Imposing Structure

Controlled Vocabularies + Ontologies

What is a controlled vocabulary?

- A controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.

  -The Getty (http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/what.pdf)

150 =
"Authorized" or preferred form of subject heading

450 = Variant terms

550 = Related terms

LC control no.: sh2007003708
LCCN Permalink: https://lccn.loc.gov/sh2007003708
HEADING: Transgender people
000 02390cz a2200385n 450
001 7193186
005 20160701055902.0
008 070530i‖ anannbabn ‖a ana
010 __ ‖a sh2007003708
035 __ ‖a (DLC)7193186
035 __ ‖a (DLC)sh2007003708
035 __ ‖a (DLC)359446
035 __ ‖a (DLC)7323129
035 __ ‖a (DLC)sp2007003708
035 __ ‖a (DLC)367975
040 __ ‖a DLC ‖b eng ‖c DLC ‖d WaU ‖d DLC
150 __ ‖a Transgender people
450 __ ‖a TG people
450 __ ‖a TGs (Transgender people)
450 __ ‖a Trans-identified people
450 __ ‖a Trans people
450 __ ‖a Transgender-identified people
450 __ ‖a Transgendered people
450 __ ‖a Transgenders
450 __ ‖a Transpeople
550 __ ‖w g ‖a Sexual minorities

Problems with controlled vocabularies

- Often Anglo-centric and patriarchal

- Non-intuitive language

- Created mostly by library communities for library systems and applications

- Can be very tightly controlled (particularly by national libraries) and difficult to change

Controlled vocabularies: Library of Congress

- LCSH = Library of Congress Subject Headings

- LCNAF = Library of Congress Name Authority File (both LCSH & LCNAF at: http://authorities.loc.gov/)

- id.lov.gov = LCSH and LCNAF in linked data form (not comprehensive & doesn't do well with pre-coordinated subject headings; we'll get more into this later)

Exclusive: New Bill Will Require Library of Congress to Continue Use of 'Illegal Alien' in Subject Headings

Apr. 13, 2016 6:31am | Kaitlyn Schallhorn

Share This   Tweet This

While the Library of Congress effectively scrubbed the terms "aliens" and "illegal aliens" from its subject headings last month, legislation set to be introduced Wednesday could change that.

The short bill doesn't require the Library of Congress to unequivocally use the phrases; rather it mandates that the research library "continue using the term 'illegal alien' just as they were previously," Rep. Diane Black (R-Tenn.) told TheBlaze Tuesday.

"The Librarian of Congress shall retain the headings 'Aliens' and 'Illegal Aliens' as well as related headings in the Library of Congress Subject Headings in the same manner as the headings were in effect during 2015," the short text of the bill, provided exclusively to the TheBlaze ahead of its introduction, reads.

Controlled vocabularies: Library and Archives Canada

- CSH = Canadian Subject Headings (http://www.bac-lac.gc.ca/eng/services/canadian-subject-headings/Pages/canadian-subject-headings.aspx)

- RVM = Repertoire des vedettes-matiére (French equivalents of CSH http://rvmweb-v2.bibl.ulaval.ca/)

- Similar problems as LCSH with non-intuitive language, outdated language, etc

Controlled vocabularies: VIAF & ISNI

VIAF = Virtual International Authority File (https://viaf.org/)

- Less Anglo-centric than LC; includes authorities from libraries world-wide

ISNI = International Standard Name Identifier (http://isni.org/)

- ISO standard; creates persistent unique identifiers for names

Controlled vocabularies: The Getty & RBMS

- These vocabularies are great for very granular description, particularly based on format

- The Getty includes AAT (Art & Architecture Thesaurus) & TGN (Thesaurus of Geographic Names) - http://www.getty.edu/research/tools/vocabularies/index.html

- RBMS = ALA's Rare Books & Manuscript Section vocabulary - http://rbms.info/vocabularies/

Which vocabularies should I use?

- it depends on the needs of your project (again, make sure your metadata is being useful!)

- Consider …

  - What are the needs of project curators?

  - What are the needs of potential users?

  - What are the needs of systems this metadata will interact with?

And Remember…

- If existing controlled vocabularies don't work for your project, you can make your own

- This can be especially helpful when working with historical data that requires making curatorial decisions

And Remember…

- If existing controlled vocabularies don't work for your project, you can make your own

- This can be especially helpful when working with historical data that requires making curatorial decisions

Exercise 1:
Create a controlled vocabulary using historical data

Ontologies

My definition: an artificially constrained intellectual universe in which terms are created and defined to model concepts and their relationships. If you want to express something in machine-readable terms, you have to draw lines around abstract ideas: ontologies define the lines. So, in a sense, controlled vocabularies are also ontologies.

A more technical definition: "An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them." — Natalya F. Noy & Deborah L. McGuinness

Why should I use ontologies?

From Ontology Development 101: A Guide to Creating Your First Ontology (http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html)

- "To share common understanding of the structure of information among people or software agents

- To enable reuse of domain knowledge

- To make domain assumptions explicit

- To separate domain knowledge from the operational knowledge

- To analyze domain knowledge"

Or in a nutshell: To make it easier for machines to understand your data

Common Ontologies

- FOAF = Friend of a Friend (http://www.foaf-project.org/)

- SKOS = Simple Knowledge Organization System (https://www.w3.org/2004/02/skos/)

- OWL = W3C Web Ontology Language (https://www.w3.org/OWL/)

- RDFS = Resource Description Framework Schema (https://www.w3.org/TR/rdf-schema/)

Note: These are expressed in RDF, which we will talk about in-depth tomorrow

Example: A representation of me using FOAF. This points to the FOAF namespace and includes RDF mappings. So if the computer didn't understand FOAF, it could still understand the RDF
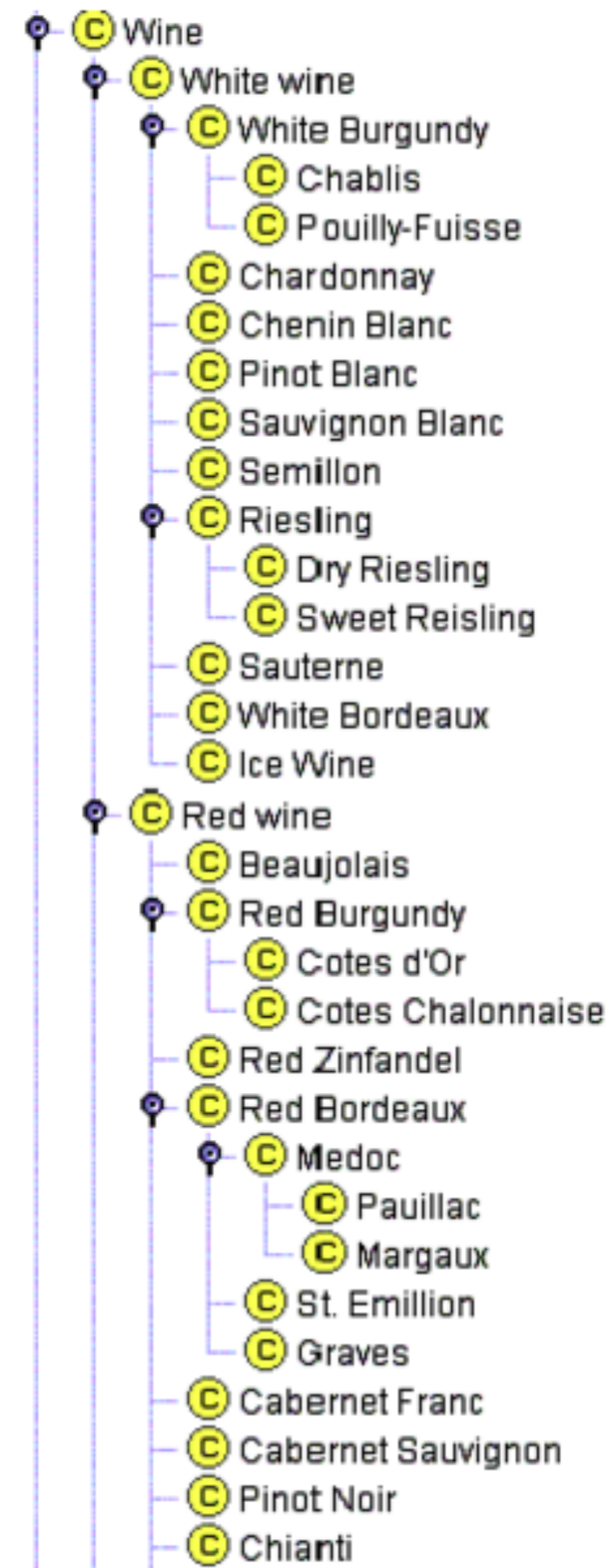
```
<foaf:Person rdf:about="#carolynhansen"
xmlns:foaf="http://xmlns.com/foaf/0.1/">
<foaf:name>Carolyn Hansen</foaf:name>
<foaf:homepage rdf:resource="http://carolynhansen.org" />
<foaf:img rdf:resource="/images/me.jpg" />
</foaf:Person>
```
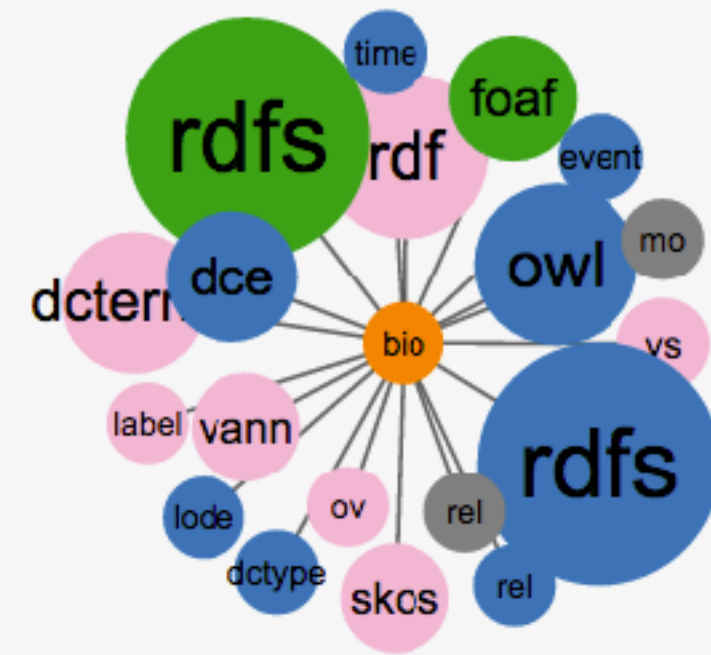
Ontologies are based on Classes & Properties

class: you establish a class so that resources you're working with can inherit a set of characteristics that are pervasive enough across a set of resources that it's worth creating a definition that can be shared

properties: describes the relationships between subject resources and object resources

Image: Classes of wine listed hierarchically. From: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html (**great resource for getting started with ontologies)

- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti

A DHish example: BIO http://vocab.org/bio/

Metadata Extends Specializes

Generalizes Has Equivalences with

Has Disjunction with Imports

Image from: http://lov.okfn.org/dataset/lov/vocabs/bio

A Project-Based Example

https://mapoflondon.uvic.ca/
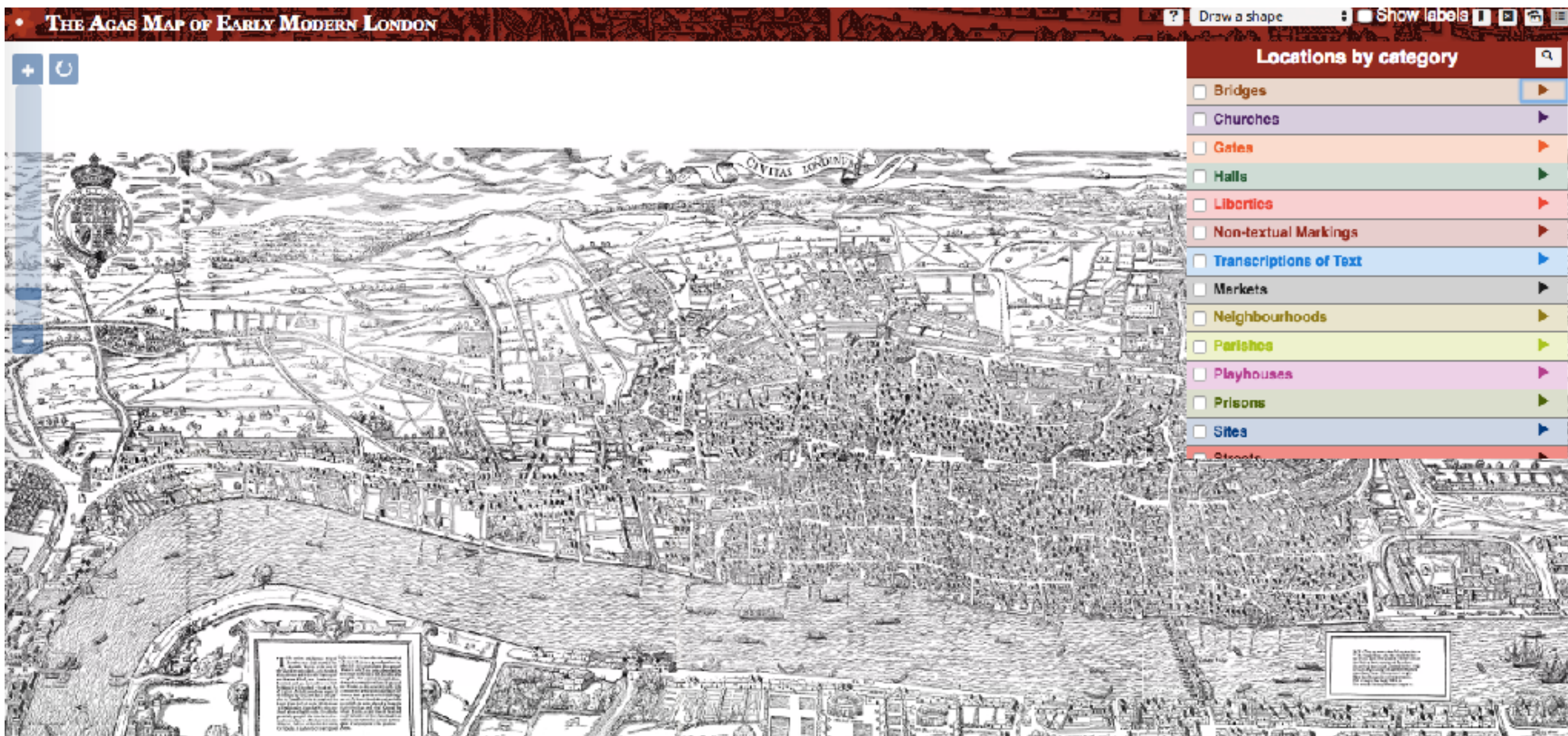
Locations by category

🔍

- [ ] **Bridges** ▼
  - [ ] **Cow Bridge, Smithfield** ◎
  - [ ] **Fleet Bridge** ◎
  - [ ] **Holborn Bridge** ◎
  - [ ] **London Bridge** ◎
  - [ ] **Queen's Bridge** ◎
- [ ] **Churches** ▶
- [ ] **Gates** ▶
- [ ] **Halls** ▶
- [ ] **Liberties** ▶
- [ ] **Non-textual Markings** ▶
- [ ] **Transcriptions of Text** ▶
- [ ] **Markets** ▶

XML for Holborn Bridge

```xml
<profileDesc>
  <textClass>
    <catRef scheme="mdt:molDocumentTypes" target="mdt:mdtBornDigital"/>
    <catRef scheme="mdt:molDocumentTypes" target="mdt:mdtEncyclopediaLocationStreet"/>
    <catRef scheme="mdt:molDocumentTypes" target="mdt:mdtEncyclopediaLocationBridge"/>
  </textClass>
  <abstract>
    <p>
      <ref target="mol:HOLB3">Holborn Bridge</ref>
      or
      <ref target="mol:HOLB3">Oldboorne bridge</ref>
      (
```
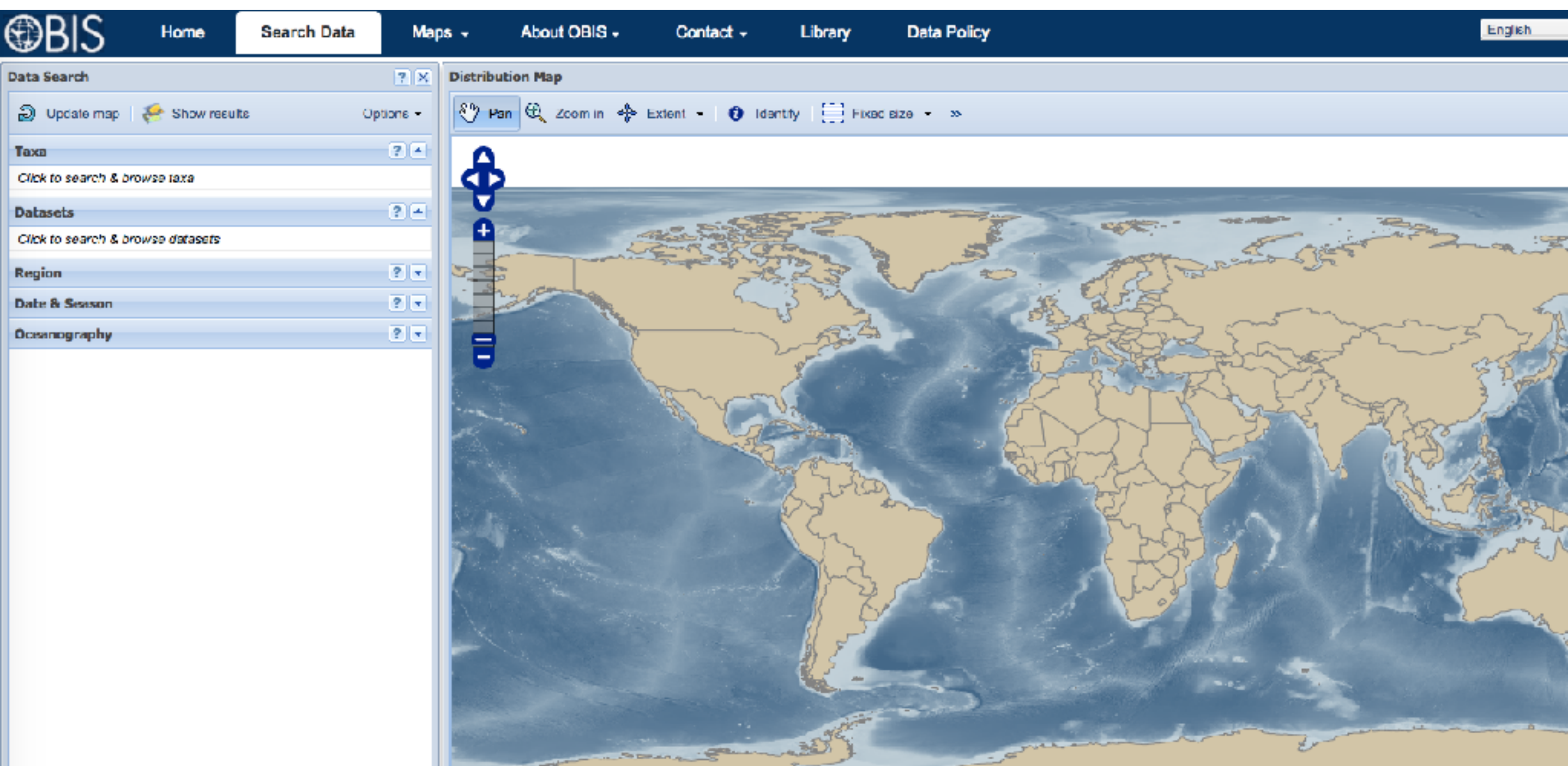
A robust ontology could map this to SKOS, DCterms, etc., improving metadata reuse & interoperability, which helps people build applications using your data

Building applications using ontologies

http://www.iobis.org/

A good tool for making robust ontologies:

protege:
http://protege.stanford.edu/

Note: You can download the application or use the Web-hosted version. If you're interested in working with Protege, there will be time on Thursday/Friday.

Exercise 2:
Create an ontology using historical data