

# **When Metadata Collides:**

Lessons on Combining Records from Multiple Repository Systems.

**Presented By:**

Steven Anderson

Boston Public Library (BPL)

[sanderson@bpl.org](mailto:sanderson@bpl.org)

**Link:**

<http://static.digitalcommonwealth.org/OR2014>

# Big Fish, Small Fish

- Great collections exist everywhere: from the largest University to the smallest local library.
- But... what of those smaller institutions?
  - They often don't have the infrastructure to digitize materials.
  - Even if they could, how would anyone find them online?
- And couldn't results from larger institutions benefit from those local materials?

# Digital Commonwealth

(<http://www.digitalcommonwealth.org>)

- “Digital Commonwealth” was founded to create a common state portal.
- Recognizing the small fish needs, BPL sought funding to:
  - Offer free digitization services (including pickup / delivery of materials).
  - Creation of “Metadata Mob”.
  - Shared repository ecosystem. BPL objects equivalent to other hosted objects.

# Digital Commonwealth

(<http://www.digitalcommonwealth.org>)

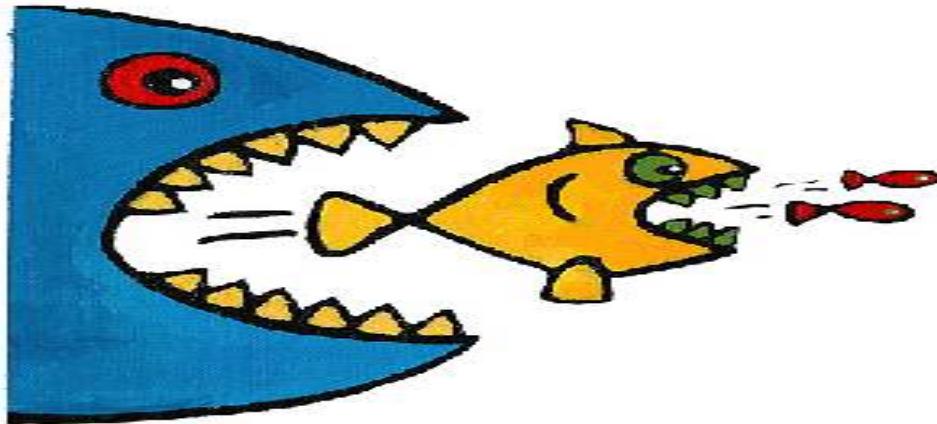
- What about those “big fish”?
  - Most already have their own institutional repositories. So Metadata records are harvested via OAI-PMH in Digital Commonwealth.
  - All records crosswalked into MODS Fedora Objects for more efficient indexing.
  - Metadata is enhanced and normalized so objects can live together in harmony.
  - As we link directly to their system for the object, better collection visibility for them.

- In the end, Cambridge Public Library's WWI plaques (hosted) can exist with Springfield College's WWI posters (OAI harvest) from the same seamless search.



View this item at Springfield College Archives and Special Collections [G](#)

# Bigger Fish



- When we ingest, we standardize and improve data from all of these local sources.
- DPLA (dp.la) can then ingest from us rather than needing separate ingestion scripts for dozens of institutions.
- Our WWI stuff exists with WWI stuff from all across the USA!

# Unexpected / Chaotic Metadata

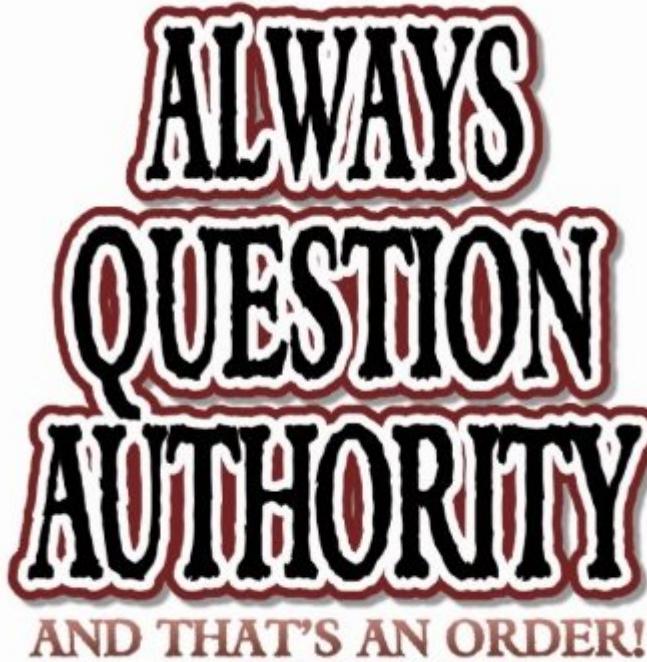
- Data lacks consistency. Following are just a few examples from one sources dc:coverage:
  - 42212N72345W
  - 1895-1905
  - The Town of Holyoke , MA in 1877
  - Springfield Hospital The Training School for Nurses, Springfield, MA; 1943.
  - Springfield Hospital School of Nursing, Springfield, MA; 1973.
  - 42 degrees 04' N 72 degrees 02'
  - 42 36' 00" N, 72 23' 55" W

# More Chaotic Metadata

- Unstructured and non-standard Subjects:
  - Horses--Massachusetts--Hadley--Hockanum
  - Congregational Church in Halifax (Halifax, Mass.)
  - Civil War
- Just plain incorrect fields
  - Coordinates in a “Rights” field.
  - Dates in a “Creator” field (Pre-1970).
  - Notes in “Creator” field (Copied from Book).



- Have to throw away assumptions on what fields should contain.
- Must parse this data. If you just dumped all of these inconsistencies into one's system based on a straight mapping, chaos will ensue.



- “Questioning Authority” Rails Gem:  
[github.com/projecthydra-labs/questioning\\_authority](https://github.com/projecthydra-labs/questioning_authority)
- “Bpl Enrich” Rails Gem:  
[github.com/boston-library/bpl\\_enrich](https://github.com/boston-library/bpl_enrich)

**Language Given:** "French"

```
$ rails c
Loading development environment (Rails 4.0.5)
2.0.0-p247 :001 > BplEnrich::Authorities.parse_language('French')
=> {:uri=>"http://id.loc.gov/vocabulary/iso639-2/fre", :label=>"French"}
2.0.0-p247 :002 > █
```

```
<mods:language>
  <mods:languageTerm authority='iso639-2b'
authorityURI='http://id.loc.gov/vocabulary/iso639-2'
type='text' valueURI='http://id.loc.gov/vocabulary/iso639-
2/fre'>French</mods:languageTerm>
</mods:language>
```

**Name Given: "Sully, François (Photographer)"**

```
$ rails c
Loading development environment (Rails 4.0.5)
2.0.0-p247 :001 > BplEnrich::Authorities.parse_name_for_role('Sully, François (Photographer)')
=> {:name=>"Sully, François", :uri=>"http://id.loc.gov/vocabulary/relators/pht", :label=>"Photographer"}
2.0.0-p247 :002 > █
```

```
<mods:name>
  <mods:role>
    <mods:roleTerm authority='marcrelator'
authorityURI='http://id.loc.gov/vocabulary/relators'
type='text'
valueURI='http://id.loc.gov/vocabulary/relators/pht'>Photographer</mods:roleTerm>
  </mods:role>
  <mods:namePart>Sully, François</mods:namePart>
</mods:name>
```

# Geographic Enhancement



- “Bplgeo” Rails Gem:  
<https://github.com/boston-library/Bplgeo>
  - Add coordinates and standardized geographic hierarchy.

# Example Parsing LCSH Subject

```
$ rails c
Loading development environment (Rails 4.0.4)
2.0.0-p247 :001 > Bplgeo.parse('Cranberry industry--Massachusetts--Yarmouth', true)
=> {:original_term=>"Cranberry industry--Massachusetts--Yarmouth", :standardized_term=>"Yarmouth,Massachusetts", :country_part=>"United States", :state_part=>"Massachusetts", :city_part=>"Yarmouth", :tgn=>{:id=>"2051052", :original_string_differs=>false}, :geonames=>{:id=>"4956335", :original_string_differs=>false}}
2.0.0-p247 :002 > █
```

# Getting TGN or Geonames entry

```
$ rails c
Loading development environment (Rails 4.0.4)
2.0.0-p247 :001 > Bplgeo::TGN.get_tgn_data('2051052')
=> {:coords=>{:latitude=>"41.70", :longitude=>"-70.2167", :combined=>"41.70,-70
.2167"}, :hier_geo=>{:city=>"Yarmouth", :county=>"Barnstable", :state=>"Massachu
setts", :country=>"United States", :continent=>"North and Central America"}, :no
n_hier_geo=>nil}
2.0.0-p247 :002 >
2.0.0-p247 :003 >   Bplgeo::Geonames.get_geonames_data('4956335')
=> {:coords=>{:latitude=>"41.70567", :longitude=>"-70.22863", :combined=>"41.70
567,-70.22863", :box=>{:west=>"-70.25624", :north=>"41.72627", :east=>"-70.20102
", :south=>"41.68506"}}, :hier_geo=>{:area=>"Earth", :cont=>"North America", :pc
li=>"United States", :adm1=>"Massachusetts", :adm2=>"Barnstable County", :ppl=>"Y
armouth"}}
2.0.0-p247 :004 > █
```

# So Given Nothing More Than: “Faneuil Hall, Boston, Massachusetts”



```
<mods:subject authority='tgn' valueURI='7013445'>
  <mods:hierarchicalGeographic>
    <mods:continent>North and Central
    America</mods:continent>
    <mods:country>United States</mods:country>
    <mods:state>Massachusetts</mods:state>
    <mods:county>Suffolk</mods:county>
    <mods:city>Boston</mods:city>
  </mods:hierarchicalGeographic>
  <mods:cartographics>
    <mods:coordinates>42.35,-71.05</mods:coordinates>
  </mods:cartographics>
</mods:subject>
<mods:subject>
  <mods:geographic>Faneuil Hall, Boston,
  Massachusetts</mods:geographic>
  <mods:cartographics>
    <mods:coordinates>42.3600619,-71.056103
  </mods:coordinates>
  </mods:cartographics>
</mods:subject>
```



- Bpl Enrich supports parsing of over 90 date formats:
  - 192-?].
  - 1943 (Spring)
  - [between 1883 and 1910]
  - 09/1962 – 10/1962
  - [20th century.]

**Date Given:** "late 1960s (Easter)"

```
$ rails c
Loading development environment (Rails 4.0.5)
2.0.0-p247 :001 > BplEnrich::Dates.standardize('late 1960s (Easter)')
=> {:date_range=>{:start=>"1967", :end=>"1969"}, :date_qualifier=>"approximate"
, :date_note=>"late 1960s (Easter)"}
2.0.0-p247 :002 > █
```

```
<mods:originInfo>
  <mods:dateCreated encoding='w3cdtf' keyDate='yes'
  point='start'
  qualifier='approximate'>1967</mods:dateCreated>
  <mods:dateCreated encoding='w3cdtf' point='end'
  qualifier='approximate'>1969</mods:dateCreated>
</mods:originInfo>
<mods:note type='date'>late 1960s (Easter)</mods:note>
```

# Caveats

- Do we do all of these for all records?
  - Short answer: no. One must review the source content first for what is appropriate.
- Still requires constant spot checking of results whenever new collections are processed.

# Geographic Parsing Caveat

## The World according to America



# **Contacts / Final Links**

**Steven Anderson (sanderson@bpl.org)**

**Eben English (eenglish@bpl.org)**

**Github:**

<https://github.com/boston-library>

**Repository:**

<https://www.digitalcommonwealth.org>

**Presentation:**

<http://static.digitalcommonwealth.org/OR2014>

# Image Credits

- **Slide 6:** <http://www.stripersonline.com/t/818407/tuna-eating-striper>
- **Slide 7:** <http://nyancat.wikia.com/wiki/Home>
- **Slide 9:** <http://www.philly.com/philly/blogs/trending/Such-Wow-Everything-you-need-to-know-about-Doge.html>
- **Slide 10:** [http://www.zazzle.com/always\\_question\\_authority\\_post\\_card-239195726898245712](http://www.zazzle.com/always_question_authority_post_card-239195726898245712)
- **Slide 13:** <http://hbi.ucalgary.ca/news-stories/where-am-i-hbi-researcher-featured-nature-things>
- **Slide 18:** <http://mbadiscussions.com/last-date-issue-cat-application-form-123.html>
- **Slide 21:**  
<http://s90.photobucket.com/user/xxxTakaraxxx/media/america.gif.html>
- (All others from DigitalCommonwealth.org)