SARDIGNA CHIRCAS
SARDEGNA RICERCHE

open:campus

Artificial Intelligence for developers

8 weekend per diventare Machine Learning Specialist

Natural Language Processing

Maurizio Atzori
Università degli Studi di Cagliari
atzori@unica.it

February 9-10 , 2024

Outline of the course

- Intro on AI, ML and NLP
- Text Processing
- Words and Corpora
- Lexical similarity
- Language Modeling
- Text Classification
- Semantic similarity
- Knowledge Graphs
- Intro to Large Language Models

Intro to Semantic Web

Knowledge Graphs

Semantic Web
(huge distributed knowledge graph)

- It is about giving semantics to information in the Web
  - In the Web, URLs correspond to resources *(mostly html pages)*
  - This forms a big network among pages *(through links)*
- In the Semantic Web a URL indicates a concept.
  - a URL may represent an **entity** or a **property**

Semantic Web: an apple

(examples of Entities)

- an apple
  - <http://dbpedia.org/resource/Apple>
  - <http://www.wikidata.org/entity/Q89>
  - <http://www.dbpedialite.org/things/18978754>
- another kind of "Apple"
  - <http://dbpedia.org/resource/Apple_Inc.>
  - <http://www.wikidata.org/entity/Q312>
  - <http://www.dbpedialite.org/things/856>
- no room for ambiguity
- we can define prefixes to shorten urls:
  - dbpedia:Apple

Semantic Web: not only entities *(examples of Properties, aka Attributes)*

- How much proteinic is something?
  - it is relevant to "apple" (Q89), but also other foods
    - <http://dbpedia.org/property/protein>
  - is it relevant to the other "apple" (Q312)?

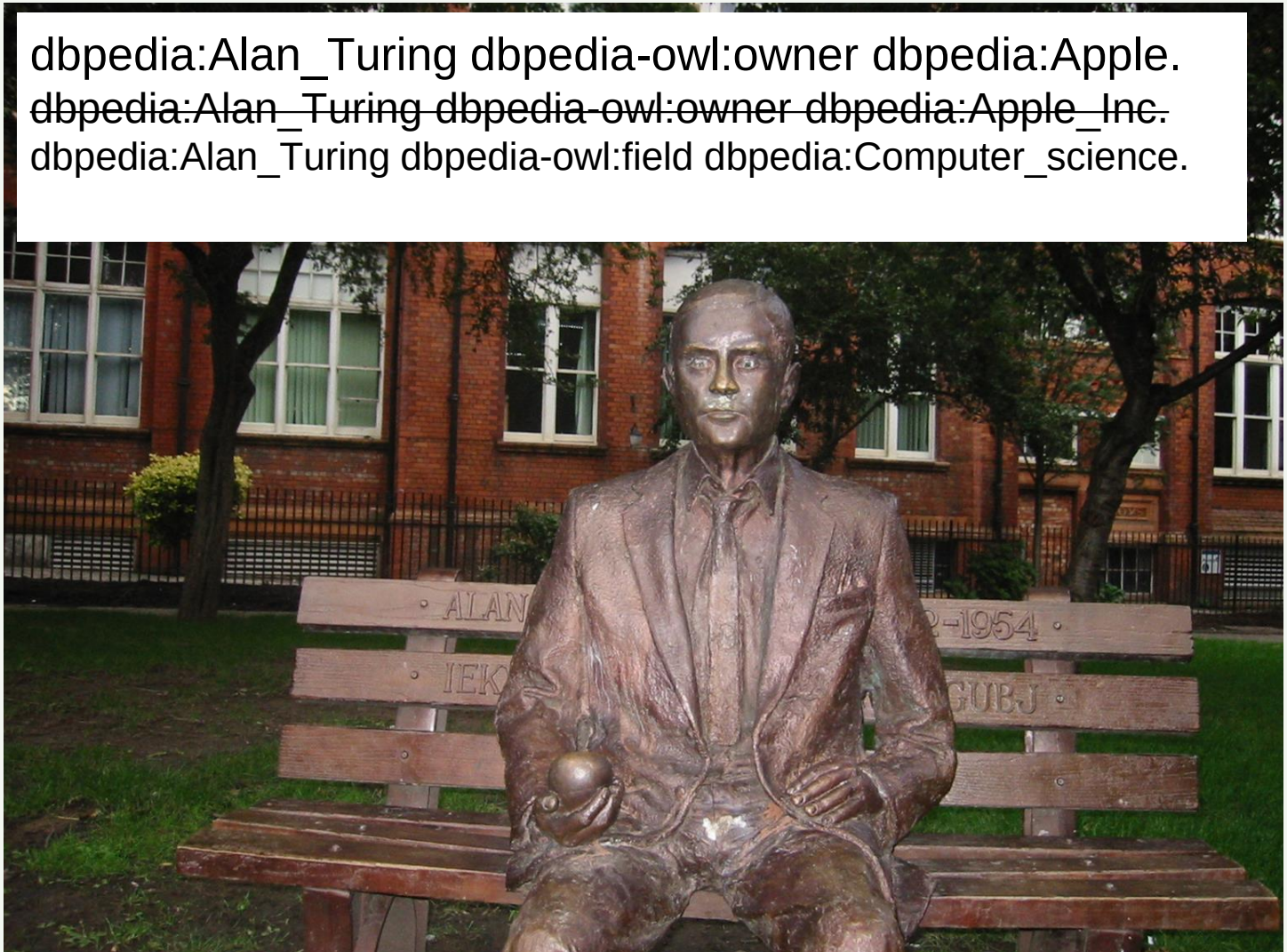
- Properties can be themselves entities.
  - What is the range, unit measure, the english word for attribute <http://dbpedia.org/property/protein> ?

Resource Description Framework (RDF)

- Entities and properties are linked together through RDF "data format"
- Multigraph-like structure made of triples
  - subject, predicate, object

not a graph (not even multigraph) in the CS sense (edge labels are also nodes in RDF)

dbpedia:Alan_Turing dbpedia-owl:owner dbpedia:Apple.
~~dbpedia:Alan_Turing dbpedia-owl:owner dbpedia:Apple_Inc.~~
dbpedia:Alan_Turing dbpedia-owl:field dbpedia:Computer_science.



An example of RDF dataset: DBpedia

http://dbpedia.org/resource/Apple (Dereferencing a URL)

About: Apple

An Entity of Type : species, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

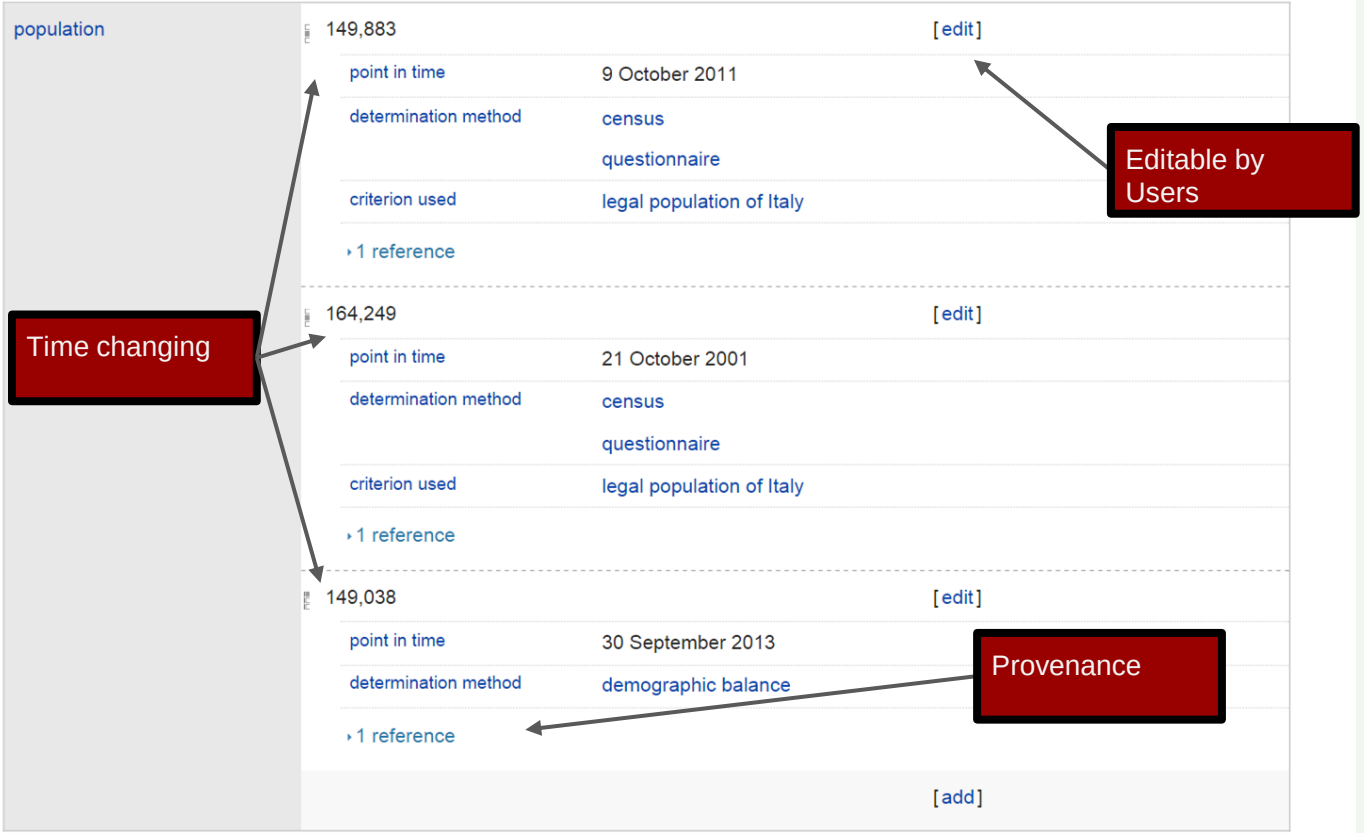
dbpprop:note	http://ndb.nal.usda.gov/ndb/search/list?qlookup=09003&format=Full
dbpprop:opt1n	dbpedia:Fluoride
dbpprop:opt1v	3.300000 (xsd:double)
dbpprop:ordo	dbpedia:Rosales
dbpprop:pantothenicMg	0.061000 (xsd:double)
dbpprop:phosphorusMg	11 (xsd:integer)
dbpprop:potassiumMg	107 (xsd:integer)
dbpprop:protein	0.26
dbpprop:q	Apples
dbpprop:regnum	Plantae
dbpprop:riboflavinMg	0.026000 (xsd:double)
dbpprop:s	1911 (xsd:integer)
dbpprop:sign	dbpedia:Plato
dbpprop:sodiumMg	1 (xsd:integer)

DBpedia

A Nucleus for a Web of Open Data (ISWC 2007, Semantic Web Journal 2014)

- Univ of Leipzig, Univ of Mannheim, OpenLink SW
- 583 million “facts” in terms of RDF triples (en14)
  - 3 billions is the union of the 125 localized versions
- 4.58 million entities (en14)
- some triples imported from other datasets
- Introduces an ontology
  - e.g. :Actor :subClassOf :Person

WikiData: Cagliari (city)



SPARQL Protocol and RDF Query Language

Quante proteine ha una mela?

- endpoint: http://dbpedia.org/sparql
- query (proteins quantity of an apple):

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbpprop: <http://dbpedia.org/property/>

SELECT ?proteins WHERE {
     dbpedia:Apple dbpprop:protein ?proteins.
}
```

- graph pattern matching
  - variables can appear in the triple in any point

SPARQL: People who were born in Berlin before 1900

```
PREFIX dbo: <http://dbpedia.org/ontology/>.
PREFIX dbr: <http://dbpedia.org/resource/>.
SELECT ?name ?birth ?death ?person WHERE {
      ?person dbo:birthPlace dbr:Berlin .
      ?person dbo:birthDate ?birth .
      ?person foaf:name ?name .
      ?person dbo:deathDate ?death .
      FILTER (?birth < "1900-01-01"^^xsd:date) .
}
ORDER BY ?name
```

SPARQL 1.1 also supports features such as aggregates, having, group by, built-in and user-defined functions, path properties, unions, intersections, difference, etc.

SPARQL on DBpedia

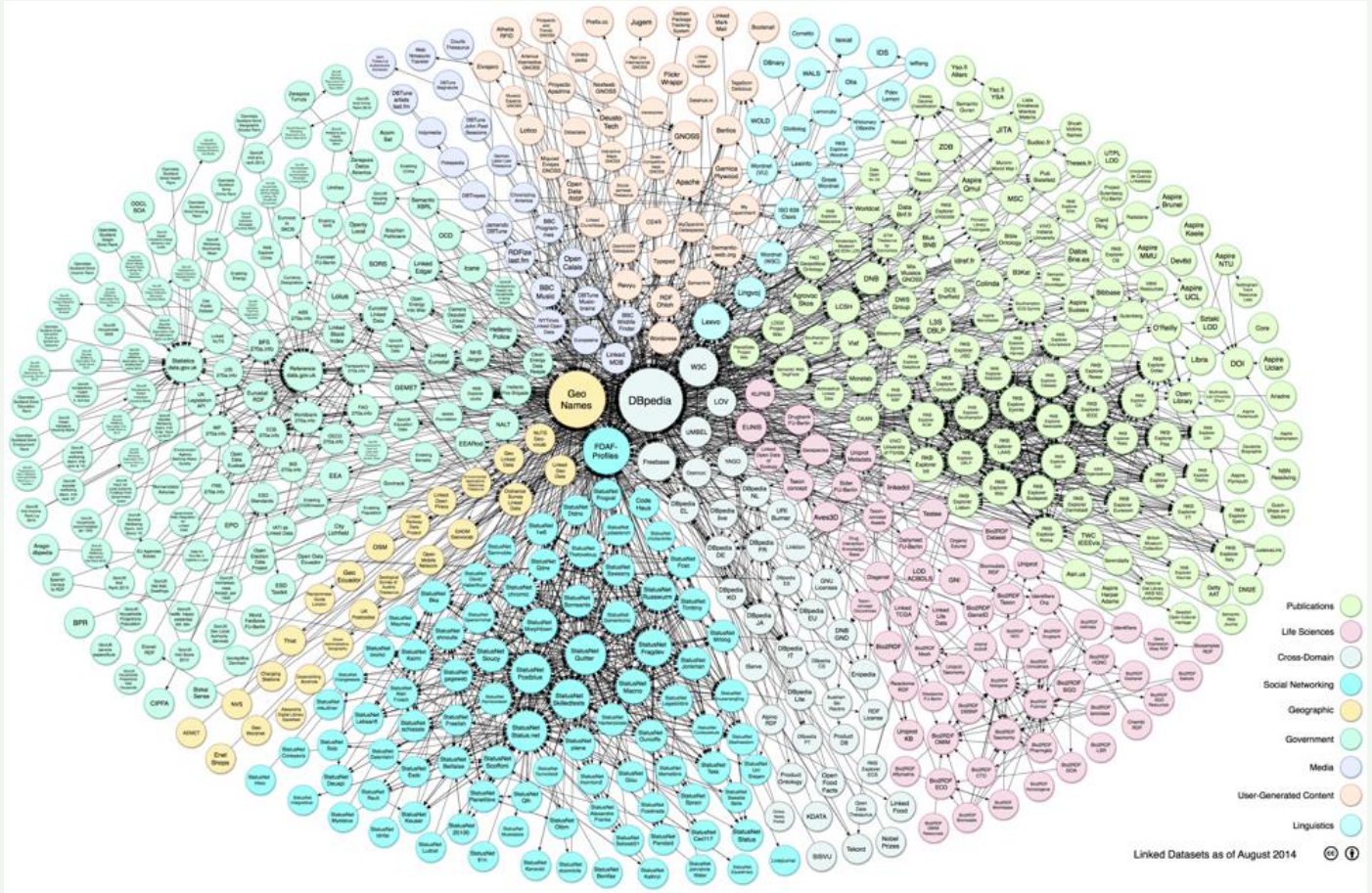
Go to: https://dbpedia.org/sparql

And query:

```
prefix dbpedia: <http://dbpedia.org/resource/>
select * where {
  dbpedia:Cagliari <http://dbpedia.org/property/mayor> ?x }
LIMIT 100
```

Moltissimi dati disponibili come Linked Data

- Do we have data in a structured form?
  - DBpedia (information from Wikipedia)
  - Wikidata (Google-funded crowdsourced structured data)
  - Musicbrainz/DBtune (encyclopedia of music)
  - SIDER (http://sideeffects.embl.de)
  - Diseasome (http://wifo5-03.informatik.uni-mannheim.de/diseasome/)
  - Drugbank (http://www.drugbank.ca)
  - plenty of other sources
  - **Yes we do have structured data!**



http://lod-cloud.net/versions/2014-08-30/lod-cloud.svg

Intro to Semantic Web

Knowledge Graphs