Biomedical semantics, information retrieval and knowledge discovery - (2)

Semantic Web and Controlled Vocabularies

Dr. Dietrich Rebholz-Schuhmann
Dr. Leyla Jael Garcia Castro

December 18, 2020

Outline

Objectives and learning outcomes

Objectives

- We will get an overview of Semantic Web followed by an introduction to the spectrum of controlled vocabularies with a focus on those where community agreements play an important role (i.e., terminologies), particularly ontologies and their main elements.

Learning outcomes

- Understanding what Semantic web is and how it relates to the Web
- Explaining the relation between semantics and controlled vocabularies
- Distinguishing differences across terminologies in terms of formality and semantics support
- Explaining what an ontology is and describing its main characteristics and elements

Semantics and Semantic Web

Literature

Hitzler, Krötzsch, Rudolph, Sure
"Foundations of Semantic Web Technologies", Springer-Verlag

Hitzler, Krötzsch, Rudolph, Sure
"Semantic Web", Springer-Verlag

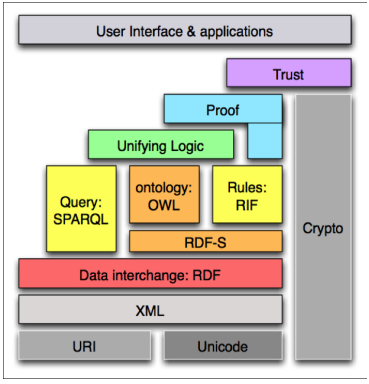
How is the Web formed?

- Computers communicate through their connections using their communication protocols.
- The main objective: exchange of data over all connected computers in the world
- The data exchange requires data standards.
- There are different layers defined by the data exchange protocols for the communication between computers.

Tasks of the Internet

- The Internet is a network composed of computers: servers and clients.
- Each computer has a unique Internet Protocol IP address to identify the machine: e.g. 201.198.168.001. – The IP address is given to the computer automatically when it registers to the internet (via domain server).
- A domain is a name (on the Web), which has been provided by a specific computer (domain server), to deliver a selection of Web pages.
- All domain servers know all other domain servers.
- IP addresses, domain names and email addresses are unique logical elements on the Web.

Data layers of the Web



- 1994: Concept of the Semantic Web
- 1999: Standardizing the data (RDF), language for ontologies (RDFS) by the W3C
- 2000: Novel and big ontologies (DAML and Ontoknowledge)
- 2002: Standardizing ontologies (OWL)
- 2004: Latest standard for data and ontologies (RDF, OWL)
- 2008: Standardizing data queries (SPARQL)
- 2009: Extension from OWL to OWL 2.0
- 2010: Standard Rule Interchange Format (RIF)

How do we keep data?

- A priori computers exchange data from flat files or more complex file formats.
- Data exchange requires data standards.
- We can distinguish different layers in the communication of computers, i.e. packages vs. full files vs. streamed information vs. Internet of Things.

Structured vs. unstructured data

	Unstructured Data	Structure Data
Example	text, image, lists	databases
Standards	Formats	Formats, data schema Data types [1]
Data storage	Keep any data	Data according to schema
Search	Simple means, e.g. string search	Data query of fields
Software (Example)	WinWord, OpenWin Text	MySQL, Oracle, ("Excel" [2])
On the Web [3]	Simple transformation possible	Complex transformation required

[1] e.g. String, Integer, Float, Boolean, Date

[2] Not a relational database, but Excel uses data types

[3] Transformation into HTML possible, but usage is more complex

Semantic Web

Semantics: the theory of meaning
Semantic Web:

- Should serve more than only the exchange of information and data.
- The data should be accompanied with meaning.
- The data should be interpretable - like the perception of the real world.
- We would like to "compute" (reason) over the data like we use logical considerations in the real world.

Web and semantic Web

World Wide Web:

- Web servers to deliver Web pages and content
- Each server can be identified by its IP address (210.217.209.005)
- Data is not interoperable, i.e. data exists in data silos (i.e. without access from the Web)

Semantic Web:

- Enable machines to use data across the Web
- Have unique representation of resources (URIs)
- Enable processing of data, i.e. data analysis, inference of new results

Controlled vocabularies: from lists to ontologies

From lists to ontologies

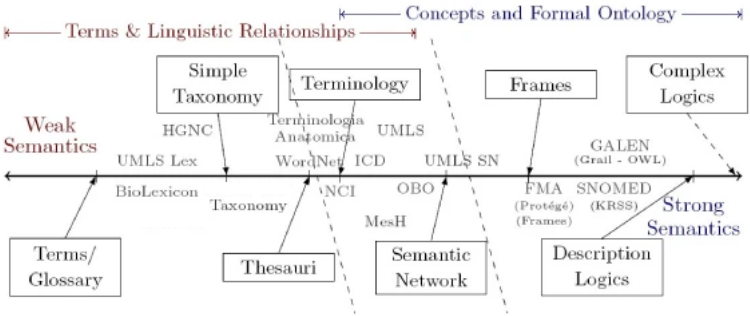


Figure: The ontology spectrum, taken from
`http://doi.org/10.1186/1471-2105-10-S10-S4`

Controlled vocabularies

- Catalogs:
  A scattered lists of terms.
- Glossaries:
  A scattered lists of terms plus glosses in natural language.
- Terminologies:
  A set of terms authorized by a community established mandate. In theory, the terms are defined excluding ambiguity (possible use of preferred terms, synonyms).

Although a term in a catalog, glossary or terminology is not necessarily accompanied by an identifier, it is still helpful to assume that each term has a unique identifier. It can be assigned explicitly or can be formed by transforming the resource into a Semantic Web resource giving each term a URI (Uniform Resource Identifier).

Taxonomies

- Taxonomies:
  A CV organized into a hierarchical structure using the basic parent-child relationship (aka: whole-part, broader-narrower, genus-species, type-instance).

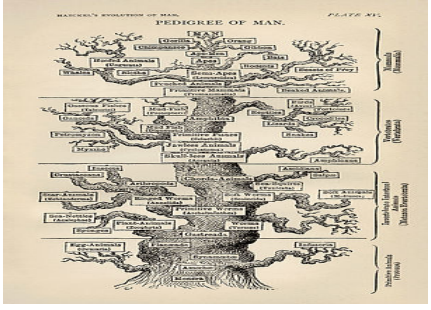


Figure: By Ernst Haeckel - Taken from Wikipedia. Public Domain

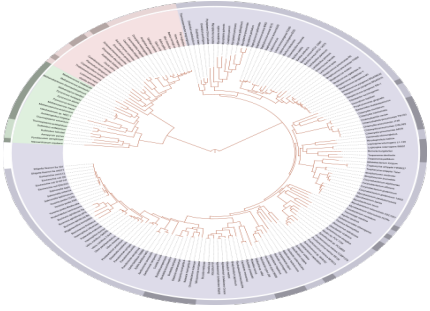


Figure: By Ivica Letunic. Taken from Wikipedia. Public Domain

Thesauri

- Thesauri:
  Taxonomies enriched by relations for equivalence or association between terms (i.e. a term being "synonym of", "related to", or "similar to" the preferred term).

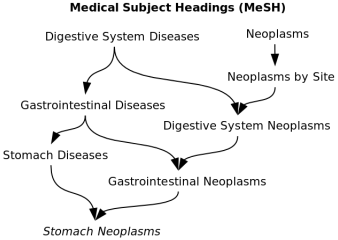


Figure: Uploader Nichtich commonswiki - from Jakob Voss. CC BY-SA 3.0. Taken from Wikipedia.

Ontologies

- Ontology:
  Taxonomies/Thesauri enriched by more complex relations beyond equivalence and determined by the domain it describes (e.g. a term "regulates the activity" of other term), the most complex type of CV with strong semantic support.

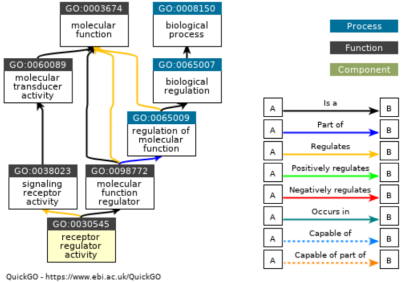


Figure: GO:0030545. Taken from QuickGO

Terminological resources used in language technologies

The different terminological resources can be exploited in several ways:

- "Term lookup": Looking up the terms in text, i.e. matching the terms (= labels) to the text, possibly considering morphological variability.
- "Data integration": Assigning the identifier from a label to a passage in the text: normalizing the text to the terminological standards (but also between data entries in a relational database).
- "Data retrieval": Using the assigned identifier (for a label) in an indexing engine that allows concept-based retrieval of data.

Introduction to ontologies

An ontology is the "explicit specification of a conceptualization" https://doi.org/10.1006/knac.1993.1008(Gruber, 1993). The conceptualization describes a part of the real world ("a domain").

The ontologist can - with the help of special data formats and formalism - describe the observations from the real world as an electronic data resource.

The ontologists denote axioms (i.e. true statements) that specify best their observations for use in computers.

Introduction to ontologies (1)

The ontologists distinguishes the following two elements

- Thing (or Being): Existing entities in the real world, typically objects such as a chair, a pipe, some living being.
- Concept or type: The abstraction of a "thing". This abstraction is represented as a specification or as information about the "thing".

Note: The debate of whether or not a "thing" exists, i.e. the question of the being, is not part of the lectures. I/we assume that things do exist.

Introduction to ontologies (2)

- Identifier: A unique number or string that helps to identify the concept / type amongst others.
- Labels: A label is the natural language short-form representation of the concept or type. It is does not specify the concept or type, but supports daily communication about the type.
- Synonym: A concept or type may be known under different labels (called "synonyms"). One commonly chosen as the "preferred term".
- Definition: The label is meaningful, but only the definition can be specific enough to fully determine the characteristics of the concept or type term.

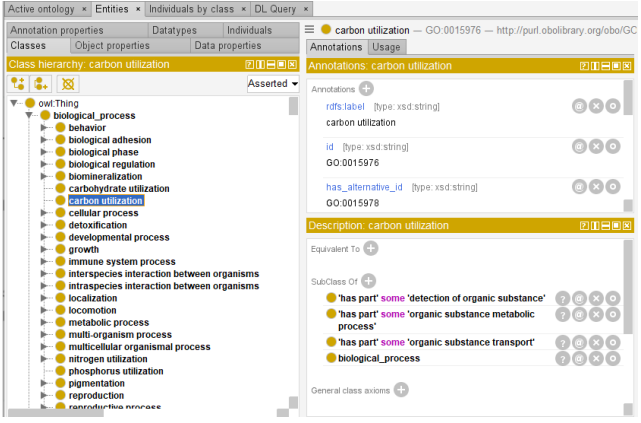
Components in an ontology

The following elements are used to form an ontology:

- Individuals: instances or objects (the basic or "ground level" objects)
- Classes: sets, collections, concepts, types of objects, or kinds of things.
- Attributes: aspects, features, characteristics, or parameters that objects (and classes) can have
- Relations: ways in which classes and individuals can be related to one another.
- Other (to be defined in the future)

Ontological components

We can distinguish: classes, individuals and axioms (or facts, statements) in a formal representation of semantics.



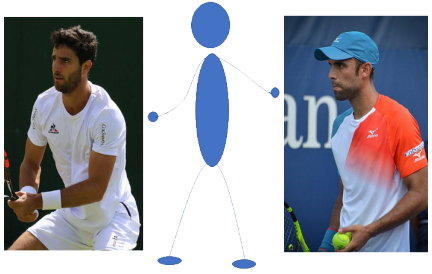
Formal representation

The semantics of classes, individuals and properties can be represented in a formal way, for example using XML, RDF or OWL (to be defined).

A practical example

- Instances / Individuals: Boris Becker, Andre Agassi, Roger Federer, Steffi Graf.
- Classes:

tennis players	golf players	tennis records
women and men	female tennis players	Wimbledon winners



Juan Sebastian Cabal (by Carine06) and Robert Farah (by si.robi)

Summary

Summary

- There is a broad spectrum of controlled vocabularies, which one is the right for you depends on your use case. The more formal semantics you need, the closer to ontologies you need to move in the spectrum.
- Ontologies are great to conceptualize and infer knowledge based on the underlying meanings.
  Ontologies make use of individuals, classes, properties and axioms to represent semantics, and we can make use of sets, trees and graphs to represent ontologies.

So where are we now?

- For biomedical semantics:
  We have an overview of different types of controlled vocabularies/terminologies with some examples for the Biomedical domain.
- For information retrieval and knowledge discovery:
  We have a basic idea on how such terminologies can be combined with language techniques: term lookup, data integration and data retrieval.
- We also have an initial understanding of ontologies, their elements, components and representation. We will go deeper into it on a future lesson.
- We can provide intuitive examples on how elements in the real world could be modelled as an ontology