Code4Lib New England, 29 May 2015

Collaborative notes! Make them better. Sharing is caring.

http://wiki.code4lib.org/NECode4lib_2015_Home

- Code4Lib New England, 29 May 2015
  - Presentations
    - Metadata Enrichment and Maps (Boston Public Library)
    - Usability, Digital Content Strategy & LibGuides (Sharon Clapp, Central Connecticut State University)
    - Using Omeka to Receive Data from Researchers and to Author Metadata (Stephen Balogh and Andrew Battista, New York University)
    - Visualizing Open Access: building a scalable infrastructure to showcase the reach of MIT research (Matt Bernhardt, MIT Libraries)
    - SCOAP3 TopicHub: a web application to allow users to discover, subscribe to, and obtain automatic delivery of article content from the SCOAP3 repository (Richard Rodgers, MIT Libraries)
    - Getting Started with Regular Expressions (Christine Moulen, MIT Libraries)
  - Lightning Talks
    - Using a script to reclassify a video collection (Steve McDonald / Tufts University)
    - Displaying library hours using Google Spreadsheets and TabletopJS (Matt Bernhardt / MIT)
    - Turning historical texts into data sources (even when they're in Russian) (Jeremy Guillette / Fung Library @ Harvard University)
    - Metadata Quality: Statistics from Digital Commonwealth on what matters in increasing a digital object's discoverability (Steven Anderson / Boston Public Library)
    - Systems & Applications Diagram: Conceptually mapping Library Technology infrastructure for planning and communication (Julia Caffrey / Simmons College)
    - HTML based Digital Signs (Jeremy Prevost / MIT)
    - Process of becoming a data harvester (Jennifer Eustis & Rick Sarvas(?), UConn)
    - Jay Luker/Apache Log Analysis in 5 Minutes with ELK & Docker
  - Discussion Topics
    - 3D Printing and the Law: A raucous uninformed melee where opinionated librarians discuss a topic they do not understand because it has not been settled yet (Edward Iglesias, Central Connecticut State University)
    - Future of Code4LibNE (Jeremy Prevost / MIT)
    - Analytics across the Discovery Environment: trying to understand how patrons move between all these platforms libraries maintain (Matt Bernhardt / MIT)

ArchivesSpace Customization & Development: If you're using ASpace and want to talk customization and development, I'd be interested! (Maura Carbone / Brandeis University)

Presentations

Metadata Enrichment and Maps (Boston Public Library)

Slides: Google Slide Presentation
Digital Commonwealth: statewide library for Massachusetts
web site
but also digitization & hosting services
administered by BPL
supports libraries that may not have the resources to support a repository, do scanning, etc
194,700 items in 440 collections from 123 institutions
They're a DPLA service hub - they ingest smaller libraries' content and DPLA in turn ingests theirs. [slide of scary toothy big fish eating smaller toothy fish eating tiny fish]

Problem: different libraries have different metadata schemes, content standards; enrichment & standardization is required

- e.g. "Books - History" vs. "Books -- History"
- dates. Ugh, dates.
- So they have standardization scripts to put everything into W3CDTF format

Authorities and locations

- Goal: separate data into more atomic forms.
- "Some guy (painter)" → "name=>'Some guy', role=>'Painter'"
- similarly city=>Boston, state=>Massachusetts, geonames=>(id, rdf link, etc.)
- Use Geomash to get coordinates, hierarchical geography, etc.

Further enhancements

- https://github.com/dpla/krikri
- https://github.com/projecthydra-labs/questioning_authority

So what do you do with this?

Blacklight maps

- blacklight = open source RoR gem; search interface to any Solr index (not just hydra!)
- The maps thing lets you add map views for items with geospatial metadata.
- You can browse maps using geospatial metadata as just another facet

Also you can view search results on a map
Uses Leaflet to generate maps

Data sources
Solr field types: `location_rpt`, `string`
You can store coordinates or bounding boxes (`location_rpt`) or GeoJSON
There's a variety of acceptable data formats

BPL, WGBH, and Northeastern are putting together a proposal to host HydraConnect 2016 - do you want to help? Talk to them!

Usability, Digital Content Strategy & LibGuides (Sharon Clapp, Central Connecticut State University)

It started with a usability test. Question: could they get down to One Search Box To Rule Them All?

Students went all over the place trying to navigate Summon - ended up on LibGuides home page, which isn't the same as the library home page..

5-second page test:
What is the purpose of this page?
This turned out to be a hard question to answer!
What's your next step/call to action?
You need to have a call to action before you can measure whether users are successful in achieving the action.
What task is to be done?

So they installed some analytics. Which LibGuides pages were most effective? Exactly the ones web pros would suggest - the ones focused on specific tasks. ("How do I tell the difference between a scholarly publication and a popular magazine? How do I access course reserves?" Stuff like that.) Also LibGuides focused on specific tasks - faculty & librarians designing those collaboratively worked.

Content authoring was convoluted…
LibGuides
WordPress
html/css/PHP
This meant that branding & navigation were inconsistent across library properties and students were confused.
It also limited the opportunities for collaboration.

One option: put everything into LibGuides

- Yay:
  - offloads hosting headaches; reduces need to bug IT
  - librarian-oriented
  - 2.0 is very template-able, and bootstrap is attractive and responsive
  - LTI integration - they can staple Blackboard/Moodle/etc. to it, which opens up options for collaborating with professors
- Boo:
  - cost (though it's not TOO costly)
  - less agile than doing it yourself
  - search/navigation/consistency

Another option: an open source CMS like Drupal or WordPress

- Yay:
  - can do anything
  - can iterate
  - can be part of larger OSS world
    - both helping librarians and students be producers, not just consumers, of the web - cf Mozilla Webmaker
    - and helping broader dev world know more about things like DPLA
  - can train students in useful skills
  - more opportunities for collaboration outside the library
- Boo:
  - you might have to run the servers (though there are some cloud options)
  - Labor overhead
  - Not what librarians want

Existential questions! for business & identity

- What are we supposed to be doing, anyway?
- Are we (web)makers?
- What's our role in web literacy?

Audience feedback

- Roll-your-own CMS is not the most efficient use of your developers when the world has created so many
- People will flee from Dreamweaver/hand editing to LibGuides; it's easier. Then you have to figure out what to do with the abandoned pages.

Using Omeka to Receive Data from Researchers and to Author Metadata (Stephen Balogh and Andrew Battista, New York University)

https://omeka.org/

spatial data files from ESRI, online sources, many researchers

What do they need?

- convenient way for users to submit data
- Easy way for people who aren't metadata experts (though they may be subject experts) to author metadata
- Ways for librarians to clean up metadata

They have a geoblacklight schema
Omeka expandable through plugins, such as to introduce new element sets or modify attributes of existing ones
Easy to query collection using APi

Visualizing Open Access: building a scalable infrastructure to showcase the reach of MIT research (Matt Bernhardt, MIT Libraries)

Slides:
http://www.slideshare.net/morphosis7/visualizing-open-access-2015-code4lib-northeast

Repositories:
https://github.com/MITLibraries/oastats-backend
https://github.com/MITLibraries/oastats-ui

Service:
http://oastats.mit.edu

MIT open access policy adopted March 18, 2009 - this resulted in a rapidly growing repository

They knew they had ~15,000 articles with ~1 million downloads, but they couldn't provide more (or more granular) information. They needed to be able to:

- better communicate with stakeholders about usage
- support the 5-year review that was part of the OA mandate

Project goals:

- author-level, article-level, and aggregated download stats
- incentivize further contributions
- support policy evaluation
- seeekrit internal goal: learn about new platforms

Harvard's repository provided a nice model, but the tech/metadata schema was too different to be adopted wholesale

Two parts to project:

- Data processing pipeline
  - Apache: filter raw server logs for qualifying downloads (from this collection, not bots, etc.)
  - Augment with additional info (e.g. paper author) - python
  - Store in MongoDB
  - Original intent was to directly query MongoDB, but this wasn't performant, so they stapled on some Solr
- Visualization interface
  - Branded like MIT's DSpace
  - Mongo, PHP, DataTables, d3.js, DataMaps

Challenges

- Pipeline
  - Content comes from a variety of places, not just authors, but they want to harvest author names
  - Department names aren't standardized, and papers may be affiliated with various parts of the MIT hierarchy (not just departments)
  - Department names are spelled inconsistently ("Physics", "Department of Physics", "Dept. of Physics")
  - The Media Lab has sixty bajillion subunits
  - Yay for OpenRefine!
  - Author names can be hard to disambiguate
  - Different fields refer to authors differently
  - They use MIT IDs as the unique ID for author data, but they need author names to get back to it
  - They have standardized forms of department names, and a whitelist of recognized forms

They email authors when they are above some (non-embarrassing) download threshold - this has generated 8500 emails. Also advertises DSpace some.

Faculty reception

- lots of excitement - "this is a fantastic tool!"
- A few people confused as to why they got the email, but not many
- "Why not more?" - people who wondered why more of their articles weren't in dspace (i.e. they totally DID incentivize deposits - both individual faculty and department chairs want more of their stuff to be recognized)

How did they do vis-a-vis their goals?

- They built the things
- Anecdotally it's incentivizing downloads - they only sent the faculty email a few weeks ago, so it's hard to tell yet

Policy announcement has not been made

Future work

automated pipeline; run more frequently (every 2 months)
postgres or something more relational than mongo
talk to faculty about making more detailed information public - do they need to keep paper-level data behind an authentication screen, or can they show everything to everybody?
improved UI
improved cataloging in DSpace - catch name variants early and save time down the road

SCOAP3 TopicHub: a web application to allow users to discover, subscribe to, and obtain automatic delivery of article content from the SCOAP3 repository (Richard Rodgers, MIT Libraries)

Repository:
https://github.com/MITLibraries/scoap3hub

Digression: his slide tool (Haiku Deck) lets you type terms into your slides and it'll automatically do a search for CC-licensed images. Sweet! Mind you, the images may not make any sense.

If you have bilateral relationships between content providers and consumers, you will be sad as you scale; each new provider requires a relationship with all the consumers and your number of graph edges becomes catastrophic.

Alternative model is a hub - centralize hub, with SourceN+TargetM (loosely coupled) connections.

Two strategies:

- push (e.g. SWORD)
- pull (e.g. API, called on demand)

Hub can implement various of these (or hybrid) strategies

Our hub harvests new content from SCOAP3 via API

- trial uses MARCXML (other formats possible)

Subscribing consumers can sign up for content via keyword - and then new content flows into their repository automatically. In theory.

Early lesson from this project - useful metadata is patchy and non-uniform (name variations, affiliations expressed in varying terms) - no good source of affiliation metadata, which is central to their project.

Getting Started with Regular Expressions (Christine Moulen, MIT Libraries)

Regex: syntax for defining patterns

14*

the * is a wildcard
this means a 14 followed by 0 or more of any character

ls 14[0-1][0-9]*

[0-1] means any one character in the range [0-1]; similarly for [0-9]
patterns like this are good for looking for filenames within a date range in your directory

mv [b-z]* temp/

will move all files that begin with letters between b and z to your temp/ directory
depending on your system setup, this might be only lowercase or it might be case-insensitive

grep 'MIT01$' sysnos.txt

the $ signifies the end of the line, so this will find lines in sysnos.txt that end in MIT01
^ will match the beginning of the line

In vi, you can use regex with the s/// substitution operator: s/$/MIT01/ will replace line endings with MIT01 (that is, it will add MIT01 to the end of all your lines)

In emacs, Mx-query-replace-regexp

She also uses them a lot in perl - look through a MARC file line-by-line and then have an if statement to do things with lines that match the regexp.

`m/^\d{9}\s260.+/` will find your 260 fields…

- `^` beginning of line
- `\d` a digit
- `{9}` 9 of previous character (aka 9 digits)
- `\s` whitespace
- `260` the MARC tag she wants
- `.+` any character at least once

Other things she does with perl and regexp…

look for deleted records - LDR position 05 = d: `$my_LDR=~ /LDR  L …..d/`

look for e-resource records: `$my_245 =~ /\$\$h\[electronic resource]\]/` (although maybe there should have been a 245 after that first dollar sign). The backslash is used to escape the $ so that it is a literal dollar sign, not a special character.
Look for OCLC numbers: `$my_035 =~ /(\(OCoLC\)\d{8,10})/`
Using substitution to move the meaningful part of an OpenURL closer to the beginning for sorting purposes.
Parsing thesis metadata - single MARC freetext field for degree/year/department needs to be 3 separate fields in DSpace/Dublin Core. This requires processing variant spellings/capitalizations/abbreviated vs non-abbreviated forms.

Lightning Talks

Using a script to reclassify a video collection (Steve McDonald / Tufts University)

Some are directly class-related. They cover lots of ground - foreign! experimental!

They've been stored in closed stacks, but the library is considering major renovations, so they might be able to move the DVDs into open stacks. (They want to get rid of VHS where possible.) Anything in public stacks needs to have LC classification, for consistency with the rest of the collection. So they're just doing that, to be ready by the time the library decides on the shelving situation.

19,000 DVDs to reclassify. They have some student help, but with that volume of work, they have to automate as much as possible or it can't be done.

Step 1: Come up with cataloging policy.

Script written in AutoIT. Emulates keystrokes/mouseclicks, so you can use it even when you don't have a useful API (say, e.g., if you're running Millennium). You feed it a bar code and it and automatically reclassifies what it can, asking for approval or y/n question answers from the humans; refers the records it can't handle to catalogers.

Reclassed over 19,000 records from May to January with minimal impact on regular work. The script turned drudge work into fun work.

Displaying library hours using Google Spreadsheets and TabletopJS (Matt Bernhardt / MIT)

Code library
https://github.com/szendeh/libhours

Implementation w/in Wordpress
https://github.com/mitlibraries-ux/MITlibraries-parent/blob/prod/js/hours-lookup.js

View live:
http://libraries.mit.edu/ (selected libraries today on front page)
http://libraries.mit.edu/barker/ (today's hours for specific libraries)
http://libraries.mit.edu/hours/ (master grid of this week's hours)

Turning historical texts into data sources (even when they're in Russian) (Jeremy Guillette / Fung Library @ Harvard University)

https://github.com/FungDavis/fungHGR
http://fungdavis.github.io/fungHGR/
Slides

Metadata Quality: Statistics from Digital Commonwealth on what matters in increasing a digital object's discoverability (Steven Anderson / Boston Public Library)

Slides: Google Slides Presentation

How do different metadata fields affect discoverability in Digital Commonwealth?

Precoordinated LCSH-style subject headers do not get as many views as non-LCSH-styled headers, or a mix of LCSH and non-LCSH. "Normalized non-LCSH style subjects soundly defeat those items that use the concatenation".

Hosted items get way more views than OAI-harvested items. Hosted items have cleaner, more uniform data - that probably helps.

Geographic subjects seem to win over even the most shared of subject terms, if you have to decide between the two. (Though this might be a function of the user interface.)

Systems & Applications Diagram (Julia Caffrey / Simmons College)

Julia Caffrey, Library Systems Assistant at Beatley (created diagram with then Systems Librarian, Amy Deschenes)
Process of conceptualizing, categorizing, visualizing library systems in use at Beatley
useful for analyzing user needs and seeing how changes in one tool affect changes in others

HTML based Digital Signs (Jeremy Prevost / MIT)

Process of becoming a data harvester (Jennifer Eustis & Rick Sarvas(?), UConn)

"We're going to become a service hub for DPLA!" ...yikes! At this point they didn't even have metadata guidelines. They had an OAI feed to provide data, but didn't have experience harvesting.

Conveniently, people like Europeana have worked on this problem already!
http://repox.sysresearch.org/

Problem: the most readily available repox was Windows-based, and they're a Red Hat shop. The installer will run if you have an emulator, but it doesn't work. There's a github repository, but it's not the same as the windows version, and they'd been doing development against that version. There's a linux installer, but it's broken too. In particular, it couldn't actually create working users file, so no one could log in.

Discrepancies between the windows and linux forks were exciting (although similarities helped them debug linux problems).

Festival of unexpected errors with the install/configure process. For instance, if you try to do XSLT for small transforms, it will helpfully assume that there's *no* difference and just not run in order to save you some time. There's a process scheduler, but it doesn't run when you tell it to - GUI editor doesn't tell you what tasks you have scheduled, and tasks are scheduled in GMT.

Jay Luker/Apache Log Analysis in 5 Minutes with ELK & Docker
https://github.com/lbjay/apache-elk-in-five-minutes

If you want to play along at home, you can, with the instructions at that link!

Relies on Docker; "you can usually just get it from your favorite package manager."

Discussion Topics

3D Printing and the Law: A raucous uninformed melee where opinionated librarians discuss a topic they do not understand because it has not been settled yet (Edward Iglesias, Central Connecticut State University)

Future of Code4LibNE (Jeremy Prevost / MIT)
Desire to keep momentum going

Proposed that we try to schedule two events per year - one that's in the Greater Boston Area and one that happens somewhere else
Discussed possibly trying to arrange for another Code4LibNE event in Fall 2015
Discussed working with Simmons GSLIS to connect Code4LibNE with library school students
Talked about continuing to use Slack and the google group as a means to organize events, and publicizing events through the general Code4Lib listserv

Analytics across the Discovery Environment: trying to understand how patrons move between all these platforms libraries maintain (Matt Bernhardt / MIT)

Platforms discussed
Cognos
Some use Google Analytics (Digital Commonwealth, Naval College)
Digital Commonwealth using async event listeners
    Can listen to lots of things, but what does that mean?
    Tracking what facets get used the most? Put them at the top of the sidebar
        (Format was thought to be popular, but not in actual fact)
Choice about de-emphasizing the under-used, versus moving it more prominently?
Notre Dame, Google Analytics, and Primo

Ivies+ Discovery Day example - poorly-labelled?
"Our 404 page is one of our most popular pages, but we aren't going to link people to it"
We may never be able to get the environment as a whole, unless it is one thing. Won't get quality referral information, for example.
Would like to see a heat map or path map between tools. What are you using, and how did you get there?

What tools do you use? Google Analytics works out of the box (never put into practice)
What do you track with external links? In specialized databases? To specialized databases? (How you use Event Tracking)
Three text strings

How do you track within proprietary databases?
(Serial Solutions?)
COUNTER stats
EZ Proxy

SUSHI (Standard Use Statstics Harvesting Initiative) (URL web protocol)

What else to track?
ILL

ArchivesSpace Customization & Development: If you're using ASpace and want to talk customization and development, I'd be interested! (Maura Carbone / Brandeis University)