A Framework for Everything:
Docs and Data in an
Open Source Environment
OLA SuperConference 2011
Session # 1018

Melissa Belvadi,
Emerging Technologies & Metadata Librarian
Donald Moses,
Collections and eResources Librarian
University of Prince Edward Island

Title slide - M - Hi, how many are systems folks? directly involved in any kind of digitization (locally digitized content)? know what drupal is? have installed a new drupal module? coming from an institution using CONTENTdm? DSpace? Have you worked with Dublin Core? how many heard Mark's presentation in 2009 at OLA or anywhere else about Islandora?
Agenda

- Institutional Context
- VREs at UPEI
- The Framework
- Example Sites (many!)
- Q&A

Agenda - M - quick read-through
UPEI and Robertson Library

- UPEI: 4,600 students, 200 Faculty
- Robertson Library:
  - 35 total staff
  - 10 full-time systems
- Shifting positions with attrition
  - Circulation
  - Acquisitions
  - Cataloguing
- Money from: Internal - Library,
  Internal - UPEI, grants, donations
- Server control
- Open source-friendly

Institutional context - M - Of the 10 FT systems, 5 are grant-funded term;
Money from: internal eg indirect costs of research, library budget; portions of
faculty research grants;
special project grants from community members/orgs, ACOA's Atlantic
Innovation Fund AIF), Inukshuk grant, ACDP, private donations
- embrace OpenSource - ie Godot, CUFTS, Evergreen, Redmine, Drupal, Moodle,
OpenOffice

The library controls its own servers and systems with minimal support
from the campus IT department aside from networking and power
infrastructure… we do cooperate with our institutional IT services, which
can be critical for developing these kinds of projects.
VREs at UPEI

- Virtual <something useful> Environment
- Administrative records
- Research
- Digital collections

What is a VRE - M -
The framework we’re talking about today was first conceived as a Virtual Research Environment for research teams, but its flexibility has morphed it into a Virtual <something useful> Environment.

It is used at UPEI to help various campus groups manage their internal Administrative records - committee agenda/minutes/forum, registrar document management,

Research groups with private data management, collaboration, and communication of results with public,

And as a repository system for Digital collections - an institutional repository for publications, digitization projects, and other community resources of interest to researchers

We’ll show you examples of all three of these uses. These are just a few examples - we actually support over 100 sites and growing every day.

This platform is suitable for use with both public and private collections of materials.
Using Open Source

- Firefox
- OpenOffice
- Drupal
- Moodle
- Evergreen
- LAMP - Linux, Apache, MySQL, PHP/Perl/Python
- Fedora
- Islandora

Melissa mentioned that we’re open source friendly and I’m sure many of you are already using open source.

We use open source software **whenever it makes sense**. So … for our library catalogue we use *Evergreen*, for our repository *Fedora*, our Library’s website is *Drupal* based, and our development of *Islandora* are all reflections of that philosophy.

The other thing we do is purchase service/maintenance agreements for our open source products … for example the catalogue - *Evergreen* - is a critical component to our information services, so we have a ‘platinum’ support package that provides 24/7 response. Similarly our ejournals openresolver KB and ERM is hosted by SFU and we pay a service fee for support and hosting.

It’s not a new idea, but for us we’ve rather invest in working with other partners to develop high quality free products rather than to pay annual software licensing fees. We’d rather build equity in our house instead of paying rent.
The core pieces in our framework are Fedora at the back-end, Drupal at the frontend. Islandora bridging the
Fedora stands for Flexible Extensible Digital Object Repository Architecture. Fedora can securely store any type of digital content and the information about the content itself. As well it has additional features that allow for the creation of relationship between object, for data or collection stewardship it has built in file integrity checking, and takes care of administrative information about your objects like creation date, and provides for versioning too. Fedora wasn’t built as a stand alone application, but was meant to be integrated with other applications.
That’s where Islandora and Drupal come in.
Drupal is an open source, web-based content management system. Drupal has many many community contributed and supported modules that can be installed. One of the features we really appreciate about Drupal is its use of Roles and Permissions to help manage access … whether it is to content or to functionality within our framework. We utilize a number of different modules depending on the needs of a particular project. For instance on many of our Administrative VREs which Melissa will talk about we use the LDAP module so that our faculty and staff can use the same credentials to access their VREs that they use for accessing their campus mail account.
Islandora is a Drupal module that connect the two software applications together to provide a robust, modular, and flexible, extensible digital asset management system.
Islandora - is a collection of Drupal modules that users will be able to turn on or off to work with a variety of collections and to provide additional functionality.

- Digital Repository module that provides all the functionality of managing and displaying digital assets.
- Fedora Attach … when you attach a file to a drupal node you can have it also have it ingested into the repo
- Fedora ImageAPI … is used to create thumbnails or generate images on the fly
- Islandora Book … provides bib item displays, a book specific search, and the viewer.

- There are many more Islandora modules … eg. Learning Object Repository Module, Melissa is going to demonstrate our Institutional Repository module, etc.

This is a group of modules that we use as part of the IslandLives project which I’ll be showing you later.
There are several different kinds of digital objects within a Fedora repository. This is an example of a data object that uses a map content model. The object is composed of several parts … an identifier or the PID, a set of properties that are set by Fedora, the Relationships the object has … for example I know that this object isMemberOf the imagined:collection, a Dublin Core datastream, an Audit Trail documenting all the changes to an object, and then the Datastreams. The images are all derivatives of an original TIFF.
Management

- Management Team
  - University Librarian
  - 3 Librarians
  - Project Manager
- Regular Meetings
  - Management - weekly
  - Tech group - bi-weekly
  - Whole - bi-weekly
- Management Tools
  - Redmine tracking (open source)
  - Merlin project management (commercial)

Management - M - quick read-through
Faculty of Education Admin VRE

Fac Educ M- basic internal site, uses campus LDAP for authentication
Fac Educ 2 M - Shared site for multiple committees to keep their agendas, minutes, other related documents.
Faculty of Education - 3

Fac Educ 3 - M - Many people are likely on more than one committee, and they can access all the material in a single place with one login
Gov Docs 1 M - the most basic of Islandora sites, and one still being built but already useful. Save local copies of important govt docs in our own secure server, for items we already have catalogue records for because we have, but will now discard, the print. Here we start with the catalogue record. Note the two full text links. The second [click]
Gov Docs 2 M - takes you directly to the PDF so at this point, the end user never sees Drupal, just straight to the document
Gov Docs 3 M - here’s what the site looks like so far - virtually no work on the front end yet, but we do have a basic search box, and about 2,000 full text books ingested. Eventually we plan to add much more sophisticated search and display features, but the priority right now is uploading, we call it “ingesting” in the Fedora world, PDFs and linking to them in the catalogue
Gov Docs 4 M - Donald explained about content models. This staff view of the “datastreams”, another Fedora word that most people would call “files”, exposes some of the content model, which in this case is a Refworks XML record, which when ingested, a DC conversion, or “crosswalk” in the “XML” world, is created automatically. Then as a separate step, a staff member uploads a PDF to go along with the RefWorks and DC metadata. Eventually we’ll add the step where the PDF’s searchable text gets added to a search engine, but for right now only the DC and RefWorks metadata is in the search engine.
IslandScholar 1 - M - when most librarians hear the word “repository” they probably think of “institutional repository” or IR, meaning the digital collection of the published works of the faculty. Here is our “IR”, using our Islandora framework. Unlike most IRs, we started by putting in the metadata, that is the citations, for all of the faculty publications, and then are adding the PDFs etc one at a time, and encouraging the faculty to do the same. Also unlike most IRs, we left student dissertations and theses for last, rather than starting with them.
IS 2 - M - Here are the first few citations for a particular faculty member. Note that the first has a “pre-print” full text, the second just a normal full text, and third has no full text.
IS 3 - M - Here is a sample full record, citation, abstract, and full text. We’ve implemented the COINS, Content Object in Span, standard so that for those citations without full text, visitors with a COINS-enabled web browser will have an easy link to see if they can access the full text in their own library’s collections. How many of you use the LibX toolbar? That’s an example of a COINS plugin. The colored button is generated by my LibX toolbar, and is not hardcoded, so if I had your LibX toolbar installed, I would get the link to your openurl resolver. As we anticipate that it will take years to accumulate all of the full text articles, in the meantime this is important to helping users find the full text when they have access to it.
IS 4 - M - One of the reasons we ingested citations first is to make it possible for the faculty to upload the full text for themselves. In many cases, the faculty have signed away their rights to provide their institutional repository with a copy of their work as finally published, and only have the right to provide a pre-print or post-print version. Since only they have access to those early versions, they need to be the one to upload it. The IR allows faculty members to login, using LDAP again, to upload full text documents. In this example, we’ve logged in as a faculty member who published in the Journal, Experimental Brain Research, opened up the citation, and have scrolled down past the citation and abstract to the second screenful. We’ve included as part of the content model a hook into the free Sherpa Romeo service, which provides information as to what the publisher allows article authors to do with the post-print or final article; so the information is right there [click] ...
IS 5 - M - for them to consult as they scroll down further to use this form to upload the full text. The form itself is defined in the “content model” for this site as well. As you can see, in addition to providing the document, they indicate to the IR at this point which version of the document they uploaded - this results in that bracketed pre-print note we saw a moment ago. A few more legalisms for them to answer, and the document is uploaded. If they upload a Word document, which is likely for a pre-print, the content model will convert it to a PDF as well as providing the full text for the search index to use. If they upload a searchable PDF, that too is fully keyword indexed.

IS is not really complete yet. We have plans to add some features in the next year that will be required by Theses Canada, to add student theses and dissertations, to turn on OAI harvestability, to include ETD metadata for the theses. ETD is the most well-known XML metadata standard, or “schema” for “electronic theses and dissertations” and the one that TC adopted. Using data standards and tools that will allow for easy harvesting and migration of data is an important part of the backend design of this framework.
How do you create a digital collection … our first digitization project was done with a large format flatbed scanner. Just as we’ve been building our application framework, we’ve also been building our digitization capacity and skill set. Again … we could choose to out source our digitization, but for us it made more sense to build those skills and infrastructure locally.

It provides us with an opportunity to learn new skills, to offer new services, and provides us with partnership options.
Digitization Workflow

Just to give a sense of how our digitization efforts tie in with the rest of the infrastructure ... we start with a book, ensure that we have permission to digitize ... and we actually digitize the permission form itself and attach that to book object ... a ticket gets created in Redmine and that initiates digitization of the book, the book gets scanned by either the ScanRobot or the Indus Book Scanner and the images adhere to our filenames convention, and then those tiffs are ingested into the repository using our book ingest script, and finally the user can have a look at the digitized work.
Island Imagined started with index cards... we mapped there information to MODS. This was a first draft...
II is a collection of historic maps, charts, atlases, and some book material drawn from the PEI Public Archives and Records Office, the PEI Museum, and the Robertson Library. This site uses a framework that works best with large format images, but could easily work for any TIFF-based collections. We made the choice to use the Metadata Object Description Schema (MODS) for describing the maps … it has a great range of geographic specific elements that allowed us to fully describe the maps.

We’re going to do a keyword search for maps related to Lot 65. Underlying the search is Solr …
On the left side we see a set of facets generated by Solr (and it is a wonderful way to find data inconsistencies) and in the right panel we can see a set of results with some basic metadata. If we select the first item …
We get to a zoomable/panible interface which important for this kind of content. We’ve extended Islandora by creating an dynamic image viewer base on OpenLayers. These images are served up dynamically to the browser by a jp2000 image server called djatoka. If you remember that digital data object example I showed previously … we are seeing a tiled version of the JP2 datastream from this object. A feature that we’re missing from our current viewer is the ability to rotate the image.
This is the image almost zoomed into its full extent.
When we select the Description tab, we get an HTML version of the MODS xml datastream … again … generated dynamically we are transforming an XML datastream and turning it into HTML. I’ll just bring your attention to the error in the metadata … Plan of Lots 65 …
We’ve been able to edit DC and QDC for some time, but we’ve only recently developed an editor for MODS metadata. It’s currently on our test server, but I thought it’d be useful to demonstrate … it fully represents the MODS standard in a tab/collapsible/web-based forms interface. It seems a little over the top … this is the ‘FULL’ version of the form … so every MODS elements is fully represented.
Here’s my correct version of the title … you can delete the incorrect version of the title, save it
it gets validated as part of save process
it transforms the MODS datastream and updates the DC datastream at the same
time so that the details remain in sync
it gets validated as part of save process
it transforms the MODS datastream and updates the DC datastream at the same time so that the details remain in sync
it gets validated as part of save process
it transforms the MODS datastream and updates the DC datastream at the same time so that the details remain in sync
it gets validated as part of save process
it transforms the MODS datastream and updates the DC datastream at the same
time so that the details remain in sync
it gets validated as part of save process
it transforms the MODS datastream and updates the DC datastream at the same time so that the details remain in sync
it gets validated against the MODS schema as part of save process
it transforms the MODS datastream and updates the DC datastream at the same time so that the details remain in sync
We’ve also got static content on the site …
Once you’ve got a bunch of digitized maps there are all kinds of things that you can do with them. In this example we’ve georeferenced selected historic maps (we hope to do all of them) and have overlayed them on Google Maps. The oldmapsonline site and the NCMaps sites were inspirational. Things that we’d like to add to this site include a map-based Geo search.
This started as a proof of concept site … a BC Public Library has implemented it and another is getting ready to.

Anyone working on a newspaper digitization project? In our community, whether it is a local genealogist or a material culture historian, newspapers are viewed as an incredibly valuable resource. This site has a few content models … a collection object, a newspaper issue object, and a newspaper page object. the pages are PartOf issues and the issues are MembersOfCollection … collection.
If we do a search for the phrase “rice point” we get a set of page results … These are default results that come with Islandora … the labels haven’t been changed or updated. There’s not much metadata associated with newspapers so it is pretty bare. If we select the first result …
IN - 3 … we get to the 4th page of this issue …

One of the things that we like to do is reuse and extend our work.

This newspaper page viewer is basically the same one used for maps, but we provide additional options ….

so I’ve selected the T icon which when selected shows the OCR’d text version of the page. We are always looking for ways to improve our interfaces and we are working with a 4th year computer science class to add search term highlighting to page images … much like the ink.ourontario.ca or chroniclingamerica sites provide. And the class is actually getting assistance from Art Rhyno …
IN - 4 … just wanted to show you that there can be different sites that point to the same content. This is UPEI’s student newspaper that was originally digitized and stored in the UPEI Archives site … it is a newspaper … so it also appears in the IslandNewspapers.ca site. It’s not a copy … it is just pointing to the content in the repository.
IslandLives - D - This was the first project we did in Drupal 6 and Fedora 3. The highlights for this project is that we start with a book … we get permission to digitize and display it if it is still under copyright … we digitize it using our ScanRobot or IndusBook Scanner, do Q/A and filenaming on the of TIFF files generated, then feed that through an automated process. Right now it is a series of scripts and processes that are chained together … that take the TIFFs generated from the digi process and OCRs them, generates an XML-based file that goes through several tranformations, including generating a TEI version of the book, that is enhanced with the named entities encoded. So we are always trying to add value to the content we produce.
IL - 2 doing a search for the phrase “rice point” returns a set of results … all blue covered books apparently … I select the second result
IL - 3 … we see the ‘full record’ with several options including Read, Checkout, Download … below that is a page thumbnail … that’s where the phrase was found in that book.
you can turn the text view on or off …
similar viewer as we saw with the newspapers …
the page image on the left and the textual version of it on the right … which is an HTML version of the TEI stored for that page object.
Island Lives - 5

IL - 5 … one thing we want to be able to do was correct the errors in the OCR or correct the encoding of the connect. We’ve got a web-based TEI editor that highlights lets authorized users do just that … the encoded terms are highlighted and the interface allows you to add or delete TEI elements … or correct the text.
IL - this shows the source view of the TEI document and it can be edited as well.
Herbarium - D

This is a collection of plant specimens that we created along with the Faculty of Science. They supplied the student labour and the content and we supplied the hardware, software, and our metadata and collections expertise.

We are looking at different ways of presenting search interfaces ... much of it depends on what metadata you have. in this case we are using DarwinCore to describe these plant samples and part of that metadata includes geo-locative data and so we can leverage that in the search interface by generating a KML file that can be overlayed on a Google Maps interface.
Herbarium 2 - on the left hand side of the panel you can see the faceted browsing options ... and facets are great for hierarchical data like tree of life type information. On the right side we see a generic list display of records. We've not had a chance to customize this display yet. We’ll select the first hit.
Herbarium -3

Herbarium 3 - this is the same viewer we use for maps and any of our high resolution content. This is an important tool for botanists and we are looking into ways of extending it by adding the option of image annotation. The metadata is in the description tab ... click
Herbarium 4 - and again we are seeing a fairly basic display of the metadata.
Herbarium 5 - If you are logged in with an administrative role you can see other options for working with the metadata including a simple forms based edit option.
Herbarium 6 - and this is where we store the latitude and longitude data for each sample … Now Melissa's going to show you some demonstrate some of our data intensive projects...
MNPL - completely functioning, uses Fedora with special metadata schema
meet with Febrice on Tuesday

MHL - "Blasts" - Peter

RPSIDD - audio - 'Debbie'

Herbarium - Darwin Core

Talk to Zac - CHBMR,

- CBMRN (mastitis)
MNPL - completely functioning, uses Fedora with special metadata schema meet with Febrice on Tuesday

MHL - "Blasts" - Peter

RPSIDD - audio - 'Debbie'

Herbarium - Darwin Core

Talk to Zac - CHBMR,

- CBMRN (mastitis)
MNPL - completely functioning, uses Fedora with special metadata schema
meet with Febrice on Tuesday

MHL - "Blasts" - Peter

RPSIDD - audio - 'Debbie'

Herbarium - Darwin Core

Talk to Zac - CHBMR,

- CBMRN (mastitis)
Scientific Data - MHL

MNPL - completely functioning, uses Fedora with special metadata schema
meet with Febrice on Tuesday

MHL - "Blasts" - Peter

RPSIDD - audio - 'Debbie'

Herbarium - Darwin Core

Talk to Zac - CHBMR,

- CBMRN (mastitis)
MNPL - completely functioning, uses Fedora with special metadata schema
meet with Febrice on Tuesday

MHL - "Blasts" - Peter

RPSIDD - audio - 'Debbie'

Herbarium - Darwin Core

Talk to Zac - CHBMR,

- CBMRN (mastitis)
Scientific Data Plus - MNPL

Digital repository

Key: 041

Taxonomy:
- Type: Algae
- Phylum: 
- SubPhylum: 

MNPL Dive Collection: Key = Key041, (G)Gyposphaeridia (G)
DiscoveryGarden/Lib2o
Booth #T4
Get Involved in Islandora

- Download the VM and try it
- Open Source Participatory Community
  - Islandora GoogleGroup
  - Code on GitHub
  - Duraspace (Fedora) Community JIRA
  - Weekly Committers Meeting
- Meet Up
  - Islandora Camp
  - RIRI

Concluding Remarks - D - benefits we've gotten - research funds, relationship with fac, community collaboration (eg maps from the Provincial Records and Archives Office), grant funding, learning opportunities for librarians, staff about these cutting edge

Islandora Open Source Community - D -

How do all these pieces get developed? Developing open source software is a community endeavour. There are a many different ways that individuals and institutions can participate. You could test the software and let us know what works and doesn’t work for you. If you’ve got programming skills you could contribute to the code. You could share your experiences with the software and document them. If you’ve got digitized content at your institution and don’t have a place to put it you could use Islandora. You could come visit us on the Island and participate in the Red Island Repository Institute or in Islandora Camp.
Thank you!

Questions?

Contact information:
Donald Moses dmoses@upei.ca
Melissa Belvadi mbelvadi@upei.ca

Questions? Also we’d appreciate any feedback you would like to give us. You are welcome to include your name or be anonymous.
Links

- http://islandora.ca
- https://wiki.duraspace.org/display/ISLANDORA/Islandora
- https://github.com/islandora
- https://jira.duraspace.org/browse/ISLANDORA
- http://groups.google.com/group/islandora
- http://fedora-commons.org/
- http://drupal.org
- http://islandscholar.ca
- http://islandnewspapers.ca
- http://islandimagined.ca
- http://islandlives.ca
- http://vre2.upei.ca/herbarium