Best Practices for Vocabulary-based Projects

Development, Standardization, Registration, Harmonization and Support

Agenda for Plenary 12

- Introductions (5 mins)
- Aims and Approach of BoF (5 mins)
- Presentations (30 mins)
- Discussion (40 mins)
- Next Steps & Follow Up (10 mins)

Introductions

- Vocabulary use increasing in data management and domain applications
  - FAIR, Open Access etc. driving need to share, discover and re-use material effectively
- Many vocabulary initiatives...and many (overlapping) vocabularies: (Juliane)
  - Libraries
  - Domain Specialists
  - Data Management
  - Publications and Scholarly Comms
  - Funders...
- Some standardisation efforts but incomplete coverage of all domains/use cases
  - Data Centric
    - RDA's DFT , The International Research Data Management (IRiDiuM) Glossary, NIST Big Data Solutions Reference Glossary, CASRAI, BioPortal, FAIRsharing....(Gary)(Peter McQuilton)
  - Domain Included
    - MIG, BioPortal (but becoming less domain-specific), FAIRshare (Koskela.....)

Aims of BoF

- Develop common best-practice approach(s) to vocabulary/ontology use
  - Avoid unnecessary re-engineering
    - Identify duplicate or strongly similar work
    - Encourage collaboration or merging (cf. Open Annotation and Annotation Ontology)
  - Encourage re-use rather than de-novo development
    - Requires an easy way to locate existing work (which does not exist yet - FAIRsharing can do this?)
  - Reduce silo-isation
    - And thus improve sharing, interdisciplinarity and discoverability
    - Also benefits code development
  - Reduce friction - Vocabularies should either be...
    - Disjoint...
    - ...or overlaps should be amenable to direct mapping...
    - ...or they should import a shared sub-vocabulary

Approach

- Bring together…
  - The (still) many siloed vocab initiatives in RDA and more widely
  - Other cross-domain activities such as Metadata 2020
  - Existing standards and de-facto standards efforts
    - With an emphasis on those in practical use!
- Leverage…
  - Expertise at crafting vocabularies and cross-walks/mappings
  - Practical domain/use-case experiences
  - Vocabulary service providers and services
- To deliver
  - Common best-practice approaches
  - Tools and services to enable/promote best practice
  - Adoption of such practice in relevant RDA work

Presentations

1. Case Studies and Initiatives
    a. Automated Subject Heading Extracttion tool (ASHE)
    b. Earthcube/RDA MIG
    c. Metadata 2020
2. Tooling and Templates
    a. FAIRSharing.org (video) - build upon rather than build anew (Dr Peter McQuilton)
    b. CEDAR
3. Data-Centric Vocabularies
    a. RDA DFT & Related Groups

Use Case

ASHE Project

The Problem

- Historic medical collection in the offsite Harvard Depository
- Have minimal catalog records with no MeSH tags
- Have digital full text copies
- Costs $5 per recalled item

Solution

- Create a tool using
  - Natural Language Processing (NLP)
  - Medical Subject Headings (MeSH)
  - Unified Medical Language System (UMLS)
- Test
  - Use a human cataloger as gold standard
  - Rate machine-extracted headings against gold standard

Initial Conclusions

- It returns correct headings
  - Possibly even correct secondary headings
- It returns misses, but even the misses are close to being correct

Interesting Unexpected Stuff

- Words with multiple meanings
  - Labor
    - MeSH term refers to birth, not physical work/occupation
- Change in language
    - Women's health - matching fails in documents from the 1890s to even the 1950s

Not-So-Interesting Stuff

- Records are siloed in proprietary ILS systems
  - Catalogers need to generate the list and then hand-edit the records
- Domain restricted
  - Needs a hierarchical, open ontology in OWL format
  - For optimal results, needs a large intial concept pool (aka UMLS)

Discussion Topics - Quick Wins

- Resources
  - Ontological/vocabulary Repositories
  - Content
  - Ontology catalogues - metadata about each ontology - domain, who maintains it, which databases implement it etc. (e.g. FAIRsharing)
- Mappings between vocabularies (and ontologies?)
- Common subsets
- Getting a handle on disparate RDA activities in the area
- What next for FAIRSharing WG?

Discussion Topics - Longer-Term Goals

- Better approaches going forward
  - Identify and promote good practice (e.g. healthcare practice disseminated to other disciplines)
  - Normalise good behaviour before projects start
- Construct cross-domain community (Juliane)
- Exit and sustainability
- Consolidation/cleanup

Actions

- WG/IG or what?
- Volunteers?

Contacts

- Gary Berg-Cross, gbergcross@gmail.com
  - Ontolog Forum, RDA US Advisory group
- Neil Jefferies, neil@data2paper.org, @NeilSJefferies
  - Head of Innovation, Bodleian Digital Libraries, University of Oxford
  - Director, Jemura Ltd (Data2Paper)
  - Community Leader, SWORDV3
  - Editor, OCFL
- Andi Ogier
  -
- Juliane Schneider
  - Team Lead, eagle-i, Harvard Catalyst
  - Co-Chair, Libraries for Research Data IG
  - Certified Carpentries Instructor

Use Case

Software Ontology

Summary | Classes | Properties | Notes | Mappings | Widgets

Details

Acronym	SWO
Visibility	Public
Description	The Software Ontology (SWO) has the scope of describing types of software used in Bioinformatics. The SWO covers areas such as the software type, the manufacturer of the software, the input and output data types and the uses (i.e. objectives) the software can be put to. The SWO intends to use BFO as an upper level ontolgoy and subclasses types from the Ontology of Biomedical Investigations. Contact James Malone for info: malone@ebi.ac.uk
Status	Production
Format	OWL
Contact	James Malone, malone@ebi.ac.uk
Categories	Biomedical Resources, Experimental Conditions

Submissions

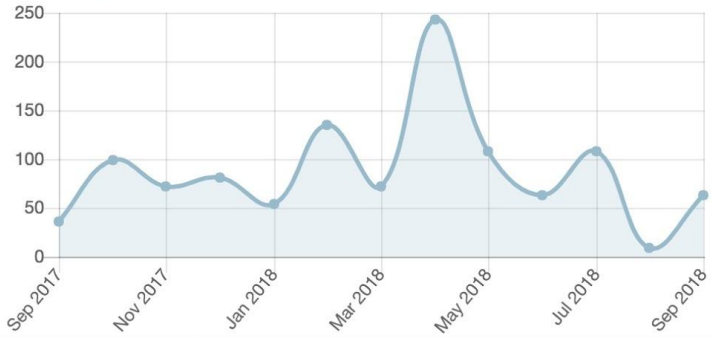
Version	Released	Uploaded	Downloads
0.4 (Parsed, Indexed, Metrics, Annotator)	02/24/2017	02/24/2017	OWL | CSV | RDF/XML | Diff
0.4 (Archived)	02/08/2017	02/08/2017	OWL | Diff
1.5 (Archived)	02/08/2015	10/10/2016	OWL | Diff
1.5 (Archived)	08/20/2015	02/17/2015	OWL | Diff
1.5 (Archived)	02/16/2015	02/16/2015	OWL | Diff

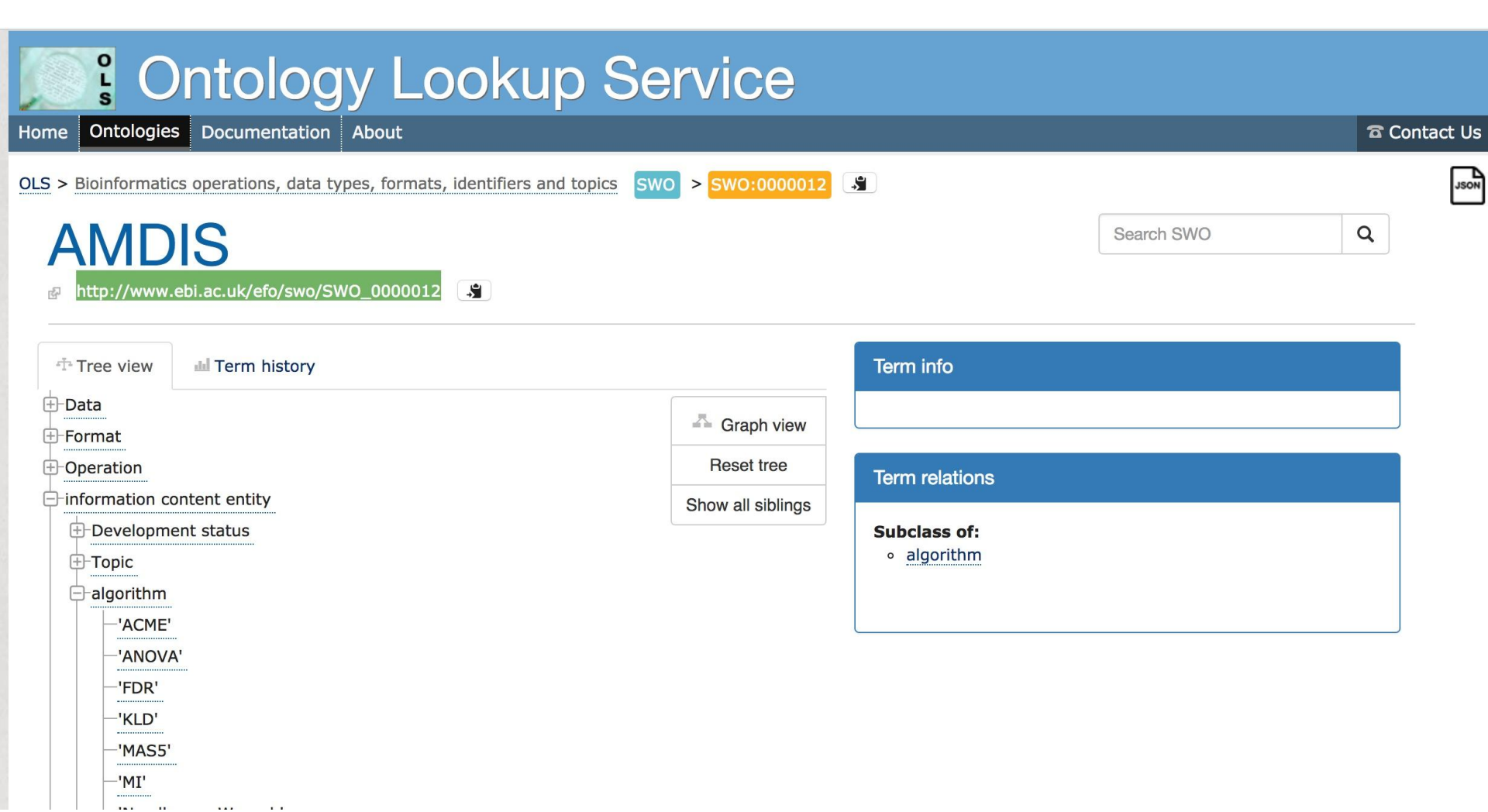
more...

Metrics

Classes	4,427
Individuals	0
Properties	27
Maximum depth	9
Maximum number of children	1,044
Average number of children	6
Classes with a single child	218
Classes with more than 25 children	24
Classes with no definition	3,056

Visits




Ontology Lookup Service
Home
Ontologies
Documentation
About
Contact Us
OLS > Bioinformatics operations, data types, formats, identifiers and topics SWO > SWO:0000012
JSON
AMDIS
Search SWO
http://www.ebi.ac.uk/efo/swo/SWO_0000012
Tree view
Term history
Data
Format
Operation
information content entity
Development status
Topic
algorithm
'ACME'
'ANOVA'
'FDR'
'KLD'
'MAS5'
'MI'
Graph view
Reset tree
Show all siblings
Term info
Term relations
Subclass of:
algorithm

HTTP Status 404 - /efo/swo/SWO_0000012

---

type Status report

message /efo/swo/SWO_0000012

description The requested resource is not available.

Apache Tomcat/7.0.55