FOXML text to a specific directory. A Fedora provided script will automatically ingest the files. Excellent batch load system.

Fedora Tour

Ingest → **Web Services** → RDF to make interconnections → External workflow

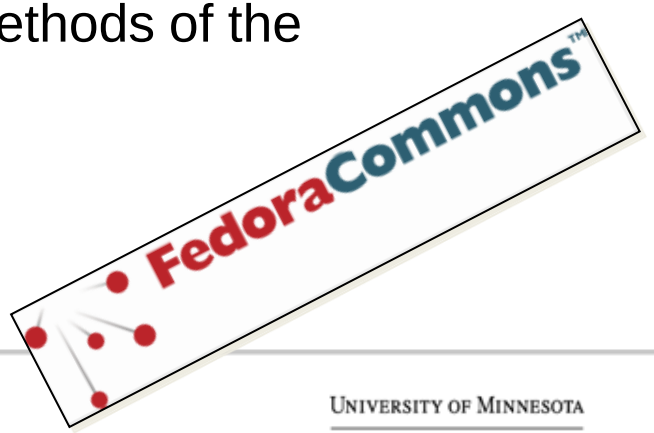
- Fedora supports both REST and SOAP.
- Rest is “URL-like” … simple.
- Fedora provides URL to access data streams.
- For instance to get a dump of the MODS xml for a data object with a pid=agecon:330 would be http://www.myrepository/agecon:330/MODS
- What does this buy us?

Fedora Tour

Ingest → **Web Services** → RDF to make interconnections → External workflow

Example: ***Batch extract/manipulation of metadata***

- Want to extract MODS for pid=agecon:330 to pid=agecon:4480, just need to write urls:
  http://www.myrepository/agecon:330/MODS
  http://www.myrepository/agecon:331/MODS
- A very short script - about 20 lines.
- Batch editing is also possible (the write methods of the Fedora API are also available to REST)



Fedora Tour

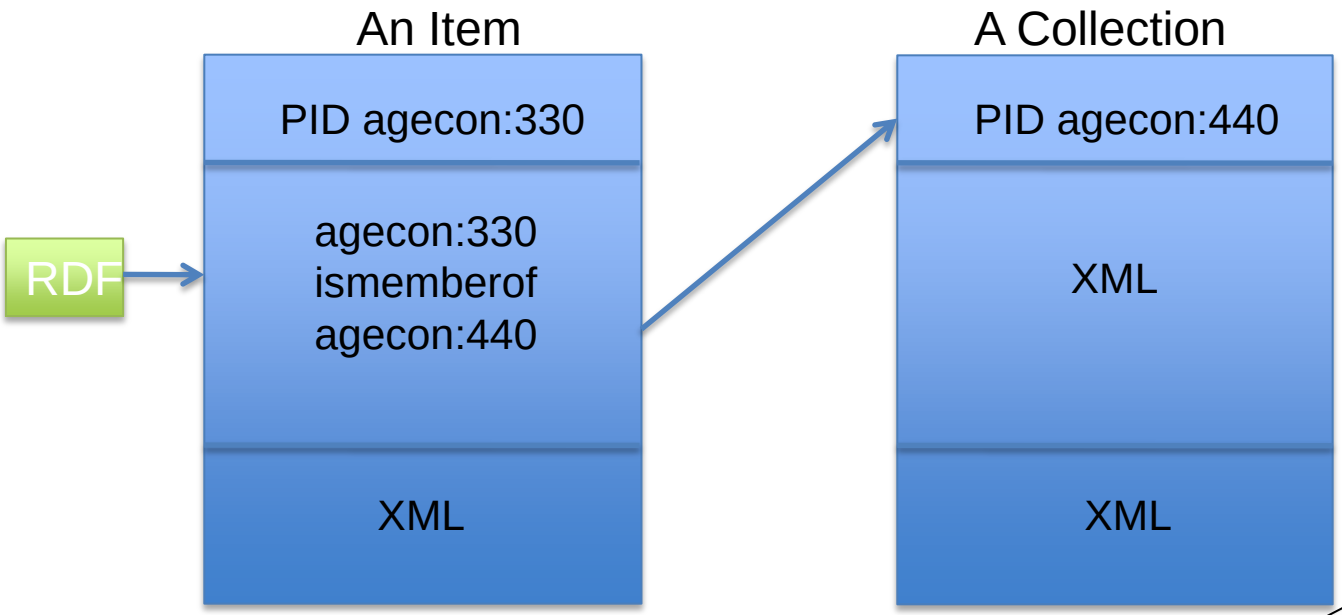
Ingest → Web Services → **RDF to make interconnections** → External workflow

- Way to link objects together on the Web.
- Uses subject-predicate-object expressions.
- For instance: Jeff *isemployedby* UMN.
- Two ways we are using RDF:
  - Defining collections.
  - Creating Content Models.

Fedora Tour

Ingest → Web Services → **RDF to make interconnections** → External workflow

RDF defines Collections



FedoraCommons™

Fedora Tour

Ingest → Web Services → **RDF to make interconnections** → External workflow

RDF to create Content Models

How to turn MODS into citations?

- With a content model we can put the instructions for this in one place and any digital object that wants to create citations.
- The RDF looks like this: agecon:330 *hasmodel* citationmaker
- Now there would be a new URL: http://www.myrepository/agecon:330/citation

Fedora Tour

Ingest → Web Services → RDF to make interconnections → **External workflow**

- Workflow functions (e.g. OCR, technical metadata extraction) can be done entirely outside of Fedora and communicate through SOAP or REST.
- These external services can be upgraded or replaced with minimal impact on Fedora.
- They can be used separately by applications other than the repository.

Drupal

About **Drupal**

- Front end for our system; connects fedora to users.
- A rich CMS, written in PHP.
- Very Modular design.
- Extremely easy to create forms.
- Many custom plugins (Modules).
- Strong support at UMN for Drupal.

How we use **Drupal**

- All displays
- Form for user input of data, collection creation and management
- User creation and management
- Managing Solr (our search engine)
- A Google Maps mashup
- Management of Google Analytics
- Islandora module allows Drupal to talk to Fedora

Questions?

New and improved **AgEcon Search** set to launch at the end of *April 2011*.

Jason Roy
jasonroy@umn.edu

Jeff Silvis
silvi003@umn.edu