7 - Bio-Ontologies

André Lamúrias

Motivation

Motivation

Major challenges in biomedical research

- Access and analyze increasing amounts of data
- Harness it for novel insights about biology or medicine
- Suggest hypotheses for further research

Amounts of data ever-increasing

- PubMed contains today over 35 million citations for biomedical literature
  - Around 20 million in 2011
- High-throughput technologies such as microarrays
- Massively parallel DNA sequencing

Motivation

Automated Search

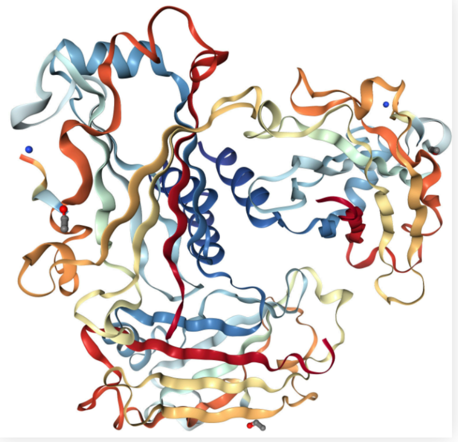
- Necessary, given the sheer amount of data and information
- Searching for strings - sequences of characters - works well
  - Text documents (.pdf, .docx, etc., also programming code)
  - Web search engines
- Search for terms corresponding to a given string difficult
  - Synonyms, abbreviations, and acronyms

Motivation

Example

- TRAF2 - gene TNF receptor-associated factor 2
- Alternative names: TRAP, TRAP3, MGC:45012
- Search (e.g. in PubMed) returns differing results

- TRAF2 - 1914 results
- TNF receptor-associated factor 2 - 1201 results
- TRAP - 51719 results (most not related to the gene)
- TRAP3 - 7 results
- MGC:45012 - 0 results



Motivation

Objective - Semantic 'Dictionary'

- Determine vocabulary of terms
- Establish alternative names and synonyms
- Provide relations between these terms

Semantic Web

Semantic Web Vision

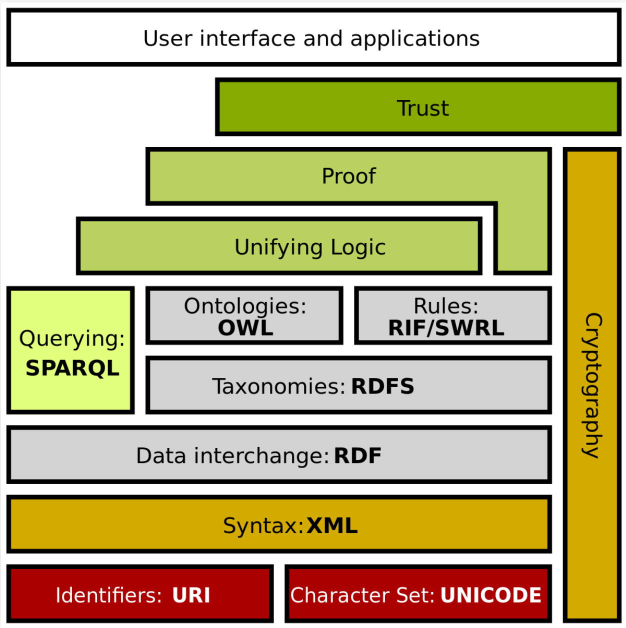
- Proposed by Tim Berners Lee in 2001
- Web of human-readable content (text and pictures - early 2000s)
- Intelligent work (selecting, combining, aggregating) delegated to human reader
- Make the web more accessible to computers

Design principles

- Make structured and semi-structured data available in standardized formats on the Web
- Make not just the datasets, but also the individual data elements and their relations accessible on the Web
- Describe the intended semantics of such data in a formalism, so that this intended semantics can be processed by machines

Semantic Web Initiative

- World Wide Web Consortium (W3C)
- Development of language standards (such as RDF, RDFS and OWL)



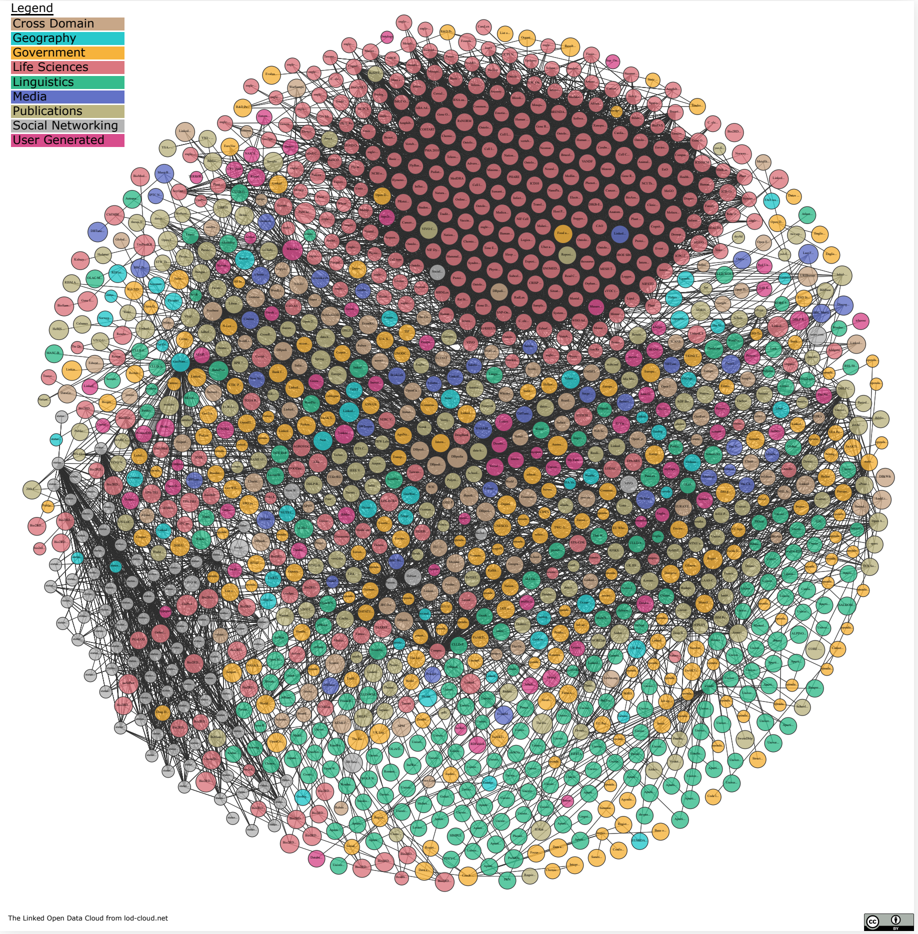
Basic technologies

- Labeled graphs as data model for objects and their relations (using RDF - Ressource Description Framework)
- Web identifiers to identify data items and their relations (using URI - Uniform Ressource Identifiers)
- Ontologies as data model for the intended semantics of data (using RDF Schema and OWL - Web Ontology Language)

Linked Open Data (LOD)

- Publication of interlinked data for a wide variety of topics and from different stakeholders
  - Cross-Domain, Geography, Government, Life Sciences, Linguistics, Media, Publications Social Networking
  - DBpedia, BBC, New York Times Company, Facebook etc.
- Linked Open Data Cloud https://lod-cloud.net/

Linked Open Data Cloud



Knowledge Graphs

- Appeared with Google Knowledge Graph in 2012
  - Represented through Google Search Engine Results Pages
  - Over 500 million objects
  - Data from Freebase/Wikidata, Wikipedia, CIA World Factbook, etc.
- Arguably a rebrand of Semantic Web technologies
- Semantic network/graph that represents real-world objects and relations between them
- Ontologies for formal representation of the entities in the graph

Ontologies

Ontologies

Origins

- Dates back (at least) to Greek Philosophy
- Study of the nature of existence
- Aristotle developed conceptual taxonomies
- The term is due to Jean le Clerc in 1692 - essence of being (On/Ontos) + suffix -logy (study of)
- In Computer Science adopted in Knowledge Representation for describing specific domains of reality

'An ontology is an explicit and formal specification of a conceptualization.' (T.R Gruber and R. Studer)

Ontologies

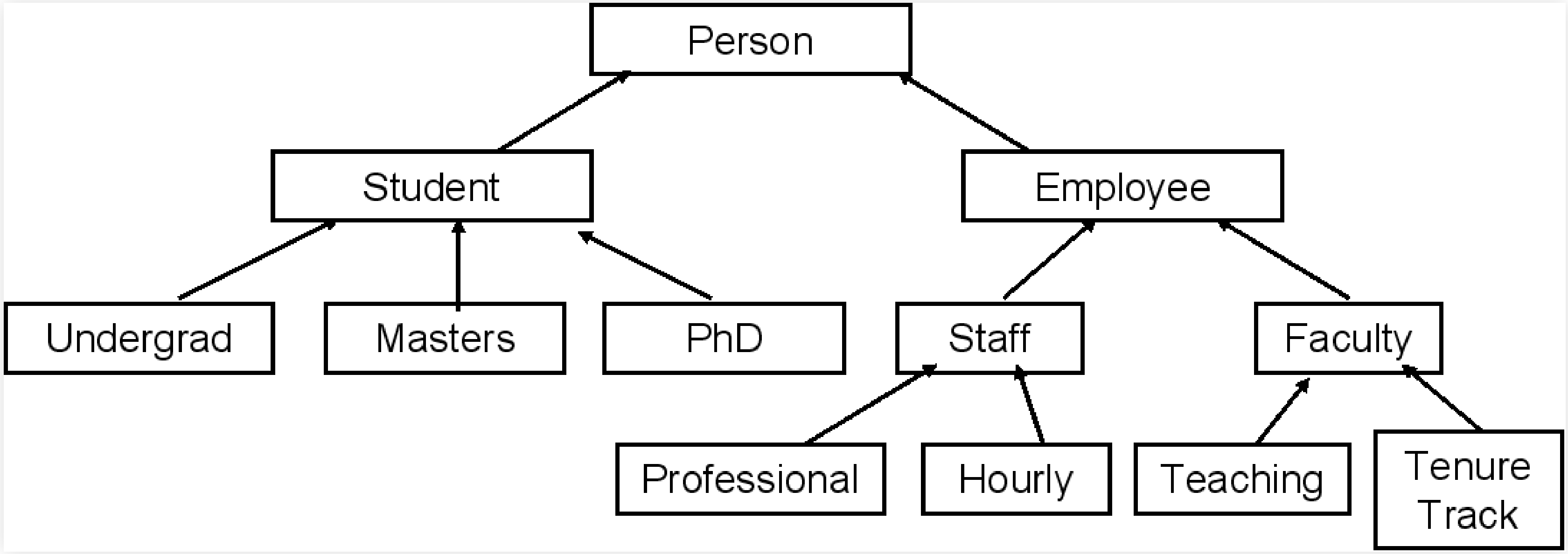
Definition

- Formal description of a domain of discourse
- Finite list of terms and relationships between these terms
- Terms represent important concepts (classes of objects) of the domain

- Example: university - students, teachers, courses, disciplines, lecture halls, etc.

Relationships in Ontologies

- Hierarchies of classes
  - e.g., undergraduates are students



Relationships in Ontologies

- Relations between classes - properties (e.g., faculty members teach courses)
- Domain and range of relations (e.g., only faculty members may teach courses)
- Disjointness of classes (e.g., faculty and staff are disjoint)
- Specifications of relationships between objects (e.g., every department must include at least ten faculty members)

Usage

- Shared understanding of a domain for disambiguation
  - Zip code vs. postal code
  - Course in university domain - degree or subject?
  - Mappings to a common ontology or between terminologies
- Improving Web Search
  - Search for concepts in the ontology instead of ambiguous keywords
  - Search for more general/special information
- Decision support systems based on logical deduction
  - Derive implicit information from explicit one
  - Explanations for inferences

Bio-Ontologies

Bio-Ontologies

Characteristics

- Describe biomedical research and medical data
- Use cases of ontologies in general applicable as well
- Additional specific usage for bio-ontologies
- Overall success story of ontology usage with a wide variety of developed ontologies

Data Integration and Search

- Bio-ontologies as controlled terminologies for terms and alternative names (recall TRAF2)
- Search for a term or synonyms

• Example - synonyms in the Gene Ontology

Name	erythrocyte development
Accession	GO:0048821
Synonyms	RBC development
	red blood cell development
Definition	The process aimed at ...

Data Integration and Search

- Standard sets of terms and synonyms

- Medical Subject Headings (MeSH) for indexing references to medical literature in PubMed
- SNOMED CT - systematic nomenclature for medicine
- Over 350,000 concepts over clinical findings, procedures, body structure, pharmaceuticals etc.
- 42 member countries (including Portugal), and many more affiliated
- National Institute of Cancer Thesaurus for unified terminology for molecular and clinical cancer research
- GoPubMed - search PubMed by indexing PubMed abstracts with Gene Ontology (GO) terms
- Retrieve abstracts according to hierarchical GO categories

Reasoning

- Apply inference rules such as transitivity
- OBO and OWL provide constructs for specifying inference rules
- For quality control of integration/aggregation of data/information
  - Detect mistakes/inconsistencies
  - Find matching terms
- Discover new facts about data or new relationships
  - New research directions or new explanations for observations

Analysis of microarray data

- Transcriptional Profiling

- Measuring the activity of all genes in the genome under multiple conditions and comparing lists of genes showing differential response (to better understand the biology of the system)
- Use the Gene Ontology to find the ontology terms that best characterize the differential genes
- Known as overrepresentation analysis/gene category analysis

Semantic similarity analysis

- Measure similarity between terms in an ontology or items annotated by these terms
  - Meaning of terms
  - Semantic structure of the ontology
  - Patterns of annotation
- Used for comparing proteins and for clinical diagnostics in medical genetics

Summary

- Challenges in data integration
- Semantic Web
- Ontologies
- Bio-ontologies

Further reading:

- Robinson and Bauer, Introduction to Bio-Ontologies, Chapter 1
- Antoniou et al., A Semantic Web Primer, Chapter 1