Creating and using ontologies

Elena Simperl AIFB, Karlsruhe Institute of Technology, Germany

With contributions from “Linked Data: Survey of Adoption”, Tutorial at the 3rd Asian Semantic Web School ASWS 2011, Incheon, South Korea, July 2011 by Aidan Hogan, DERI, IE

Session 2, Day 2, 14:30 – 17:00
Ontologies in Computer Science

- An ontology defines a domain of interest
- … in terms of the things you talk about in the domain, their features and characteristics, as well as relationships between them
Ontologies in Computer Science (2)

- Ontologies are used to
  - Share a common understanding about a domain among people and/or machines
  - Enable reuse of domain knowledge

- This is achieved by
  - Agreeing on meaning and representation of domain knowledge
  - Making domain assumptions explicit
  - Separating domain knowledge from the operational knowledge

- They are used (under different names) in various areas
  - Data management and integration
  - Digital libraries
  - Multimedia analysis
  - Software engineering
  - Machine learning
  - Natural language processing
  - …
Are Semantic Web ontologies just UML?

- **Ontologies vs ER schemas**
  - Semantic Web ontologies represented in Web-compatible languages, using Web technologies
  - They represent a shared view over a domain

- **Ontologies vs UML diagrams**
  - Formal semantics of ontology languages defined, languages with feasible computational complexity available

- **Ontologies vs thesauri**
  - Formal semantics, domain-specific relationships

- **Ontologies vs taxonomies**
  - Richer property types, formal semantics of the is-a relationship

- **Ontologies vs Linked Data vocabularies**
  - Well…
Process overview

Requirements analysis
motivating scenarios, use cases, existing solutions, effort estimation, competency questions, application requirements

Conceptualization
conceptualization of the model, integration and extension of existing solutions

Implementation
implementation of the formal model in a representation language
Requirements analysis (1): Domain and scope

- What is the ontology going to be used for?
- Who will use the ontology?
- How it will be maintained and by whom?
- What kind of data will refer to it? And how will these references be created and maintained?
- Are there any information sources available that could be reused?
- What questions should the ontology be able to answer?

To answer these questions, talk to domain experts, users, and software designers

- Domain experts don’t need to be technical, they need to know about the domain, and help you understand its subtleties
- Users teach you about the terminology that is actually used and the information needs they have
- Software designers tell you about the type of use cases you need to handle, including the data to be described via the ontology
Semantic technologies at BestBuy

- Goal: “to provide more visibility to products, services and locations to humans and machines”
- Search engines identify the data more easily and put it into context (30% increase in search traffic)
- Improved consumer experience

Due to “Increasing product and service visibility through front-end semantic web” by Jan Myers, SemTech 2010
Semantic technologies at BestBuy(2)

- Data is marked-up using RDFa and refers to concepts from a pre-defined eCommerce ontology.
- Markup is entered by BestBuy staff via online forms that produce RDFa.

Due to “Increasing product and service visibility through front-end semantic web” by Jan Myers, SemTech 2010
Requirements analysis (2): Domain vs task-oriented ontologies

- **Domain-oriented**
  - Ontology models the types of entities in the domain of the application
    - Example: content and features of movies, points of interest in a city, different types of digital camera’s…
  - Cover the terminology of the application domain
    - Example: classifications, taxonomies, folksonomies, text corpora
  - Used for annotation and retrieval

- **Task-oriented**
  - Ontology serves a purpose in the context of an application
    - Example: finding movies with certain features, recommending sightseeing tours matching my interests, finding and comparing products matching user preferences
  - Define the structure to a knowledge base that can be used to answer competency questions
  - Used for automated reasoning and querying

Content due to Valentina Pressuti and Eva Blomqvist
Requirements analysis (3): Competency questions

- A set of queries which place demands on the underlying ontology

- Ontology must be able to represent the questions using its terminology and the answers based on the axioms

- Ideally, in a staged manner, where consequent questions require the input from the preceding ones

- A rationale for each competency question should be given
## Requirements analysis (4): Examples

<table>
<thead>
<tr>
<th>Competency Questions</th>
<th>Other requirements</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>■ Which wine characteristics should I consider when choosing a wine?</td>
<td>■ Concepts in the ontology should be bi-lingual.</td>
<td>■ An ontology reflects an abstracted view of a domain of interest. You should not model all possible views upon a domain of interest, or to attend to capture all knowledge potentially available about the respective domain.</td>
</tr>
<tr>
<td>■ Is Bordeaux a red or white wine?</td>
<td>■ The ontology should not have more than 10 inheritance levels.</td>
<td>■ Even after the scope of the ontology has been defined, the number of competency questions can grow very quickly → modularization, prioritization.</td>
</tr>
<tr>
<td>■ Does Cabernet Sauvignon go well with seafood?</td>
<td>■ The ontology should be extended and maintained by non-experts.</td>
<td>■ Requirements are often contradictory → prioritization.</td>
</tr>
<tr>
<td>■ What is the best choice of wine for grilled meat?</td>
<td>■ The ontology should be used to build an online restaurant guide.</td>
<td></td>
</tr>
<tr>
<td>■ Which characteristics of a wine affect its appropriateness for a dish?</td>
<td>■ The ontology should be usable on an available collection of restaurant descriptions written in German.</td>
<td></td>
</tr>
<tr>
<td>■ Does a flavor or body of a specific wine change with vintage year?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Competency Questions

- Which wine characteristics should I consider when choosing a wine?
- Is Bordeaux a red or white wine?
- Does Cabernet Sauvignon go well with seafood?
- What is the best choice of wine for grilled meat?
- Which characteristics of a wine affect its appropriateness for a dish?
- Does a flavor or body of a specific wine change with vintage year?
Requirements analysis (5): Finding existing ontologies

Where to find ontologies

- Swoogle: over 10 000 documents, across domains
  - [http://swoogle.umbc.edu/](http://swoogle.umbc.edu/)
- Protégé Ontologies: several hundreds of ontologies, across domains
- Open Ontology Repository: work in progress, life sciences, but also other domains
- Tones: 218 ontologies, life sciences and core ontologies.
  - [http://owl.cs.manchester.ac.uk/repository/browser](http://owl.cs.manchester.ac.uk/repository/browser)
- Watson: several tens of thousands of documents, across domains
  - [http://watson.kmi.open.ac.uk/Overview.html](http://watson.kmi.open.ac.uk/Overview.html)
- Talis repository
- Ontology Yellow Pages: around 100 ontologies, across domains
- OBO Foundation Ontologies
  - [http://www.obofoundry.org/](http://www.obofoundry.org/)
- AIM@SHAPE
  - [http://dsw.aimatshape.net/tutorials/ont-intro.jsp](http://dsw.aimatshape.net/tutorials/ont-intro.jsp)
- VoCamps
  - [http://vocamp.org/wiki/Main_Page](http://vocamp.org/wiki/Main_Page)
Life sciences and healthcare

The Open Biological and Biomedical Ontologies

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The groups developing ontologies who have expressed an interest in this goal are listed below, followed by other relevant efforts in this domain.

In addition to a listing of OBO ontologies, this site also provides a statement of the OBO Foundry principles, discussion fora, technical infrastructure, and other services to facilitate ontology development. We welcome feedback and encourage participation.

Click any column header to sort the table by that column. The link to the term request trackers for the listed ontologies.

<table>
<thead>
<tr>
<th>Title</th>
<th>Domain</th>
<th>Prefix</th>
<th>File</th>
<th>Last changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological process</td>
<td>biological process</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
<td>2010/07/26</td>
</tr>
<tr>
<td>Cellular component</td>
<td>anatomy</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
<td>2010/07/26</td>
</tr>
<tr>
<td>Chemical entities of biological interest</td>
<td>biochemistry</td>
<td>CHEBI</td>
<td>chebi.obo</td>
<td>2010/07/07</td>
</tr>
<tr>
<td>Molecular function</td>
<td>biological function</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
<td>2010/07/26</td>
</tr>
<tr>
<td>Phenotypic quality</td>
<td>phenotype</td>
<td>PATO</td>
<td>quality.obo</td>
<td>2010/07/25</td>
</tr>
<tr>
<td>Protein Ontology (PRO)</td>
<td>proteins</td>
<td>PRO</td>
<td>pro.obo</td>
<td>2010/07/24</td>
</tr>
<tr>
<td>Xenopus anatomy and development</td>
<td>anatomy</td>
<td>XAO</td>
<td>xenopus_anatomy.obo</td>
<td>2009/12/02</td>
</tr>
<tr>
<td>Zebrafish anatomy and development</td>
<td>anatomy</td>
<td>ZFA</td>
<td>zebrafish_anatomy.obo</td>
<td>2010/05/14</td>
</tr>
</tbody>
</table>
Freebase

http://www.freebase.com

An entity graph of people, places and things, built by a community that loves open data.

Freebase is joining Google! Learn more »
# Dublin Core

**Properties in the `/terms/` namespace**

- abstract
- accessRights
- accrualMethod
- accrualPeriodicity
- accrualPolicy
- alternative
- audience
- available
- bibliographicCitation
- conformsTo
- contributor
- coverage
- creator
- created
- date
- dateAccepted
- dateCopyrighted
- dateSubmitted
- description
- educationLevel
- extent
- format
- hasFormat
- hasPart
- hasVersion
- identifier
- instructionalMethod
- isFormatOf
- isPartOf
- isReferencedBy
- isReplacedBy
- isRequiredBy
- issued
- isVersionOf
- language
- license
- mediator
- medium
- modified
- provenance
- publisher
- references
- relation
- replaces
- requires
- rights
- rightsHolder
- source
- spatial
- subject
- tableOfContents
- temporal
- title
- type
- valid

**Properties in the legacy `/elements/1.1/` namespace**

- contributor
- coverage
- creator
- date
- description
- format
- identifier
- language
- publisher
- relation
- rights
- source
- subject
- title
- type

**Vocabulary Encoding Schemes**

- DCMIType
- DDC
- IMT
- LCC
- LCSH
- MESH
- NLM
- TGN
- UDC

**Syntax Encoding Schemes**

- Box
- ISO3166
- ISO639-2
- ISO639-3
- Period
- Point
- RFC1766
- RFC3066
- RFC4646
- RFC5646
- URI
- W3CDTF

**Classes**

- Agent
- AgentClass
- BibliographicResource
- FileFormat
- Frequency
- Jurisdiction
- LicenseDocument
- LinguisticSystem
- Location
- LocationPeriodOrJurisdiction
- MediaType
- MediaTypeOrExtent
- MethodOfAccrual
- MethodOfInstruction
- PeriodOfTime
- PhysicalMedium
- PhysicalResource
- Policy
- ProvenanceStatement
- RightsStatement
- SizeOrDuration
- Standard

Table from [http://dublincore.org/documents/dcmi-terms/](http://dublincore.org/documents/dcmi-terms/)
Friend Of A Friend
Semantically Interlinked Online Communities

Image from [http://rdfs.org/sioc/spec/]: Bojārs, Breslin et al.
Simple Knowledge Organization System

Image from [http://www.w3.org/TR/swbp-skos-core-guide](http://www.w3.org/TR/swbp-skos-core-guide). Miles, Brickley
Description Of A Project

Image from http://code.google.com/p/baetle/wiki/DoapOntology; Breslin
Music Ontology

Image from http://musicontology.com/; Raimond, Giasson
DBpedia

- Classes and properties for Wikipedia export (infoboxes)

---

```{template:Infobox Town AT
name = Innsbruck
image = Innsbruckwappen.png
image_map = Karte-tirol-1.png
state = [(Tyrol)]
region = [(Statutory city)]
population = 117,342
population_as_of = 2006
pop_dens = 1,113
area = 104.91
elevation = 674
lat_deg = 47
lat_min = 16
lat_bom = N
lon_deg = 11
lon_min = 28
lon_bom = E
postal_code = 5010-5080
area_code = 0512
license = Free
mayor = Hilde Zach
website = [http://innsbruck.at]
```
schema.org

Collection of schemas to mark-up structured content in HTML pages
Additional resources

http://vocamp.org/wiki/Where_to_find_vocabularies
Requirements analysis (5): Selecting relevant ontologies

- What will the ontology be used for?
  - Does it need a natural language interface and if yes in which language?
  - Do you have any knowledge representation constraints (language, reasoning)?
  - What level of expressivity is required?
  - What level of granularity is required?

- What will you reuse from it?
  - Vocabulary++

- How will you reuse it?
  - Imports: transitive dependency between ontologies
  - Changes in imported ontologies can result in inconsistencies and changes of meanings and interpretations, as well as computational aspects
Auxiliary support methods

Folksonomy Import (step 1)

- Browse Files | Start Upload
  - Overall Progress (94.6 MB) 64%
  - Uploading "data.xml" 64%
  - Upload: 60.4 MB with 4.6 MB/s, 7 s remaining.

- data.xml
  - Remove

Folksonomy Import (step 2)

- start importing

details:
- initiating connection ✓
- preparing data ✓
- select relevant tags ✓
- do tag grouping ✓
- lookup wordnet ✓
- search in wikipedia ✓
- do spellchecking ✓
- regroup tags ✓
- search coocurrences ✓
- search coactors ✓
- create cluster pairs ✓
- group clusters to ontologies ✓
Conceptualization (1): Vocabulary

- What are the terms we would like to talk about?
- What properties do those terms have?
- What would we like to say about those terms?
- Competency questions provide a useful starting point
- Go int out too far vs. going down too far
- Investigate homonyms and synonyms
Conceptualization (2): Classes

- Select the terms that describe objects having independent existence rather than terms that describe these objects
  - These terms will be classes in the ontology

- Classes represent concepts in the domain and not the words that denote these concepts
  - Synonyms for the same concept do not represent different classes

- Typically nouns and nominal phrases, but not restricted to them
  - Verbs can be modeled as classes, if the emphasis is on the process as a whole rather than the actual execution
    - Visit as an event rather than an action performed by an actor
Conceptualization (3): Class hierarchy

- A subclass of a class represents a concept that is a “kind of” the concept that the superclass represents.

- It has:
  - Additional properties
  - Restrictions different from those of the superclass, or
  - Participates in different relationships than the superclasses.

- All the siblings in the hierarchy (except for the ones at the root) must be at the same level of generality.

- If a class has only one direct subclass there may be a modeling problem or the ontology is not complete.

- If there are more than a dozen subclasses for a given class then additional intermediate categories may be necessary.

- Functional inclusion:
  - A chair is a piece of furniture
  - A hammer is a tool

- State inclusion:
  - Polio is a disease
  - Hate is an emotion

- Activity inclusion:
  - Tennis is a sport
  - Murder is a crime

- Action inclusion:
  - Lecturing is a form of talking
  - Frying is a form of cooking

- Perceptual inclusion:
  - A cat is a mammal
  - An apple is a fruit
Conceptualization(4): Properties

- We selected classes from the list of terms in a previous step
  - Most of the remaining terms are likely to be properties of these classes
- For each property in the list, we must determine which class it describes
  - Properties are inherited and should be attached to the most general class in the hierarchy
- Two types of principal characteristics
  - Measurable properties: attributes
  - Inter-class connections: relationships.
    - Use relationships to capture something with an identity
- Arrest details as attribute of the suspect vs. arrest as an relationship
  - Do we measure degrees of arrestedness or do we want to be able to distinguish between arrests?
- Color of an image as attribute vs. class
  - A „pointing finger“ rather than a „ruler“ indicates identity
Conceptualization (5): Domain and ranges

- Refine the semantics of the properties
  - Cardinality
  - Domain and range
    - When defining a domain or a range for a slot, find the most general classes or class that can be respectively the domain or the range for the slots
    - Do not define a domain and range that is overly general
  - General patterns for domain and range
    - A class and a superclass – replace with the superclass
    - All subclasses of a class – replace with the superclass
    - Most subclasses of a class – consider replacing with the superclass
Conceptualization (6): Ontology Design Patterns

Content from http://ontologydesignpatterns.org/
ONTOLOGIES AND LINKED DATA
Ontologies and Linked Data

- Global, distributed dataspace build on a simple set of standards
  - RDF, URIs, HTTP

- Entities are connected by links
  - creating a global data graph that spans data sources and
  - enables the discovery of new data sources

- Provides for data-coexistence
  - Everyone can publish data to the Web of Linked Data
  - Everyone can express their personal view on things
  - Everybody can use the vocabularies/schema that they like

Tasks:
1. Make data available as RDF via HTTP
2. Set RDF links pointing at other data sets
3. Make your data self-descriptive

Aspects of self-descriptiveness
1. Enable clients to retrieve the schema
2. Reuse terms from common vocabularies
3. Publish schema mappings for proprietary terms
4. Provide provenance metadata
5. Provide licensing metadata
6. Provide data-set-level metadata using voID
7. Refer to additional access methods using voID

Content due to Chris Bizer
Ontologies and Linked Data

- Model pre-defined through the (semi-) structure of the data to be published
- Emphasis on alignment, especially at the instance level
- Stronger commitment to reuse instead of development from scratch

- Human vs machine-oriented consumption (using specific technologies)
- Trade-off between acceptance/ease-of-use and expressivity/usefulness

- Publication according to Linked Data principles
Example: BBC

- Various micro-sites built and maintained manually.
- No integration across sites in terms of content and metadata.
- Use cases
  - Find and explore content on specific (and related) topics.
  - Maintain and re-organize sites.
  - Leverage external resources.
- Ontology: One page per thing, reusing DBpedia and MusicBrainz IDs, different labels...

"Design for a world where Google is your homepage, Wikipedia is your CMS, and humans, software developers and machines are your users."

http://www.slideshare.net/reduxd/beyond-the-polar-bear
STATE OF THE ART AND OPEN TOPICS
Ontology engineering today

- Various domains and application scenarios: life sciences, eCommerce, Linked Open Data
- Engineering by reuse for most domains based on existing data and vocabularies
  - Alignment of data sets
  - Data curation
  - Human-aided computation (e.g., games, crowdsourcing)
- Most of them much simpler and easier to understand than the often cited examples from the 90s
  - However, still difficult to use (e.g., for mark-up)
Open topics

- Meanwhile we have a better understanding of the scenarios which benefit from the usage of semantics and the technologies they typically deploy
  - Guidelines and how-to’s
  - Design principles and patterns
  - Schema-level alignment (data-driven)
  - Vocabulary evolution
  - Assessment and evaluation

- Large-scale approaches to knowledge elicitation based on combinations of human and computational intelligence
The current configuration of the “Red Hot Chili Peppers” are: Anthony Kiedis (vocals), Flea (bass, trumpet, keyboards, and vocals), John Frusciante (guitar), and Chad Smith (drums). The line-up has changed a few times during they years, Frusciante replaced Hillel Slovak in 1988, and when Jack Irons left the band he was briefly replaced by D.H. Peligo until the band found Chad Smith. In addition to playing guitars for Red hot Chili Peppers Frusciante also contributed to the band “The Mars Volta” as a vocalist for some time.

From September 2004, the Red Hot Chili Peppers started recording the album “Stadium Arcadium”. The album contains 28 tracks and was released on May 5 2006. It includes a track of the song “Hump de Bump”, which was composed in January 26, 2004. The critic Crian Hiatt defined the album as "the most ambitious work in his twenty-three-year career". On August 11 (2006) the band gave a live performance in Portland, Oregon (US), featuring songs from Stadium Arcadium and other albums.
Assignment: Alignment

- The aim is to reach a 'shared conceptualization' of all participants at the ESWC2011 summer school on the ontology developed in the previous assignment.
  - Assumption: every group is committed to their conceptualization.
  - Procedure: each group selects a representative, representatives agree on an editor, and on the actual steps to be followed.