UZH BIO390

Semantic web, RDF, Ontologies and Knowledge Graphs in biomedical sciences

Ahmad Aghaebrahimian

Zurich University of Applied Sciences
agha@zhaw.ch

Introduction

- Ahmad Aghaebrahimian

- Research Associate at ZHAW

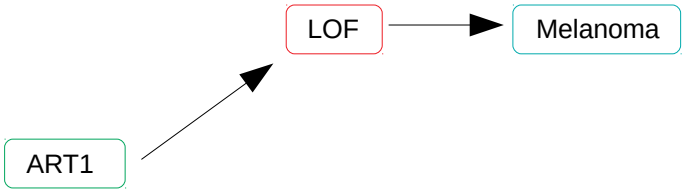
- Ph.D. Computer Sciences focusing on Computational Linguistics

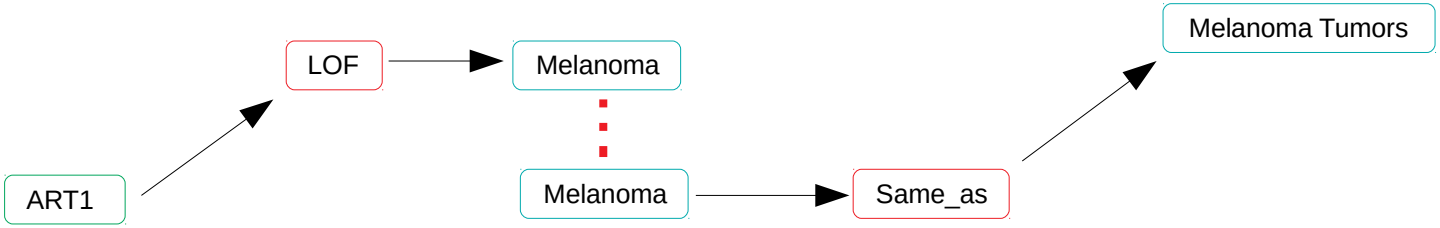
- Area of interests:
  - Machine Learning
  - Deep Neural Networks
  - Biomedical text analytics
  - Natural Language Processing
  - Semantic Web

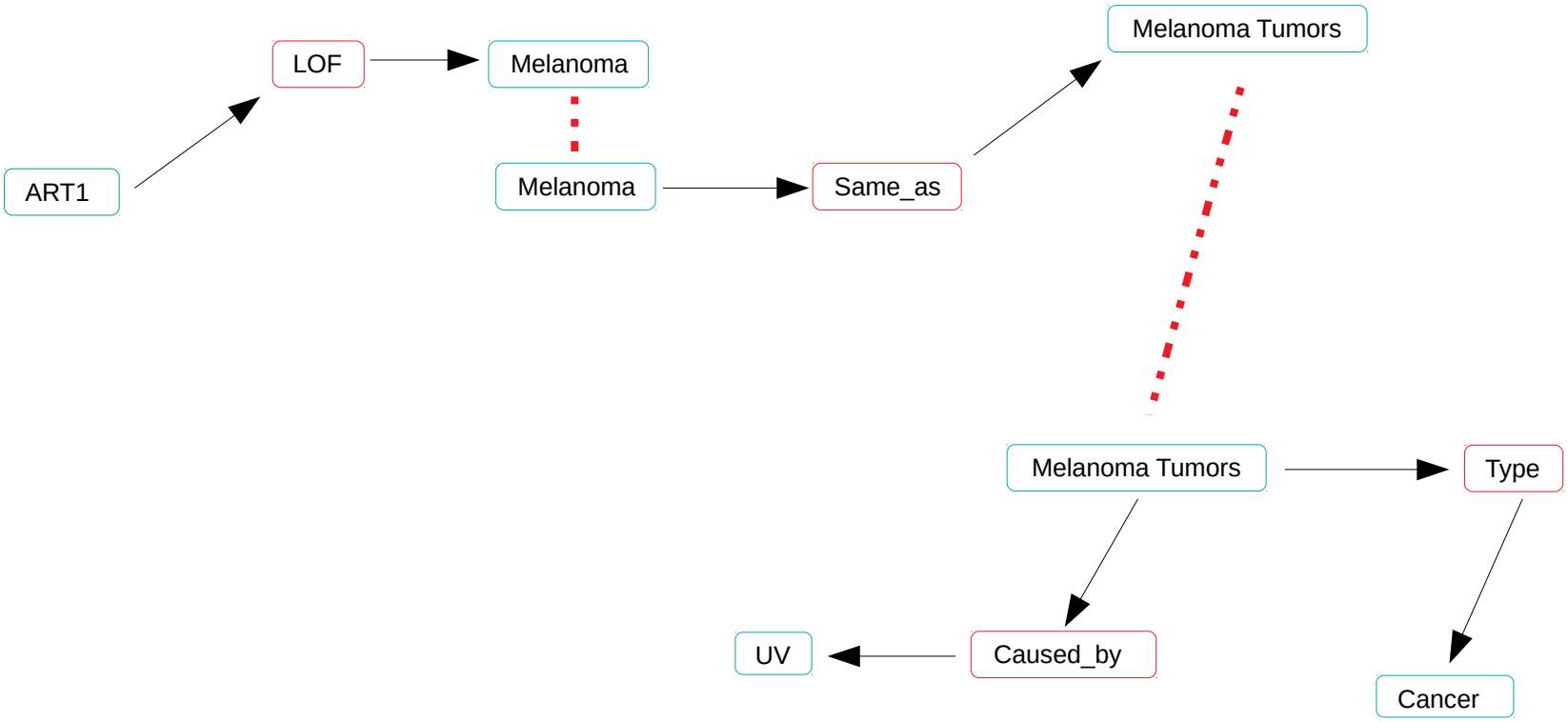
Email: agha@zhaw.ch

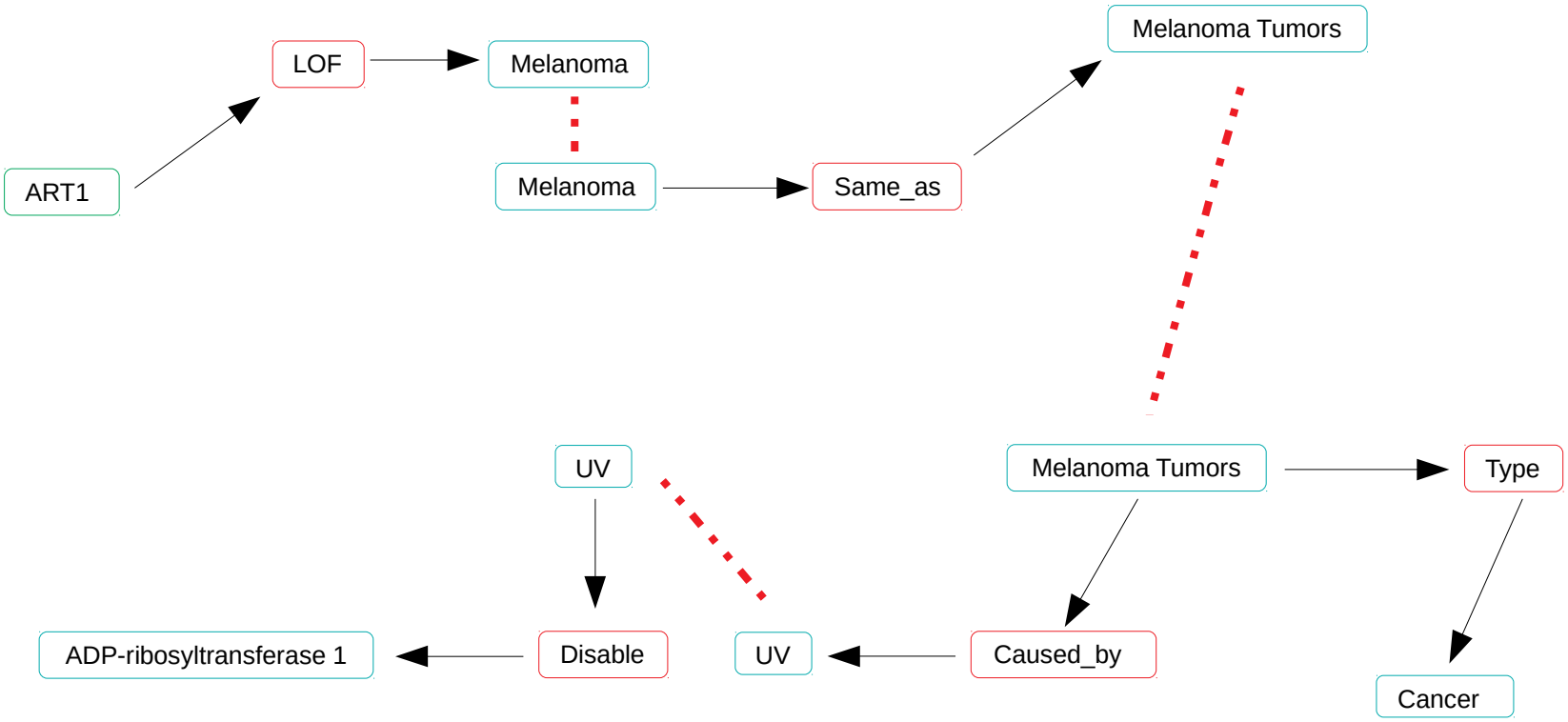
Session Content

- Introduction
- Stack of standards (URI, XML, RDF, SPARQL, OWL, …)
- RDF: Entities and Relationships
- Ontology
- Knowledge graphs


LOF
Melanoma
ART1

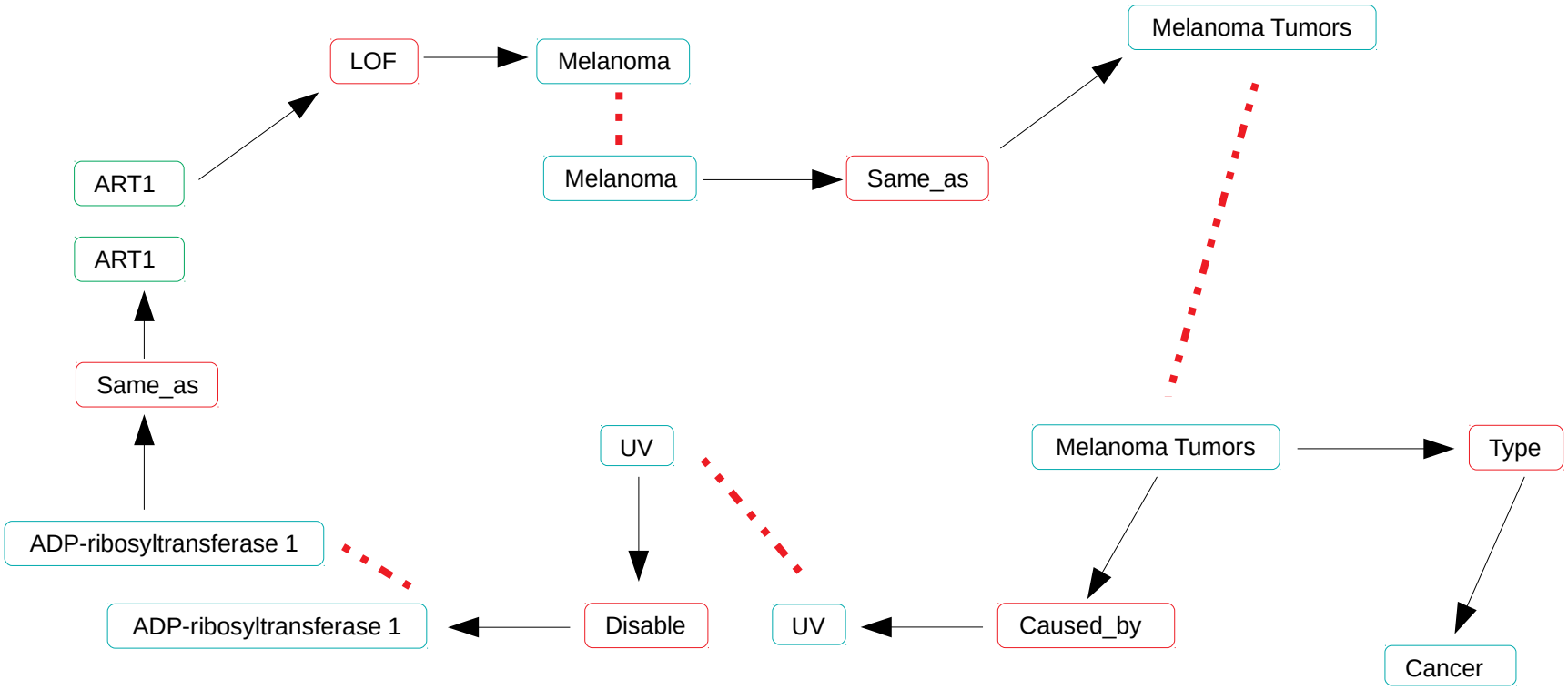

Melanoma Tumors
LOF
Melanoma
ART1
Melanoma
Same_as


ART1
LOF
Melanoma
Melanoma
Same_as
Melanoma Tumors
Melanoma Tumors
Type
UV
Caused_by
Cancer

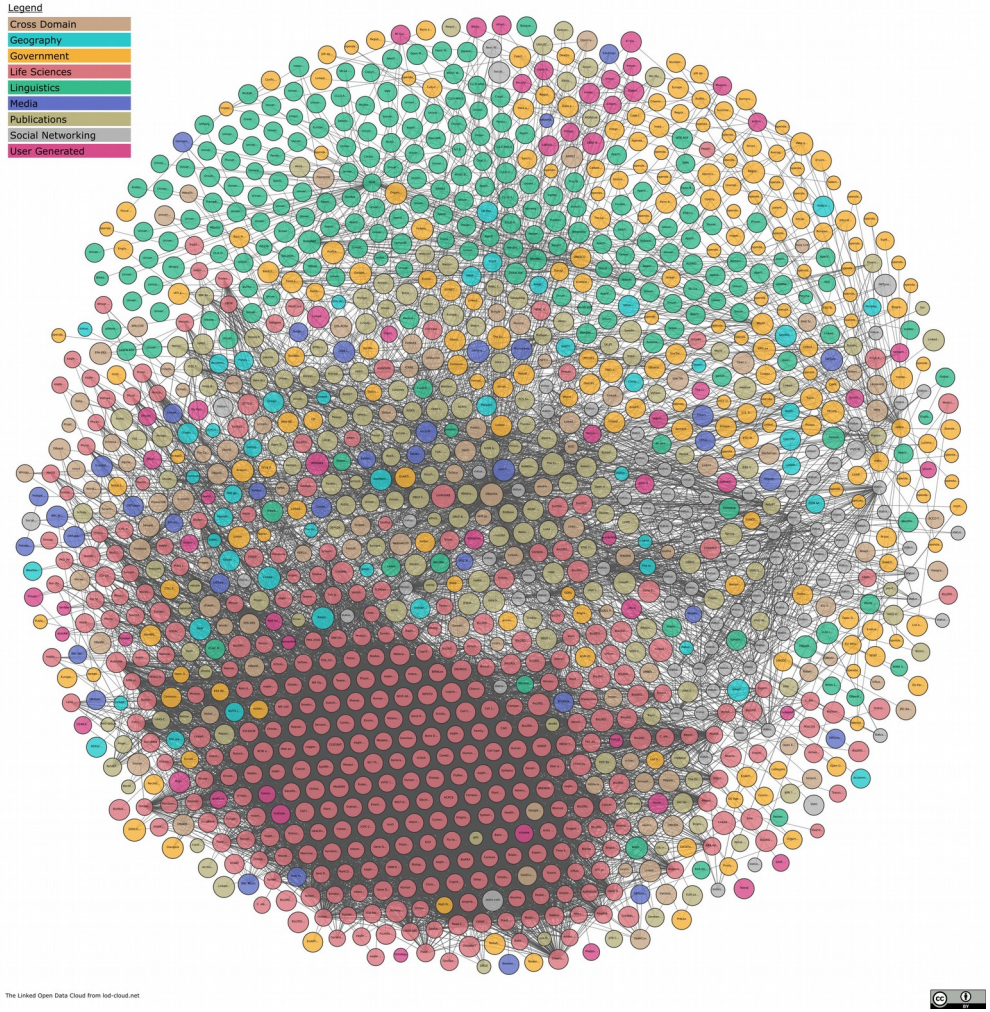

ART1
LOF
Melanoma
Melanoma
Same_as
Melanoma Tumors
Melanoma Tumors
Type
Cancer
Caused_by
UV
UV
Disable
ADP-ribosyltransferase 1

The linked open data

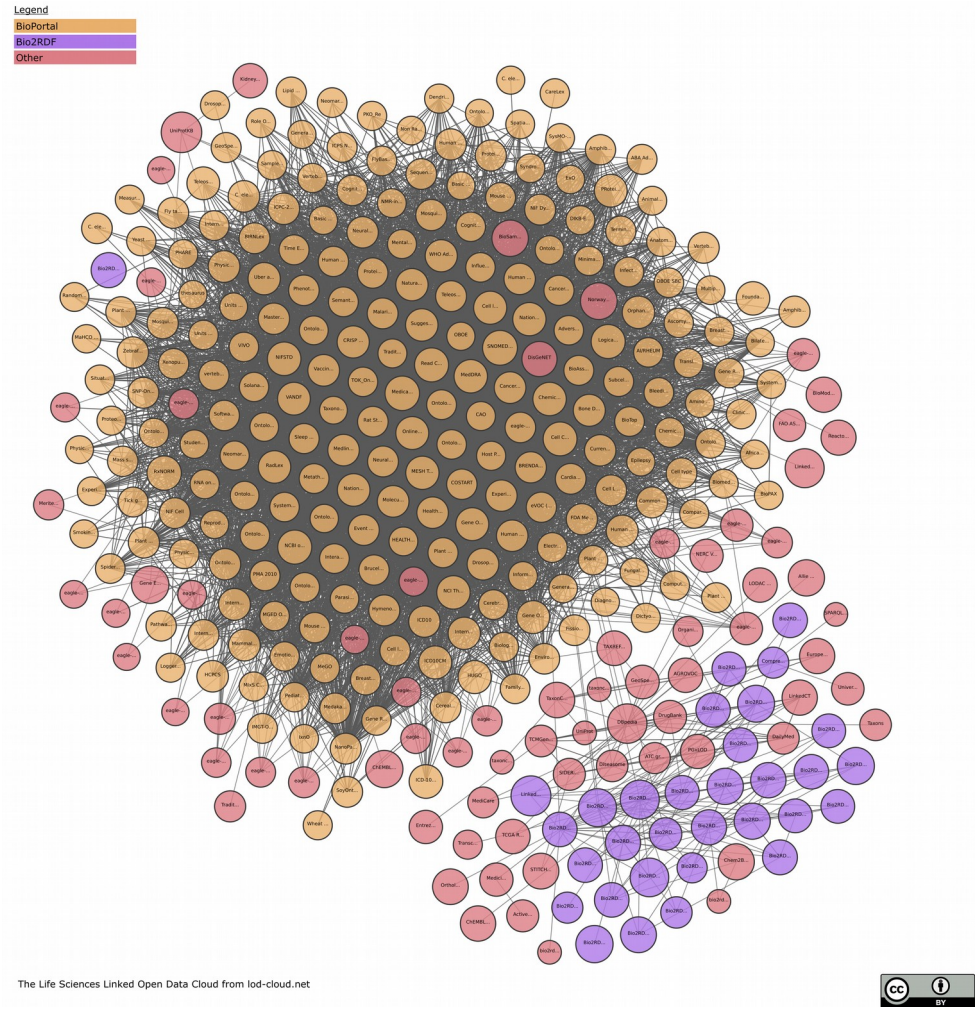
- Linked open data example



The linked open data cloud



The life sciences data cloud



Basics of the web

- Web structure:
    Server vs. Client

Basics of the web

- Web structure:
    - Server vs. Client

- Web Components:
    - Uniform Resource Locator (URL): identify document
    - Hypertext Markup Language (HTML): access document
    - Hypertext Transfer Protocol (HTTP): transfer document

Basics of the web

- Web structure:
    - Server vs. Client

- Web Components:
    - Uniform Resource Locator (URL): identify document
    - Hypertext Markup Language (HTML): access document
    - Hypertext Transfer Protocol (HTTP): transfer document

- Moving from pages to resources
    - Interactive web, Web 2.0 or semantic web

Semantic Web

Semantic Web

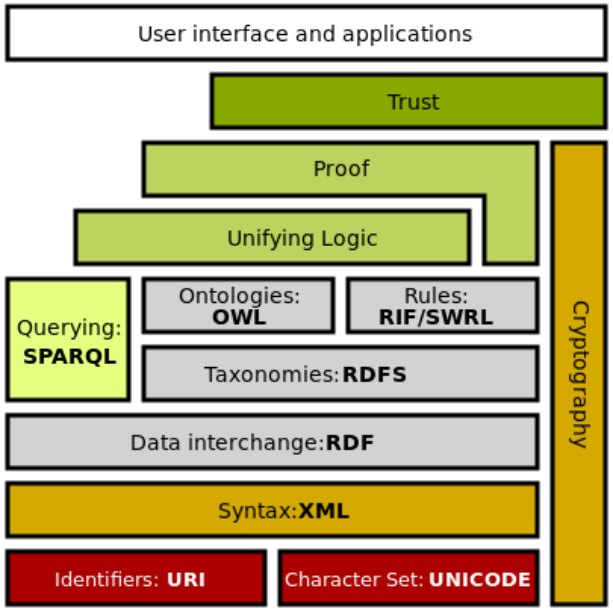
What?

Semantic Web (SW) is an extension of the World Wide Web that uses the **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)**, among other standards, to make the Internet machine-readable.

Semantic Web

What?

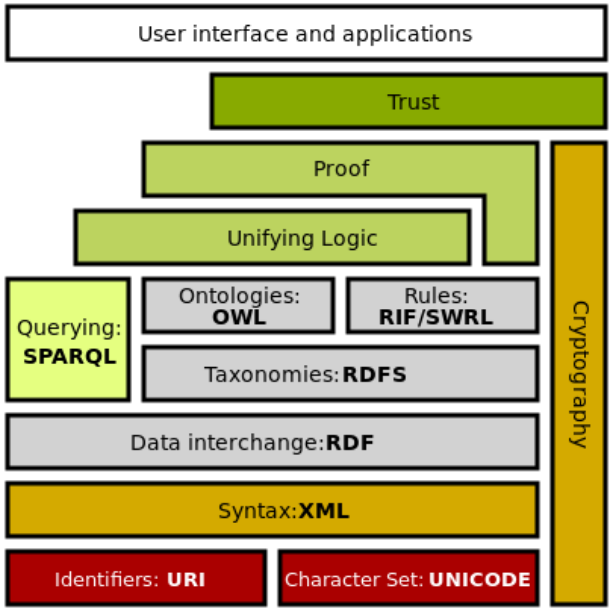
Semantic Web (SW) is an extension of the World Wide Web that uses the **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)**, among other standards, to make the Internet machine-readable.



Semantic Web

What?

Semantic Web (SW) is an extension of the World Wide Web that uses the **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)**, among other standards, to make the Internet machine-readable.

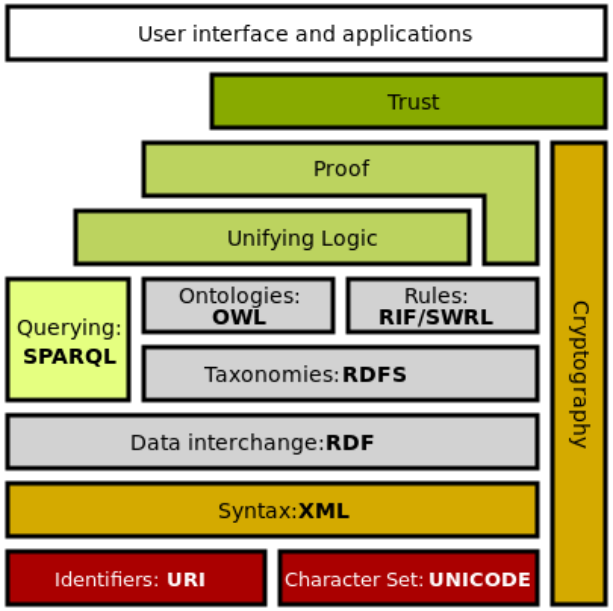


Why?

Semantic Web

What?

Semantic Web (SW) is an extension of the World Wide Web that uses the **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)**, among other standards, to make the Internet machine-readable.



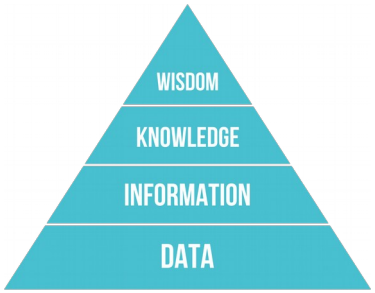
Why?

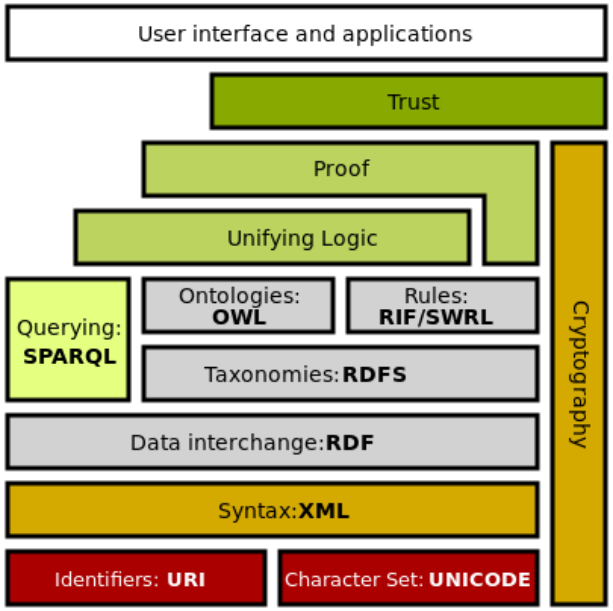
- Presenting knowledge about data

Semantic Web

What?

Semantic Web (SW) is an extension of the World Wide Web that uses the **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)**, among other standards, to make the Internet machine-readable.





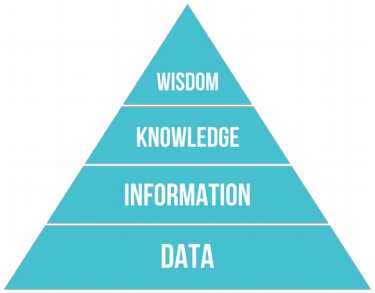
Why?

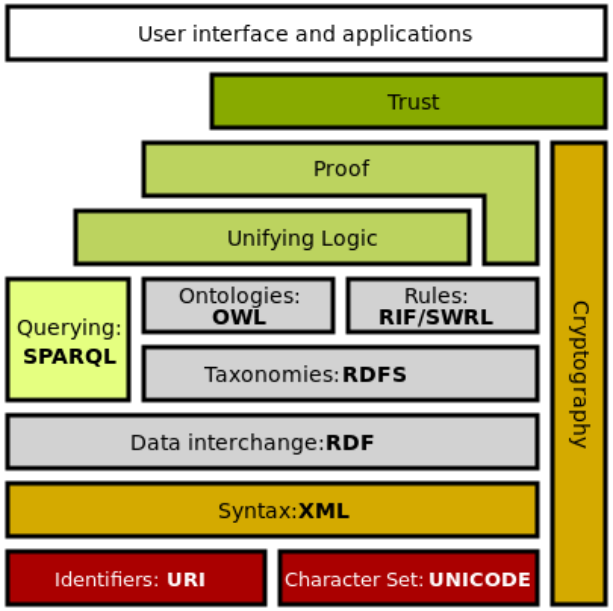
- Presenting knowledge about data

Semantic Web

What?

Semantic Web (SW) is an extension of the World Wide Web that uses the **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)**, among other standards, to make the Internet machine-readable.





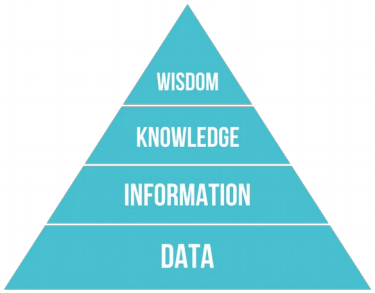
Why?

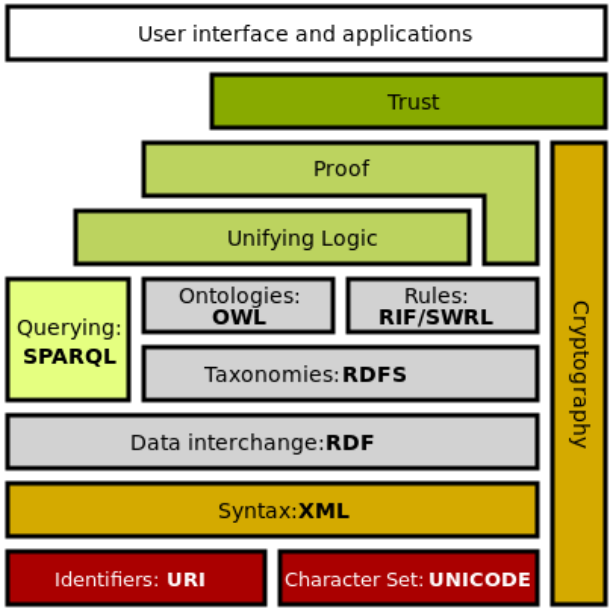
- Presenting knowledge about data
- Allowing data integration from data silos

Semantic Web

What?

Semantic Web (SW) is an extension of the World Wide Web that uses the **Resource Description Framework (RDF)** and **Web Ontology Language (OWL)**, among other standards, to make the Internet machine-readable.





Why?

- Presenting knowledge about data
- Allowing data integration from data silos
- Introduce intelligence to systems

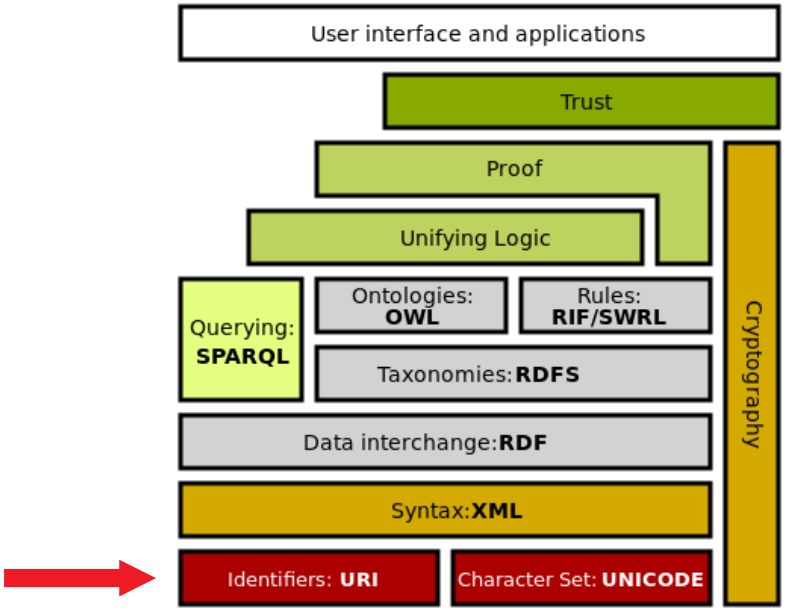
Semantic Web Standards

URI:

What is a Resource?

URL → URI → IRI

Physically located → conceptually identified → conceptually identified in all languages



Semantic Web Standards

URI:

What is a Resource?

URL → URI → IRI

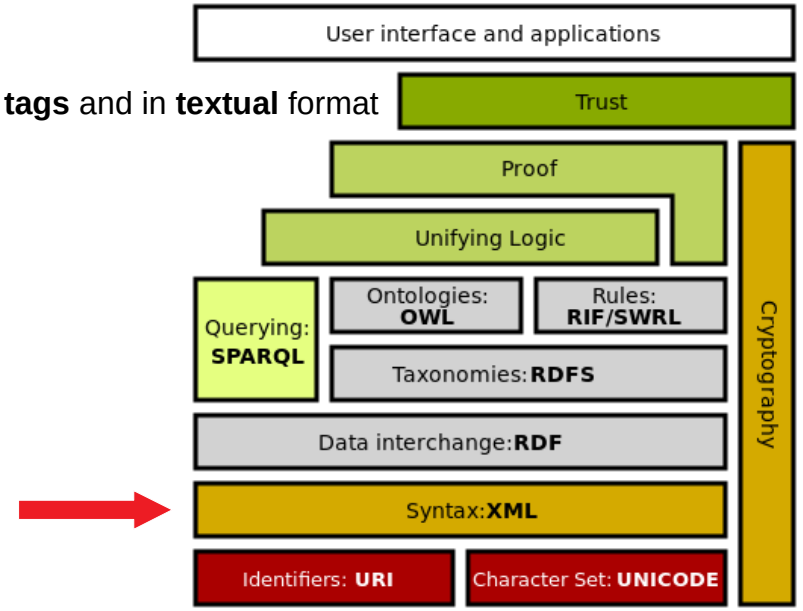
Physically located → conceptually identified → conceptually identified in all languages

XML:

Open family of languages represent **structured** data using **tags** and in **textual** format

Rules:

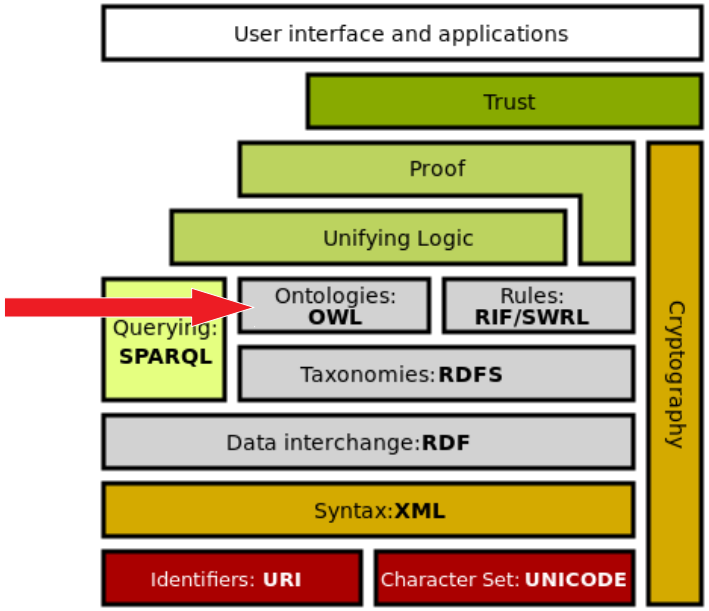
- Only one root <root> </root>
- Opening with closing <Gene></Gene>
- no tag begin with number or xml
- Case sensitive <Gene> != <gene>
- Order matters <Gene> <nucl> </nucl></Gene>
- Tags may have attributes <Gene inherited='true' />



Semantic Web Standards

OWL:

OWL provides a rich vocabulary to add semantics and context and allow reasoning and inference



Ontology

- **Ontology is**
  - A model of a domain
  - A vocabulary consisting of classes and properties
  - Machine-readable knowledge representation

Ontology

- **Ontology is**
  - A model of a domain
  - A vocabulary consisting of classes and properties
  - Machine-readable knowledge representation

- **How to build an ontology?**
  - Define a domain
  - Define the classes and properties
  - Extend existing ontology (RDF schema, dbpedia,...)

Ontology

- **Ontology is**
  - A model of a domain
  - A vocabulary consisting of classes and properties
  - Machine-readable knowledge representation

- **How to build an ontology?**
  - Define a domain
  - Define the classes and properties
  - Extend existing ontology (RDF schema, dbpedia,...)

- **Benefits of an ontology in Biomedical research? (And why they are important)**
  - Data integration
  - Language processing via domain vocabulary
  - Defining the precise meaning of classes
  - Automated processing

Ontology Continued

- **Ontology as a set of:**
  - Definitions
  - Terms and their synonyms
  - Relationships

Ontology Continued

- **Ontology as a set of:**
  - Definitions
  - Terms and their synonyms
  - Relationships
- **OBO : ChEBI** Access via : ‘https://github.zhaw.ch/agha/D-Heath’

```
[Term]
id: CHEBI:60871
name: selenium(2+)
def: "The selenium ion with two positive charges." []
synonym: "Se(2+)" RELATED [UniProt:]
synonym: "selenium dication" RELATED [ChEBI:]
synonym: "Se2+" RELATED [SUBMITTER:]
synonym: "Se" RELATED FORMULA [ChEBI:]
synonym: "[Se++]" RELATED SMILES [ChEBI:]
synonym: "InChI=1S/Se/q+2" RELATED InChI [ChEBI:]
synonym: "InChIKey=MFSBVGSNNPNWMD-UHFFFAOYSA-N" RELATED InChIKey [ChEBI:]
is_a: CHEBI:60250
is_a: CHEBI:30412

[Term]
id: CHEBI:60250
name: selenium ion
def: "A selenium atom having a net electric charge." []
is_a: CHEBI:36904
is_a: CHEBI:36914
```

Ontology Continued

- **Ontology as a set of:**
  - Definitions
  - Terms and their synonyms
  - Relationships
- **OBO : ChEBI** Access via : ‘https://github.zhaw.ch/agha/D-Heath’

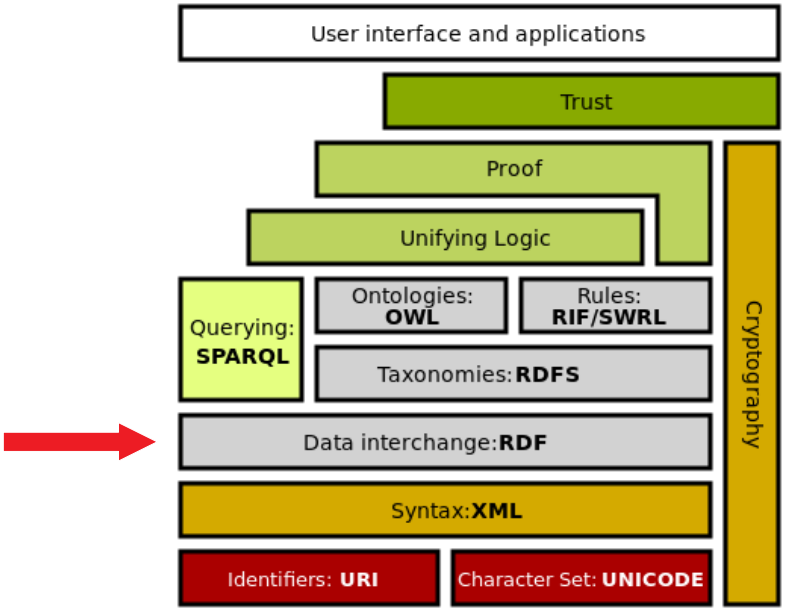
```
[Term]
id: CHEBI:60871
name: selenium(2+)
def: "The selenium ion with two positive charges." []
synonym: "Se(2+)" RELATED [UniProt:]
synonym: "selenium dication" RELATED [ChEBI:]
synonym: "Se2+" RELATED [SUBMITTER:]
synonym: "Se" RELATED FORMULA [ChEBI:]
synonym: "[Se++]" RELATED SMILES [ChEBI:]
synonym: "InChI=1S/Se/q+2" RELATED InChI [ChEBI:]
synonym: "InChIKey=MFSBVGSNNPNWMD-UHFFFAOYSA-N" RELATED InChIKey [ChEBI:]
is_a: CHEBI:60250
is_a: CHEBI:30412

[Term]
id: CHEBI:60250
name: selenium ion
def: "A selenium atom having a net electric charge." []
is_a: CHEBI:36904
is_a: CHEBI:36914
```

- **UMLS:**
  - Metathesaurus
  - Semantic network
  - Specialized Lexicon

Semantic Web Standards

RDF:

RDF is a **graph-based data model** and the set of **syntax** that allows us to write **description** about the resources on the web and to exchange them. It presents data in the **triple format** and gives it structures and unique identifiers so that data can be easily linked



Semantic Web Standards

RDF:

RDF is a **graph-based data model** and the set of **syntax** that allows us to write **description** about the resources on the web and to exchange them. It presents data in the **triple format** and gives it structures and unique identifiers so that data can be easily linked.

Principles:

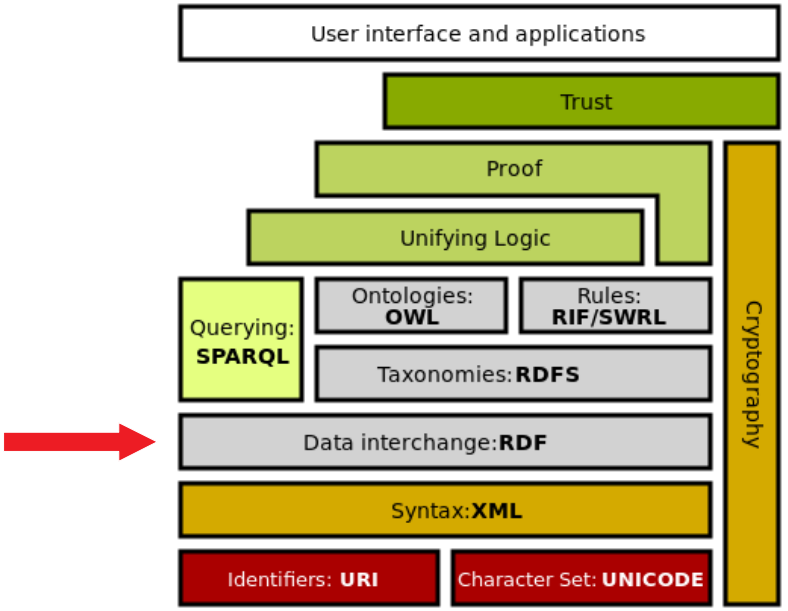
Triple structure: (subject, predicate, object)

- subject → a URI resource
- predicate → binary type URI
- object → a URI resource or literal

Predicates are labeled

Predicates are directed

RDF is a graph model



Semantic Web Standards

RDF:

RDF is a **graph-based data model** and the set of **syntax** that allows us to write **description** about the resources on the web and to exchange them. It presents data in the **triple format** and gives it structures and unique identifiers so that data can be easily linked.

Principles:

Triple structure: (subject, predicate, object)

- subject → a URI resource
- predicate → binary type URI
- object → a URI resource or literal

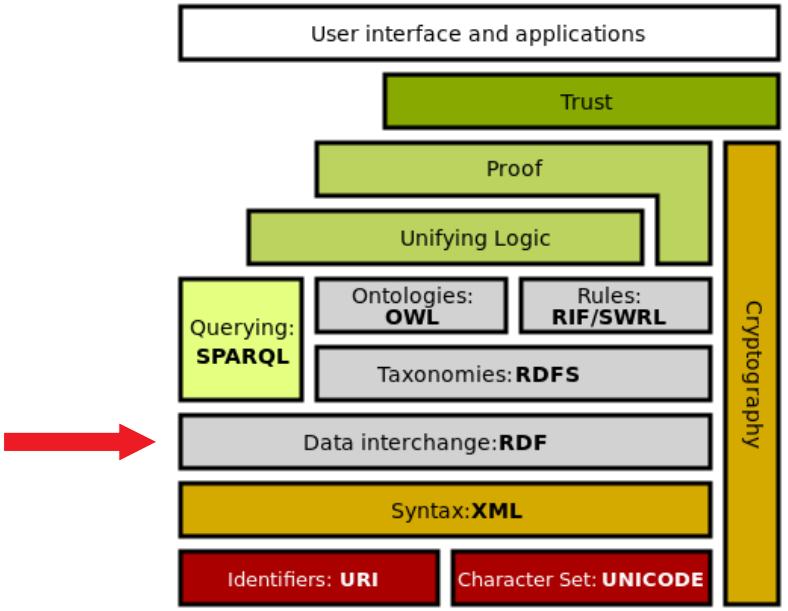
Predicates are labeled

Predicates are directed

RDF is a graph model

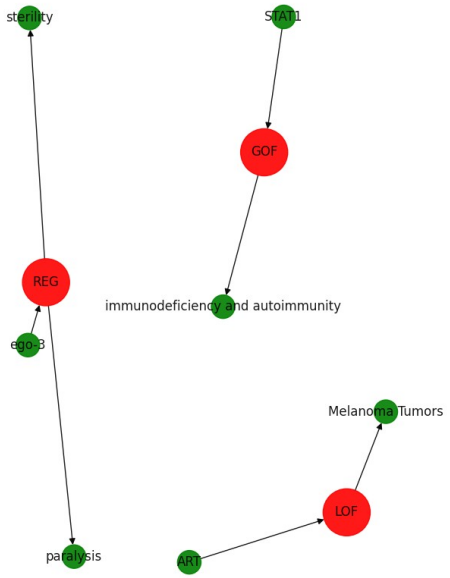
RDF serialization:

XML, N-triple, Turtle, TriG, JSON-LD



The Graph data model

- Storing data in form of triplets: (Subject, Predicate, Object)
  e.g. (ART, LOF, Melanoma_Tumors)
  Subject and Predicate must be in URI form



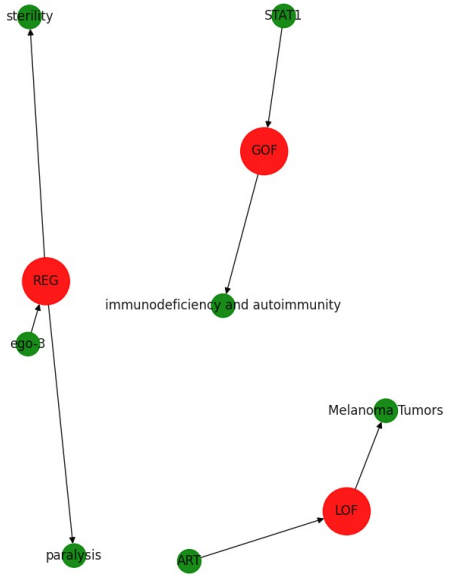
The Graph data model

- Storing data in form of triplets: (Subject, Predicate, Object)
  e.g. (ART, LOF, Melanoma_Tumors)
  Subject and Predicate must be in URI form

- Triplets follow the **RDF** standard.

- Triplets are easily **Expanded** and **Interlinked**.

- Triplets can be queried via **SPARQL**:



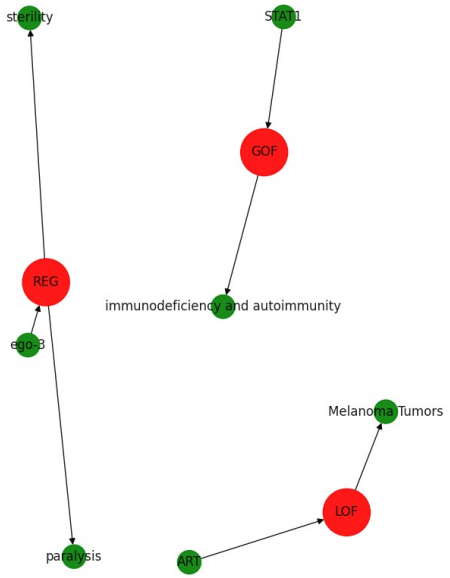
The Graph data model

- Storing data in form of triplets: (Subject, Predicate, Object)
  e.g. (ART, LOF, Melanoma_Tumors)
  Subject and Predicate must be in URI form

- Triplets follow the **RDF** standard.

- Triplets are easily **Expanded** and **Interlinked**.

- Triplets can be queried via **SPARQL**:




```
SELECT ?gene ?relation
WHERE {
        ?gene ?relation Melanoma_Tumors .
}
```

Subjects, Predicates, Objects

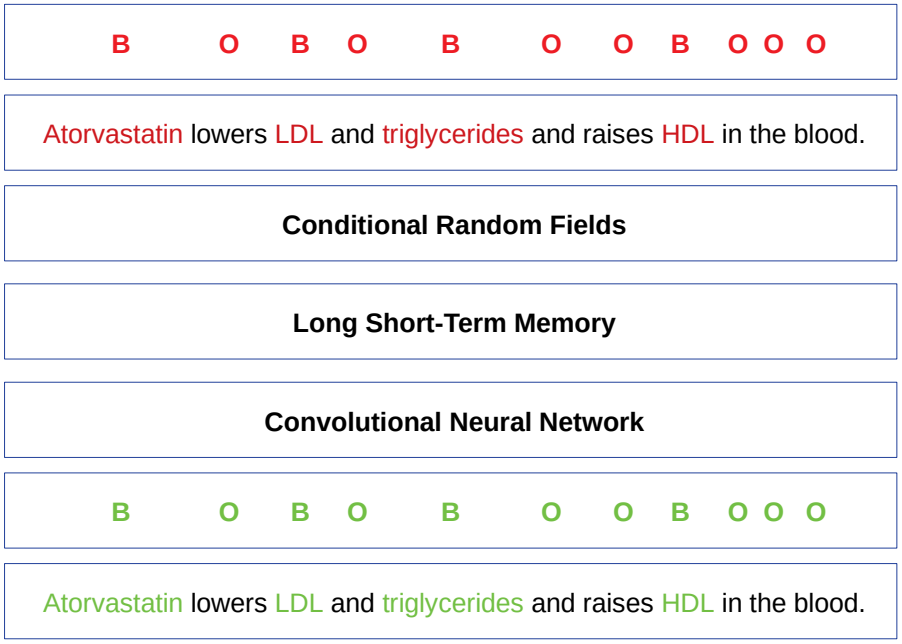
Named Entity

Relationship

NLP components:

- Named Entity Recognition (NER)
- Named Entity Disambiguation (NED)
- Relation Extraction (RE)

Named Entity Recognition (NER)



NER Evaluation:

Accuracy:

$$Accuracy = \frac{|Correct\ answers|}{|test\ data|}$$

NER Evaluation:

Accuracy:

$$Accuracy = \frac{|Correct\ answers|}{|test\ data|}$$

F1 score:

Example: Cancer Diagnostics

True labels

Predicted labels	Cancer	No-cancer
Cancer	TP	FP
No-cancer	FN	TN

NER Evaluation:

Accuracy:

$$Accuracy = \frac{|Correct\ answers|}{|test\ data|}$$

F1 score:

Example: Cancer Diagnostics

$$Precision = \frac{|predicted\ correct\ answers|}{|all\ detected\ answers|}$$

$$Recall = \frac{|predicted\ correct\ answers|}{|all\ correct\ answers|}$$

True labels

Predicted labels	Cancer	No-cancer
Cancer	TP	FP
No-cancer	FN	TN

NER Evaluation:

Accuracy:

$$Accuracy = \frac{|Correct\ answers|}{|test\ data|}$$

F1 score:

Example: Cancer Diagnostics

$$Precision = \frac{|predicted\ correct\ answers|}{|all\ detected\ answers|}$$

$$Recall = \frac{|predicted\ correct\ answers|}{|all\ correct\ answers|}$$

True labels

Predicted labels	Cancer	No-cancer
Cancer	TP	FP
No-cancer	FN	TN

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Named Entity Disambiguation (NED)

Atorvastatin lowers LDL and triglycerides and raises HDL in the blood.

CHEBI:39548 lowers CHEBI:47774 and IUPAC:46823 and raises CHEBI:47775 in the blood.

Named Entity Disambiguation (NED)

Atorvastatin lowers LDL and triglycerides and raises HDL in the blood.

CHEBI:39548 lowers CHEBI:47774 and IUPAC:46823 and raises CHEBI:47775 in the blood.

Problem:

- Different order
- Morphological forms
- Synonymous names
- Abbreviation

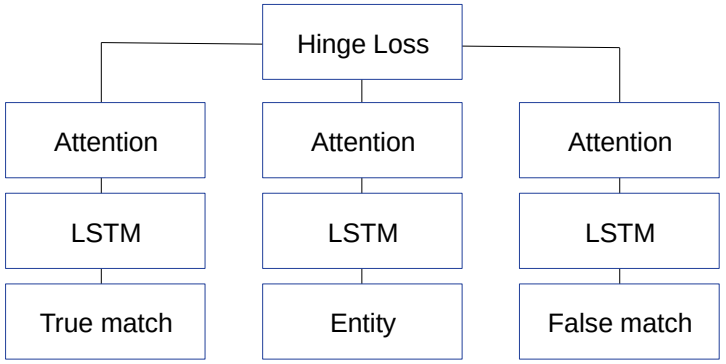
Named Entity Disambiguation (NED)

Atorvastatin lowers LDL and triglycerides and raises HDL in the blood.

CHEBI:39548 lowers CHEBI:47774 and IUPAC:46823 and raises CHEBI:47775 in the blood.

Problem:

- Different order
- Morphological forms
- Synonymous names
- Abbreviation



Aghaebrahimian, A., Cieliebak, M.(2020), Named Entity Disambiguation at Scale, **ANNPR**, Winterthur, Switzerland

Relation Extraction (RE)

Atorvastatin lowers LDL and triglycerides and raises HDL in the blood.

Atorvastatin lowers LDL and triglycerides and raises HDL in the blood.

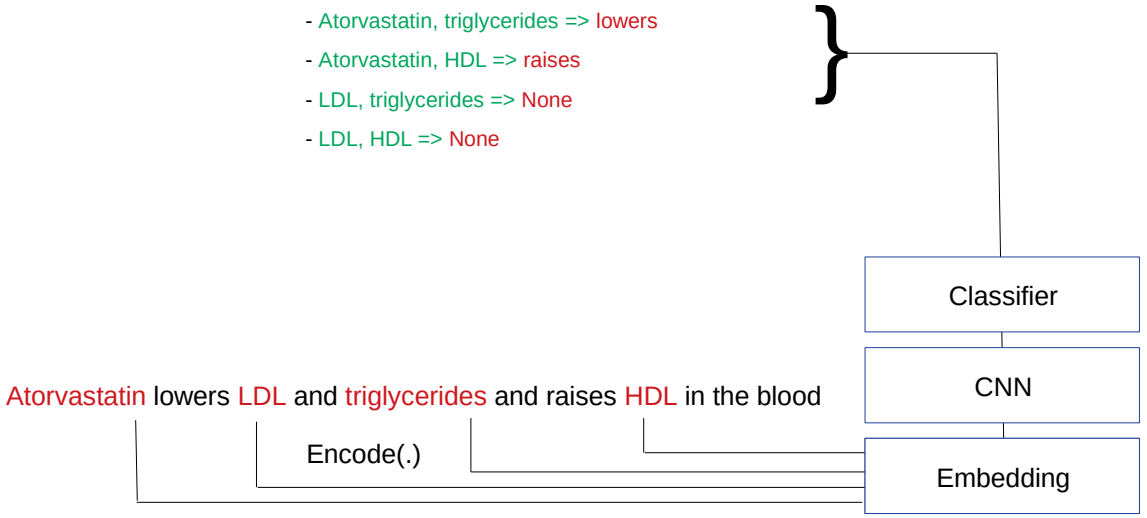
Relation Extraction (RE)

Atorvastatin lowers LDL and triglycerides and raises HDL in the blood.

Atorvastatin lowers LDL and triglycerides and raises HDL in the blood.

Single hop RE:

- Atorvastatin, LDL => lowers
- Atorvastatin, triglycerides => lowers
- Atorvastatin, HDL => raises
- LDL, triglycerides => None
- LDL, HDL => None



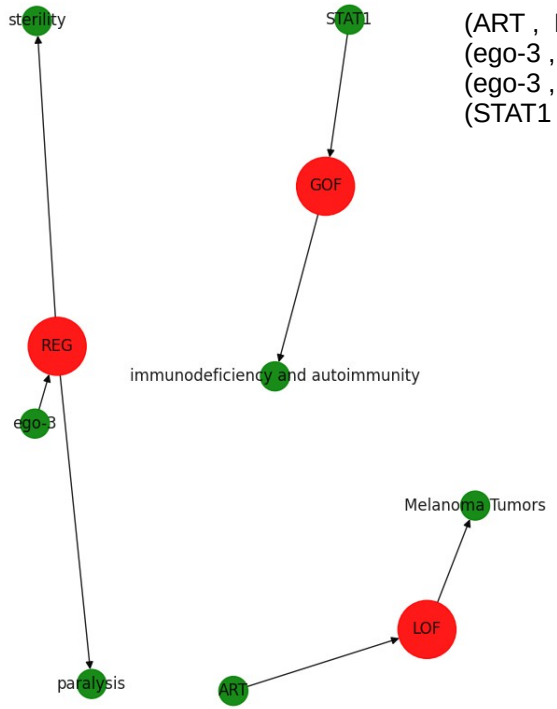
Aghaebrahimian, A. and Jurcicek, F., (2016), Open-domain Factoid Question Answering via Knowledge Graph Search, **NAACL**, San Diego, USA

Knowledge Graph

Collection of billions of triplet graph structures known as **Assertion** modeled in the **RDF** model

Knowledge Graph

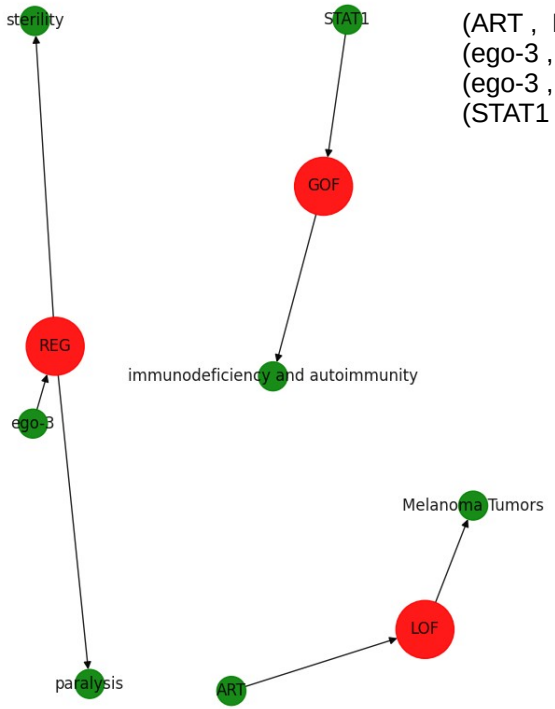
Collection of billions of triplet graph structures known as **Assertion** modeled in the **RDF** model



(ART , LOF, Melanoma Tumors)
(ego-3 , REG , paralysis)
(ego-3 , REG , sterility)
(STAT1 , GOF , immunodeficiency and autoimmunity)

Knowledge Graph

Collection of billions of triplet graph structures known as **Assertion** modeled in the **RDF** model

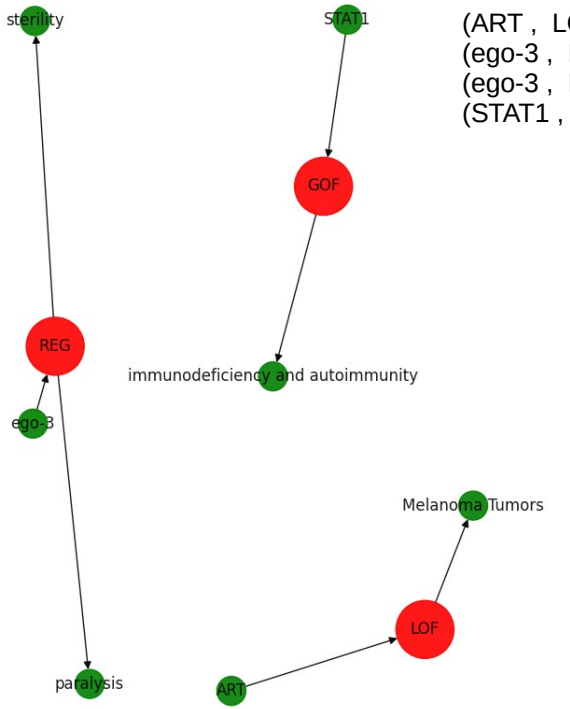


(ART , LOF, Melanoma Tumors)
(ego-3 , REG , paralysis)
(ego-3 , REG , sterility)
(STAT1 , GOF , immunodeficiency and autoimmunity)

Which genes are related to paralysis?
How STAT1 impacts the immune system?

Knowledge Graph

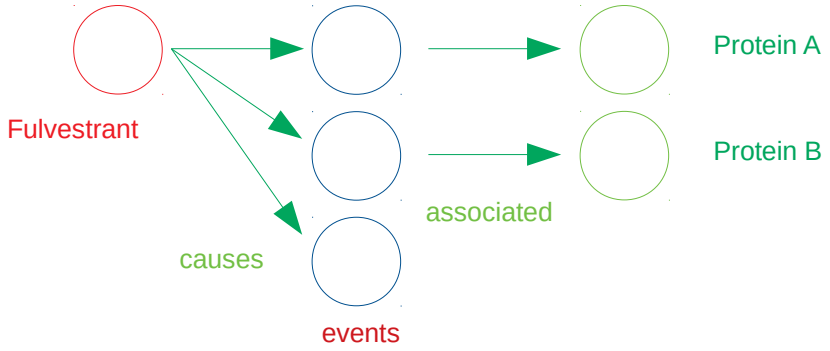
Collection of billions of triplet graph structures known as **Assertion** modeled in the **RDF** model



(ART , LOF, Melanoma Tumors)
(ego-3 , REG , paralysis)
(ego-3 , REG , sterility)
(STAT1 , GOF , immunodeficiency and autoimmunity)

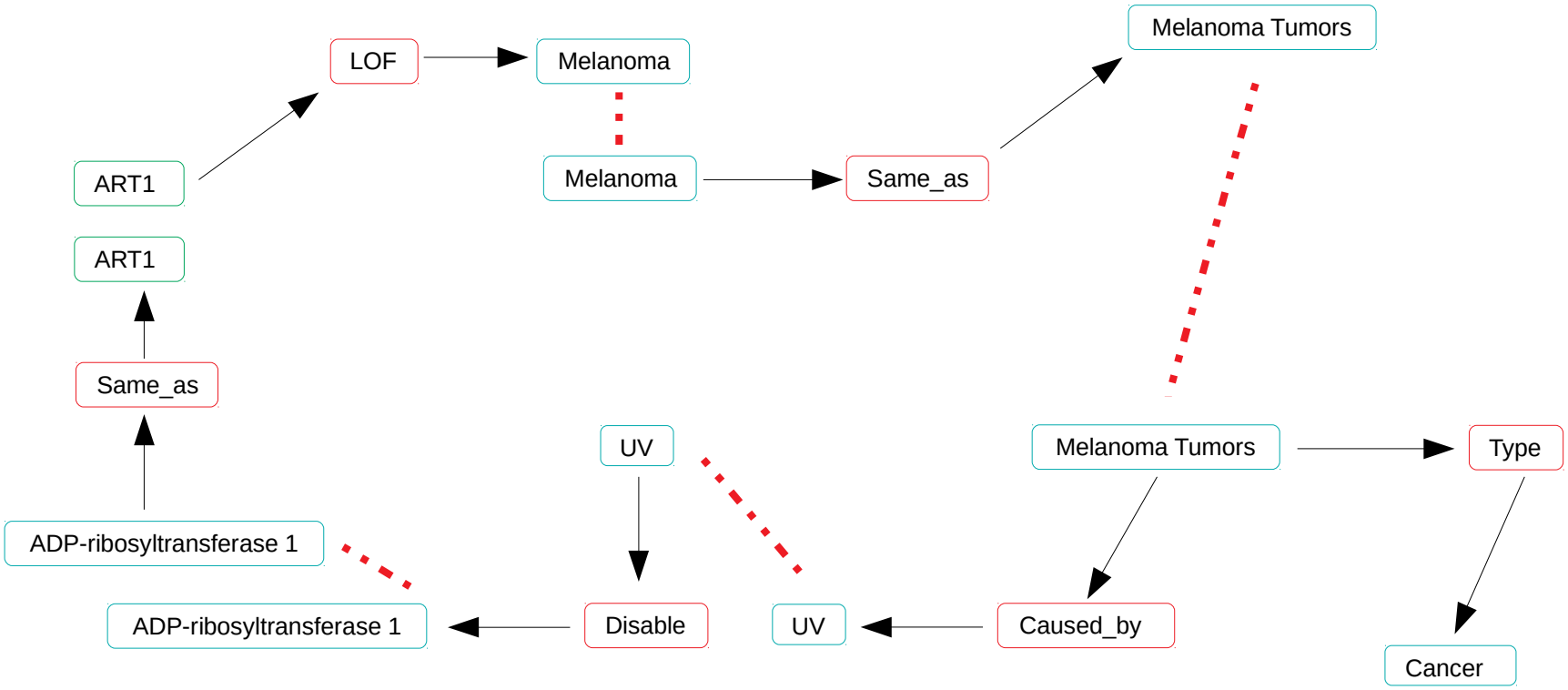
Which genes are related to paralysis?
How STAT1 impacts the immune system?

What proteins are associated with adverse events caused by Fulvestrant?



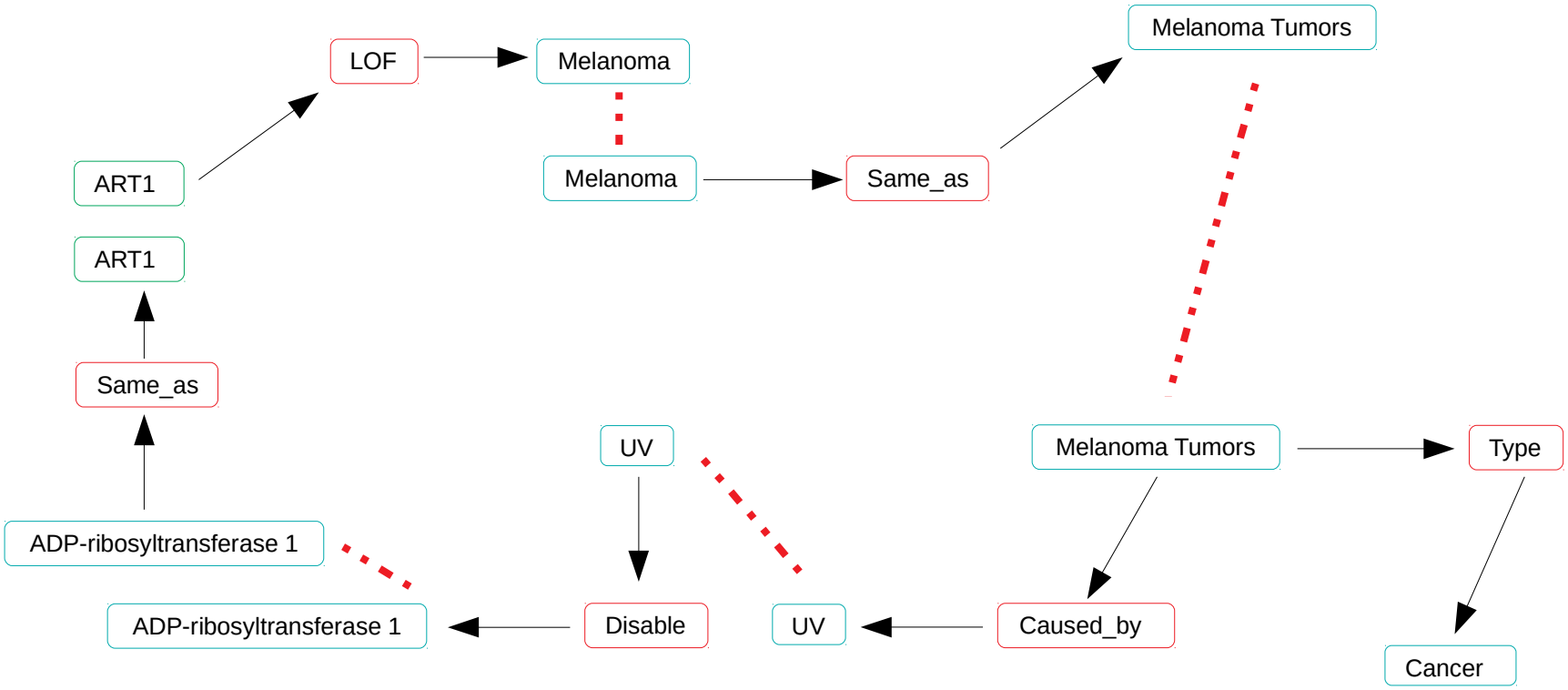
The linked open data

- Linked open data example



The linked open data

- Linked open data example



Question: How do we know that the dotted entities are the same entities.

Semantic Web tools

- RDFa:

Extracting triples from HTML pages via markups

https://rdfa.info/play/

- Gleaning Resource Descriptions from Dialects of Languages (GRDDL):

Algorithms instead of markups

<link rel="transformation" href="http://www.w3.org/2000/06/dc-extract/dc-extract.xsl" />

- JSON for Linked Data: JSON-LD

Attaching context to JSON files

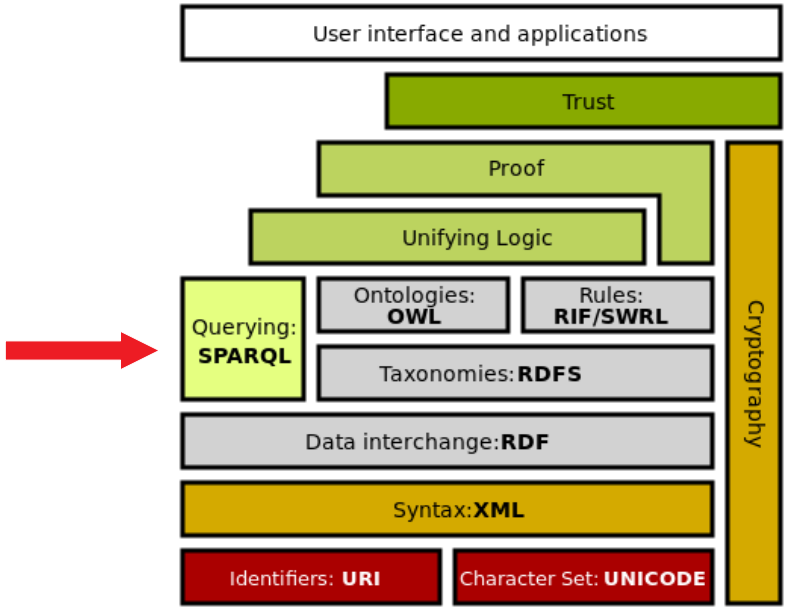
- R2RML: Transforming tables to RDF

SPARQL

W3C standard

SPARQL Protocol And RDF Query Language

Lab work: https://bit.ly/3wjyHpf



Life Sciences RDF data and SPARQL Endpoints

A SPARQL endpoint gets queries and returns their results using HTTP protocol

- Generic
  - http://sparql.org/sparql.html
  - http://demo.openlinksw.com/sparql
- Specific
  - Dbpedia
    - https://dbpedia.org/sparql
  - SIB Swiss Institute of Bioinformatics
    - UniProt: http://sparql.uniprot.org
    - neXtProt: http://snorql.nextprot.org
  - EBI European Bioinformatics Institute:
    - BioSamples, BioModels, ChEMBL, Expression Atlas, Reactome, Ensembl
    - https://www.ebi.ac.uk/rdf/services/sparql
  - NCBI National Center for Biotechnology Information:
    - PubChemRDF (rdf only, no SPARQL endpoint)
    - https://pubchem.ncbi.nlm.nih.gov/rdf/
  - http://sparql-playground.sib.swiss/