Enhancing descriptive metadata with geospatial properties

Christina Harlow, @cm_harlow
Increasing Geospatial Awareness Workshop
DLF Forum 2015

Slides, Examples, & Other docs

http://bit.ly/DLF15geowkshop

github.com/cmh2166/DLF15GeoMetadata

Agenda

1. Introduction
2. Existing Best Practices
3. Standards & Vocabularies
4. Tools & Scripts

Goals for This Workshop:

1. Take awesome work for geospatial data
2. Explore how we can leverage in descriptive metadata for non-geospatial objects
3. See how it changes our work
4. Begin some broad best practices

Why This Workshop?

Build new methods - or make consistent application of existing methods - for metadata supporting enhanced geographic discovery & identification of all resources in discovery layer

Is This a Workshop?

- Follow along
- **Give feedback**
- Tinker with examples
- Try out some tools, methods for generating
- Discuss building interoperability
- Please interrupt with questions, additions, corrections

Existing Best Practices #1

Mountain West Digital Library (MWDL)

- http://bit.ly/MWDLgeo
- Expands MWDL DC Application Profile
- Focus on DC (core & terms) in XML
- Primarily for MWDL partner institutions
- Question of Authorities

Best Practices: MWDL cont.

- Geonames preferred authority
- Use dct:spatial, not dc:coverage
- Recommended dct:spatial usage:

```
place name hierarchy, URI, Lat., Long. ; [repeat]
```

```
Phoenix, Maricopa County, Arizona, United States,
http://sws.geonames.org/5308655/; Aurora (historical),
Mineral County, Nevada, United States,
http://sws.geonames.org/5499519/;
```

Best Practices: MWDL cont.

- Coordinates: decimal degrees only
- Put in (or map to) dct:spatial
- Other geospatial metadata => unmapped fields
- These fields get parsed in Primo

```
Boise, Ada County, Idaho, United States,
http://sws.geonames.org/5586437/, 43.6135, -116.20345
```

Best Practices: MWDL cont.

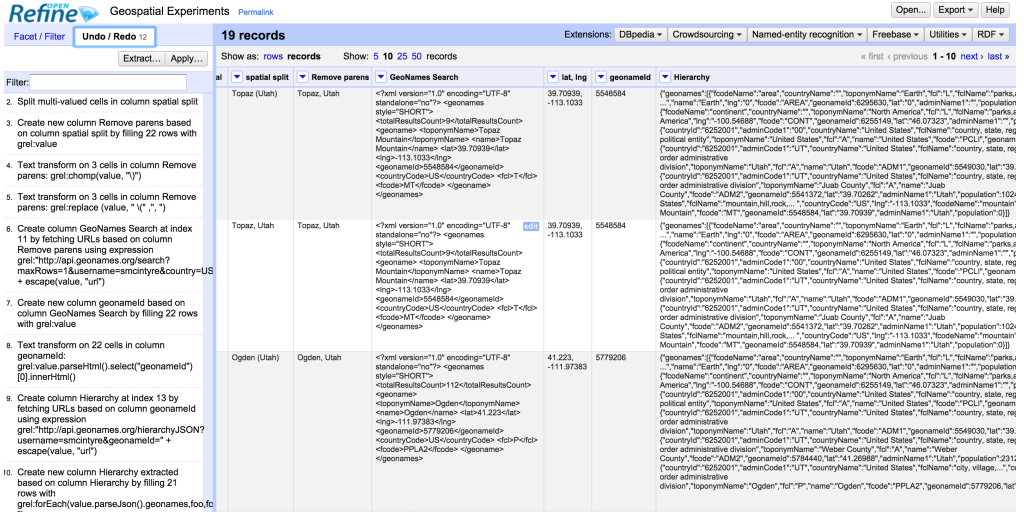
- Develop gazetter for regional names not in Geonames?
- Reviewing ways of handling point & box coordinates in DC/XML
- Working with OpenRefine reconciliation options

MWDL OpenRefine Process

1. Pull data into OpenRefine
2. Hit Geonames API for each term
3. Return from XML output for top match into OpenRefine
4. Parse that XML return in OpenRefine for preferred term, hierarchy, URI, coordinates

Link to their test OpenRefine workflow

MWDL Sample Data/Process



Existing Best Practices #2

MODS (kinda)

Existing options:

- originInfo/place/placeTerm
- subject/geographic/*
- subject/cartographics/*
- subject/hierarchicalGeographic/*

MODS:placeTerm

originInfo/place/placeTerm

- Heavily MARC-related
- Validates only with authority=ISO3166 or MARC Country Codes
- Can use preferred authorities' valueURI
- RDA: transcribe as found

MODS:placeTerm

```
<originInfo>
  <place supplied="yes">
    <placeTerm type="text" valueURI=
    "http://id.loc.gov/authorities/names/n79007751">
    New York (N.Y.)
    </placeTerm>
  </place>
  <dateCreated encoding="edtf" keyDate="yes">
  1862-07-28
  </dateCreated>
</originInfo>
```

MODS:geographic

subject/geographic

- Capture authority at subject level
- LC or TGN, growing use of Geonames
- Can use for LCSH complex subject strings
- Can couple with MODS:cartographics

MODS:geographic

```
<subject>
  <geographic authority="naf" valueURI=
  "http://id.loc.gov/authorities/names/n78095779">
  Memphis (Tenn.)
  </geographic>
  <cartographics>
    <coordinates>
    35.14944N, 90.04889W
    </coordinates>
  </cartographics>
</subject>
```

MODS:cartographic

subject/cartographic/*

- Gives us: mods:coordinates, mods:scale, mods:projection
- Can be extended with non-MODS elements (as of 3.6)

MODS:coordinates

- subchild of MODS:cartographic
- no encoding attribute
- repeating statements is way of handling points, lines, polygons, so:
  - 1 coordinates statement == point
  - 2 coordinates statements == line
  - n coordinates statements == n-sided polygon, points supplied in polygon-traversal order

MODS:scale, projection

- subchildren of MODS:cartographic
- no type attributes
- in speaker's experience used rarely -- mainly MARC records transform for cartographic objects

MODS:cartographic

Similar to previous example...

```
<subject>
  <geographic authority="geonames" valueURI=
  "http://sws.geonames.org/4613427/">
  Chota (historical)
  </geographic>
  <cartographics>
    <coordinates>
    35.55508, -84.12991
    </coordinates>
  </cartographics>
</subject>
```

MODS:cartographic

As of 3.6...

```
<subject>
<geographic>Madison County (Tenn.)</geographic>
<cartographics>
  <cartographicExtension displayLabel="geo"
  xmlns:rdf="..." xmlns:dcterms="..."
  xmlns:rdfs="..." xmlns:dc="...">
    <rdf:RDF >
      <rdf:Description rdf:about="objectURI">
        <dc:spatial>
          <dcterms:Point>
            <rdfs:label>Madison County</rdfs:label>
            <rdf:value>
            name=Madison County; east=35.60814;
            north=-88.83847
            </rdf:value>
          ...
```

MODS:geographicCode

- Like MODS:placeTerm, focused on country
- Only validates with marcgac, marccountry, iso3166
- Redundant possibly with use of valueURIs (in my opinion/limited use cases)

MODS:hierarchicalGeographic

subject/hierarchicalGeographic/*

- Gives us many different levels
- Questions of Authorities

With 3.6:

- Can indicate new place types, hierarchy levels (via attributes)
- Can indicate relevant time period
- Sorry Canada, 'Province' being depreciated for 'State'

MODS:hierarchicalGeographic

```
<subject>
  <hierarchicalGeographic>
    <country>United States</country>
    <state>Tennessee</state>
    <region regionType="range">
      Great Smoky Mountains
    </region>
    <county>Blount County</county>
    <city period="1800-1914">Cades Cove</city>
  </hierarchicalGeographic>
</subject>
```

What I do

- Subjects:
  - geographic & cartographics
  - geographic has LoC valueURI, form
  - cartographic from Geonames
  - coordinates are decimal form
  - avoid complex subject strings
- hierarchical avoided unless legacy data
  - If equal interest in town + county, e.g., geographic repeated
- geographicCode not used
- extensions used when cartographic objects

Existing Best Practices #3

DPLA MAP v.4

- more method then best practice
- http://dp.la/info/developers/map/
- JSON-LD
- Uses dcterms:spatial with object belonging to dpla:place
- Extends Europeana Data Model- edm:Place

DPLA MAP v.4

dpla:Place in DPLA MAP:

- Name > skos:prefLabel
- Latitude > wgs84_pos:lat
- Longitude > wgs84_pos:long
- Altitude > wgs84_pos:alt
- Geometry > geojson:geometry
- Parent Feature > gn:parentFeature
- Country Code > gn:countryCode
- … (other properties)

DPLA MAP v.4

dpla:Place Depreciated in DPLA MAP v.4:

- City > dpla:city
- State > dpla:state
- County > dpla:county
- Region > dpla:region
- Country > dpla:country
- Coordinates > wgs84_pos:lat_long

DPLA dpla:Place

```
sourceResource: {
  title: "SUMMARY OF AIR FORCE EXPERIENCE, GEN OTTO
  P. WEYLAND, WASHINGTON, D.C",
  spatial: [
    {
    name: "Washington (D.C.)",
    state: "District of Columbia",
    coordinates: "38.8903694153, -77.0319595337",
    country: "United States"
    }
  ],
...
```

Standards & Vocabularies:

ISO3166

Country codes. Really, that's it. Very MARC-centric.

Standards & Vocabularies:

WGS84

World Geodetic System 1984. Coordinates, Altitude mainly (for library metadata data)

Standards & Vocabularies:

DCMI Point, Box Datatypes

Could be used in RDF-situations to further refine objects of dcterms:spatial predicates

Standards & Vocabularies:

DCMI Point, Box Datatypes

```
<rdf:RDF >
  <rdf:Description rdf:about="objectURI">
    <dc:spatial>
      <dcterms:Point>
        <rdfs:label>Madison County</rdfs:label>
        <rdf:value>
        name=Madison County; east=35.60814;
        north=-88.83847
        </rdf:value>
      </dcterms:Point>
    </dc:spatial>
  </rdf:Description>
</rdf:RDF>
```

Standards & Vocabularies:

DCMI Point, Box Datatypes

```
@prefix dc11: <http://purl.org/dc/elements/1.1/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf:
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://njh.me/objectURI> dc11:spatial [
    a dc:Point ;
    rdfs:label "Madison County" ;
    rdf:value """
            name=Madison County; east=35.60814;
            north=-88.83847
          """
  ] .
```

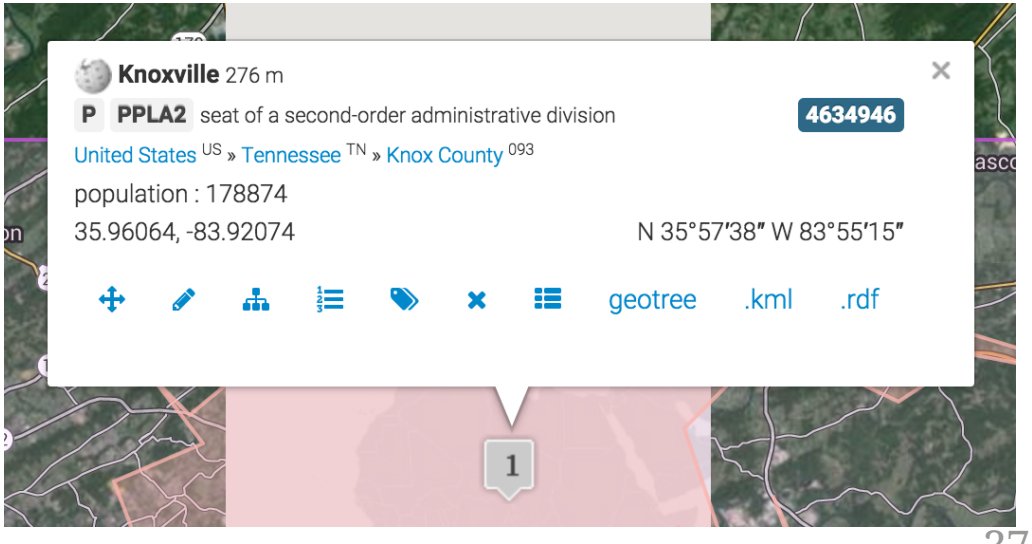
Standards & Vocabularies:

Geonames

Becoming the favorite at present. Offers coordinates, hierarchy, RDF records for entities. PrefLabels don't play well with LoC headings.

Standards & Vocabularies:

Geonames



Standards & Vocabularies:

Getty TGN

Covers hierarchies; similar term forms to AAT; Available as LOD via SPARQL Endpoint but not API; Gives coordinates with datatypes

Standards & Vocabularies:

Getty TGN

```
@prefix gvp: <http://vocab.getty.edu/ontology#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix tgn: <http://vocab.getty.edu/tgn/> .
@prefix wgs: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
...
tgn:7013841 a gvp:Subject , skos:Concept , ... ;
  rdfs:label "Knoxville" , ... ;
  ...
  gvp:broader tgn:2001883 ;
  ...
tgn:7013841-place a schema:Place , wgs:SpatialThing ;
  wgs:lat "35.95"^^xsd:decimal ;
  wgs:long "-83.9167"^^xsd:decimal ;
  ...
  schema:geo tgn:7013841-geometry .
```

Standards & Vocabularies:

id.loc.gov

Pulling in geographic information from GeoNames. Captured some coordinates, constantly under review for further linking (base MARC, transformed into SKOS/RDF, MADS/RDF).

Standards & Vocabularies:

id.loc.gov

Knoxville (Tenn.)

URI(s)
- http://id.loc.gov/authorities/names/n79109786

Instance Of
- MADS/RDF Geographic
- MADS/RDF Authority
- SKOS Concept

Scheme Membership(s)
- Library of Congress Name Authority File

Collection Membership(s)
- Names Collection - Authorized Headings
- LC Names Collection - General Collection

Sources
- found: Its A detailed itemized statement ... 1916-18.
- found: GeoNames, algorithmically matched, 2009 (ppl; 35°57'38"N 083°55'15"W)

Standards & Vocabularies:

id.loc.gov

```
@prefix ns0: <http://www.loc.gov/mads/rdf/v1#> .

<http://id.loc.gov/authorities/names/n79109786>
  a ns0:Geographic, ns0:Authority> ;
  ns0:authoritativeLabel "Knoxville (Tenn.)"@en ;
  ns0:isMemberOfMADSScheme
    <http://id.loc.gov/authorities/names> ;
  ns0:hasSource [
    a ns0:Source ;
    ns0:citation-source
      "GeoNames, algorithmically matched, 2009" ;
    ns0:citation-note
      "(ppl; 35°57’38”N 083°55’15”W)"@en ;
    ns0:citation-status "found"
  ] .
```

Standards & Vocabularies:

German National Library

```
<datafield tag="034" ind1=" " ind2=" ">
  <subfield code="d">W 000 07 32</subfield>
  <subfield code="e">W 000 07 32</subfield>
  <subfield code="f">N 051 30 30</subfield>
  <subfield code="g">N 051 30 30</subfield>
  <subfield code="2">geonames</subfield>
  <subfield code="0">
    (uri)http://sws.geonames.org/2643743
  </subfield>
  <subfield code="9">A:agx</subfield>
</datafield>
<datafield tag="034" ind1=" " ind2=" ">
  <subfield code="d">W000.125740</subfield>
  <subfield code="e">W000.125740</subfield>
  <subfield code="f">N051.508530</subfield>
  ...
```

Tools & Scripts: BoundingBox

By Klokan (Swiss company that focuses on Geodata), takes bounding box & encodes for choice of MARC, DC, few other schema/frameworks:

http://boundingbox.klokantech.com/

Tools & Scripts: OpenRefine

OpenRefine Geonames Reconciliation Service

Extended with LCparse, Stanford local Gazetteer for better id.loc.gov <=> Geonames matchines

Tools & Scripts: Catmandu

- getJson module with GeoNames API
- LDF/Aggregated RDF Reconciliation
  - DBpedia
  - id.loc.gov
  - TGN

Links + Contact

http://bit.ly/DLF15geowkshop

github.com/cmh2166/DLF15GeoMetadata

@cm_harlow, cmharlow@gmail.com