Ontologies for data integration

MARCO ROOS
27 SEPTEMBER, 2016, ROME

4th International Summer School on: “Rare Disease and Orphan Drug Registries”, ISS-CNMR, Rome, Italy
Semantic Reference model for rare disease data integration

External source: Semantic reference model for rare disease data integration

References Orphanet Rare Disease Ontology (ORDO), Human Phenotype Ontology (HPO), Biobank ontology (OBIB), ontology of bioinformatics data types (EDAM)
Ontology

Dates back to Aristotle

Realist theory of categories

Central role for organisms

Porphyrian tree
What is Ontology?

Webster Dictionary:

- D1: a branch of metaphysics concerned with the nature and relations of being;
- D2: a particular theory about the nature of being or the kinds of existents;
- D3: a theory concerning the kinds of entities and specifically the kinds of abstract entities that are to be admitted to a language system;
What is Ontology?

Webster Dictionary:

- D1: a branch of metaphysics concerned with the nature and relations of being.
- D2: a particular theory about the nature of being or the kinds of existents.
- D3: a theory concerning the kinds of entities and specifically the kinds of abstract entities that are to be admitted to a language system.

What about managing and integrating registry data?
What is the role of ontologies for managing registry data?

Rare disease summer school participants say...

□ ...

□ ...
<table>
<thead>
<tr>
<th>Disease</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial deletion of the long arm of chromosome 14</td>
<td>Frequent seizures in early childhood</td>
</tr>
<tr>
<td>Partial deletion of q14</td>
<td>Mild to moderate developmental delay and difficulty with learning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disease</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;ORDO URI for “Partial deletion of the long arm of chromosome 14”&gt;</code></td>
<td><code>&lt;HPO URI for “Frequent seizures in early childhood”&gt;</code></td>
</tr>
<tr>
<td><code>&lt;ORDO URI for “Partial deletion of q14”&gt;</code></td>
<td><code>&lt;HPO URI for “Mild to moderate developmental delay and difficulty with learning”&gt;</code></td>
</tr>
</tbody>
</table>
## Chromosome 14

Partial deletion of the long arm of chromosome 14

<table>
<thead>
<tr>
<th>Disease</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial deletion of the long arm of chromosome 14</td>
<td>0.01</td>
</tr>
<tr>
<td>Partial deletion of q14</td>
<td>0.001</td>
</tr>
</tbody>
</table>

### Disease Ontologies as Source for Data Types

<table>
<thead>
<tr>
<th>ORDO URI for Disease</th>
<th>EFO URI for Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;ORDO URI for “Partial deletion of the long arm of chromosome 14”&gt;</td>
<td>0.01</td>
</tr>
<tr>
<td>&lt;ORDO URI for “Partial deletion of q14”&gt;</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Disease Frequency
### Ontologies to make data linkable

<table>
<thead>
<tr>
<th>Disease</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial deletion of the long arm of chromosome 14</td>
<td>Frequent seizures in early childhood</td>
</tr>
<tr>
<td>Partial deletion of q14</td>
<td>Mild to moderate developmental delay and difficulty with learning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>&lt;ORDO URI for class Disease&gt;</th>
<th>&lt;HPO URI for class Phenotype&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;ORDO URI for “Partial deletion of the long arm of chromosome 14”&gt;</td>
<td>&lt;HPO URI for “Frequent seizures in early childhood”&gt;</td>
</tr>
<tr>
<td>&lt;ORDO URI for “Partial deletion of q14”&gt;</td>
<td>&lt;HPO URI for “Mild to moderate developmental delay and difficulty with learning”&gt;</td>
</tr>
</tbody>
</table>
Ontologies are a source of well-defined, machine readable identifiers (for values and classes)
Guideline 1

Look for concepts in *existing* ontologies
Where to find ontology concepts

1. NCBO Bioportal
   http://bioportal.bioontologies.org/

2. EBI Ontology Lookup Service
   http://www.ebi.ac.uk/ols/beta/
BioPortal

Class Search
Search for a class in multiple ontologies

blue iris

advanced options

Blue iris - Systematized Nomenclature of Medicine - Clinical Terms (SNOMEDCT)
http://purl.bioontology.org/ontology/SNOMEDCT/301952009
details - visualize

Blue iris - Read Codes, Clinical Terms Version 3 (CTV3) (RCD)
http://purl.bioontology.org/ontology/RCD/8a88f
details - visualize - 2 more from this ontology

Iris spuria - National Center for Biotechnology Information (NCBI) Organismal Classification (NCBITAXON)
http://purl.bioontology.org/ontology/NCBITAXON/93026
details - visualize - 1 more from this ontology

Additional References from other Ontologies

Iris germanica - Eagle-1 Research Resource Ontology (ERO)
http://purl.obolibrary.org/obo/NCBITaxon_34205
details - visualize - 2 more from this ontology

Zygonyx iris - NeuroMorpho.Org species ontology (NMOSP)
http://purl.obolibrary.org/obo/NCBITaxon_451506
details - visualize
### Systematized Nomenclature of Medicine - Clinical Terms

**Details**

<table>
<thead>
<tr>
<th>Preferred Name</th>
<th>Blue iris</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonyms</td>
<td>Blue iris (finding)</td>
</tr>
<tr>
<td>ID</td>
<td><a href="http://purl.bioontology.org/ontology/SNOMEDCT/301952009">http://purl.bioontology.org/ontology/SNOMEDCT/301952009</a></td>
</tr>
<tr>
<td>Active</td>
<td>1</td>
</tr>
<tr>
<td>altLabel</td>
<td>Blue iris (finding)</td>
</tr>
<tr>
<td>CASE SIGNIFICANCE ID</td>
<td>900000000000020002</td>
</tr>
<tr>
<td>CTV3ID</td>
<td>Xa88f</td>
</tr>
<tr>
<td>cui</td>
<td>C0578626</td>
</tr>
<tr>
<td>DEFINITION STATUS ID</td>
<td>900000000000074008</td>
</tr>
<tr>
<td>Effective time</td>
<td>20020131</td>
</tr>
<tr>
<td>Has finding site</td>
<td>Iris structure</td>
</tr>
<tr>
<td>interprets</td>
<td>Color of iris</td>
</tr>
<tr>
<td>notation</td>
<td>301952009</td>
</tr>
<tr>
<td>prefLabel</td>
<td>Blue iris</td>
</tr>
</tbody>
</table>

*BioPortal*

---

The National Center for Biomedical Ontology is one of the National Centers for Biomedical Computing supported by the NHGRI, the NHLBI, and the NIH Common Fund.

Copyright © 2005-2015, The Board of Trustees of Leland Stanford Junior University. All rights reserved.

NCBO Website, Release Notes, Terms of Use, Privacy Policy, How to Site.
Search results for **blue iris**

**Iris spuria**  [NCBITaxon](http://purl.obolibrary.org/obo/NCBITaxon_92026)

http://purl.obolibrary.org/obo/NCBITaxon_92026

The blue of a iris.

**blue iris**  [OBA](http://purl.obolibrary.org/obo/OBA_1000124)

http://purl.obolibrary.org/obo/OBA_1000124

A markedly blue coloration of the iris.

**Blue irides**  [HP](http://purl.obolibrary.org/obo/HP_0000025)

http://purl.obolibrary.org/obo/HP_0000025
blue iris

http://purl.obolibrary.org/obo/OBA_1000124

The blue of a iris.

**Synonyms:** iris blue, blue iris trait

---

**Class relations**

**Equivalent to:**
- blue and inheres in some iris

**Subclass of:**
- iris color
- blue
- inheres in some iris
OVERLOAD!!!!!
bioontology spectrum

- Ad-hoc Hierarchies
- XML DTDs
- Data Models (UML, STEP)
- DB Schema
- Logic Programming
- Structured Glossaries
- Description Logics
- "Ordinary" Glossaries
- Thesauri
- Terms

E.g. NCIT
National Cancer Institute Thesaurus

E.g. all OBO Foundry ontologies in OWL,
e.g. HPO, OBIB, OBI
Thesaurus vs Ontology with axioms

```
"Body part"

is_a

"Eye"

is_a

"Blue eye"
```
Thesaurus vs Ontology with axioms

For all instances:

*a Eye
*has_colour* = Blue
Thesaurus vs Ontology with axioms

```

For all instances:

\( a \text{ Eye} \quad has\_colour = \text{Blue} \)
```
Guideline 2

Choose an ontology that has properties
Systematized Nomenclature of Medicine - Clinical Terms

Jump To:

- Iris finding
- Anterior segment of eye
- Iris color - finding
- Anomaly of iris - finding
- Iris pigmentation - finding
- Iris hypopigmentation
- Iris hyperpigmentation
- Pale spots in peripheral iris
- Position of iris - finding
- Focal lesion in iris
- Iris bleeding
- Iris normal
- Iris transillumination
- Mass in iris
- O/E - iris
- Blue iris
- Brown iris
- Green iris
- Heterochromic iris

Details

- Preferred Name: Blue iris
- Synonyms: Blue iris (finding)
- ID: http://purl.bioontology.org/ontology/SNOMEDCT/301952009
- Active: 1
- altLabel: Blue iris (finding)
- CASE SIGNIFICANCE ID: 9000000000000020002
- CTV3ID: C9a88f
- cui: C0578626
- DEFINITION STATUS ID: 9000000000000074008
- Effective time: 20020131
- Has finding site: Iris structure
- interprets: Color of iris
- notation: 301952009
- prefLabel: Blue iris
blue iris

http://purl.obolibrary.org/obo/OBA_1000124

The blue of a iris.

**Synonyms:**  iris blue, blue iris trait

---

**Class relations**

**Equivalent to:**
- blue and inheres in some iris

**Subclass of:**
- iris color
- blue
- inheres in some iris
Guideline 3

Choose an ontology that uses a commonly used upper ontology

Look for OBO Foundry or EFO
Upper level ontologies

- **Thing**
  - **Abstract class with high-level properties**
    - e.g. Continuant
  - e.g. Disease
    - e.g. Rare disease, frequency < 1/10000
  - e.g. Cystic Fibrosis, type B, mutations X, Y
  - **More general domain class with fewer properties**
  - **Specific class with many properties**
Upper level ontologies

- **Thing**
  - Abstract class with high-level properties
    - ~“in terms of”
      - Abstract class with high-level properties
        - ~“in terms of”
          - More general domain class with fewer properties
            - ~“in terms of”
              - Specific class with many properties
Upper level ontologies

- **Thing**
  - Abstract class with high-level properties
    - e.g. **Continuant**
    - **Disease**
      - e.g. **Rare disease, frequency < 1/10000**
        - e.g. **Cystic Fibrosis, type B, mutations X, Y, modifiers A, B**
  - More general domain class with fewer properties
  - Specific class with many properties

How does this relate to data integration?
Upper level ontologies

- **Thing**
  - Ontology used by all
    - Abstract class with high-level properties
      - Ontology usable by most, if not all
        - Ontology used in large community, in many ontologies
          - Ontology used in small community, in few ontologies
            - My ontology used by just me
        - Ontology used in large community, in many ontologies
    - Abstract class with high-level properties
      - More general domain class with fewer properties
        - Specific class with many properties
Upper ontologies

Benefits

1. **Shared conceptualization = more paths (relations) between linked data sets**

2. **Computer reasoning**
   (automatic classification, query extension, consistency checking, e.g. a blue horse can never be an eye)
Ignore guideline 3 if

- You only need annotation for visualization in a user interface
  (i.e. never for data integration or reasoning)
- You would only like to do something with the term hierarchy itself
  (e.g. term enrichment analysis)
Guideline 5, 6, 7

Use ontologies defined in OWL and RDF
(*most* bioontologies in bioportal are)

Choose ontologies that are being used
(check usage statistics in Bioportal)

Send feature requests and feedback to ontology creators
Ask advice

e.g. EBI (Helen Parkinson et al.), Katy Wolstencroft, Luiz Bonino, Marco Roos,...
Summary: Why do we need ontologies?

- RDF + SPARQL allows querying across triple stores
- Ontologies provide a common vocabulary for describing a domain – less ambiguity
- Ontologies provide the ability to logically explore data – inference and reasoning
Summary: Components of an Ontology

- **Classes**
  - Concepts, entities, ontology terms
- **Individuals**
  - Instances of those concepts or entities
- **Properties**
  - Data type
    - Relates individuals to a data value
  - Object type
    - Relates individuals to each other
- **Axioms, relations and restrictions**
  - Logical assertions (rules)
Guidelines for choosing ontology concepts

1. Look for concepts in existing ontologies
2. Use ontologies defined in OWL and RDF
3. Choose ontologies that are being used
4. Choose an ontology that has properties
5. Choose a curated ontology e.g. OBO Foundry, BFO
   Upper ontologies can provide additional benefits
6. Ignore guideline 4 & 5 if
   (i) you need annotation for visualization, not integration or reasoning;
   (ii) you have plans with the terms themselves (e.g. enrichment)
   > a vocabulary is sufficient
7. Send requests and feedback to ontology creators
8. Ask advice
What to use for instances and values?

1. When instances are terms, use the consensus vocabulary for your community (e.g. HPO for human phenotypes)

2. When instances are database identifiers
   1. Stimulate source to provide URIs (+provenance)
   2. Use Identifiers.org

3. When an instance is needed to link a literal value to:
   1. create a new URI by following ‘10 simple rules for persistent identifiers’
   2. define value as a ‘datatype’ property of the instance

ALWAYS ADD A RDF:LABEL to ANY CLASS, PROPERTY, INSTANCE
Thank you