Ollscoil na Gaillimhe
University of Galway

Introduction to Linked Data

Oksana Dereza
Digital Library Developer
oksana.dereza@universityofgalway.ie

University *of* Galway.ie

What is Linked Data?

Ollscoil na Gaillimhe
University of Galway

Linked Data is structured data published and interlinked using standard Web technologies to make the connections readable by computers, enabling data from different sources to be connected and queried.

Linked Open Data (LOD) must be publicly available under an open license.

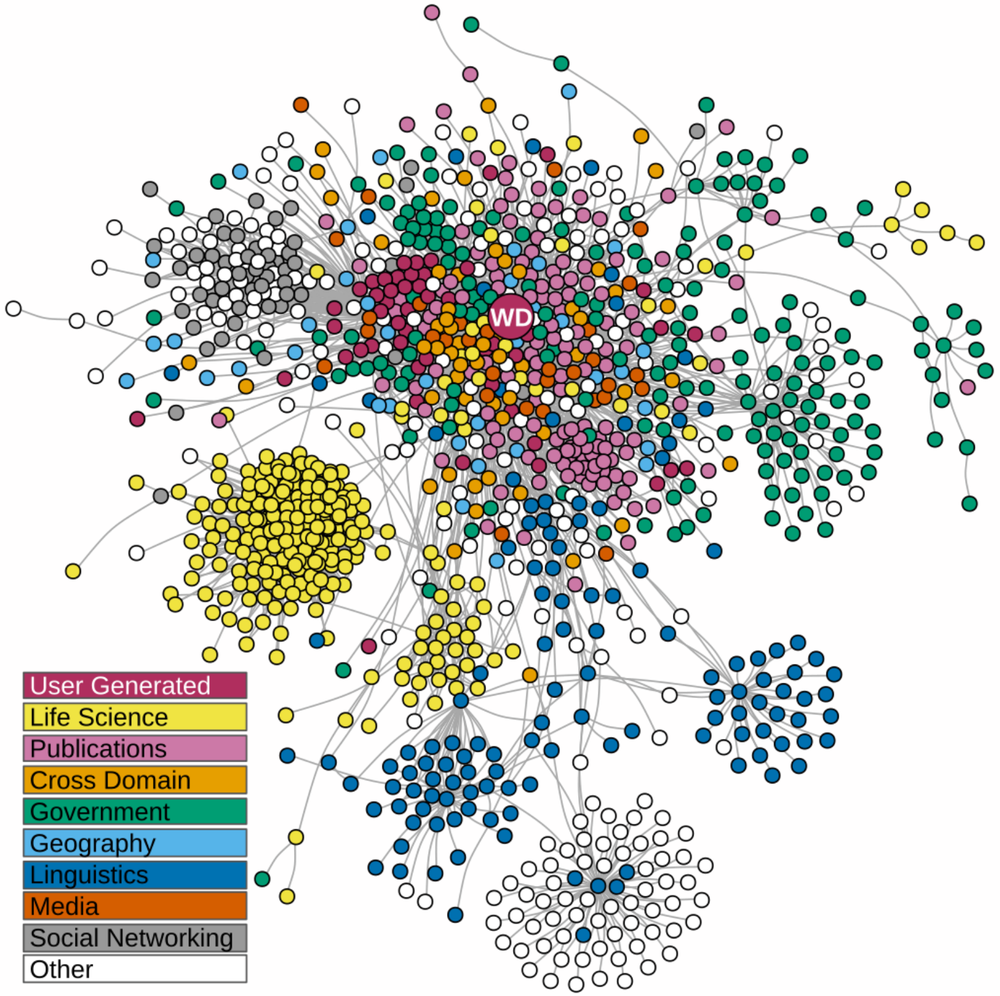


Image source: Wikipedia

How It Started

Ollscoil na Gaillimhe
University of Galway

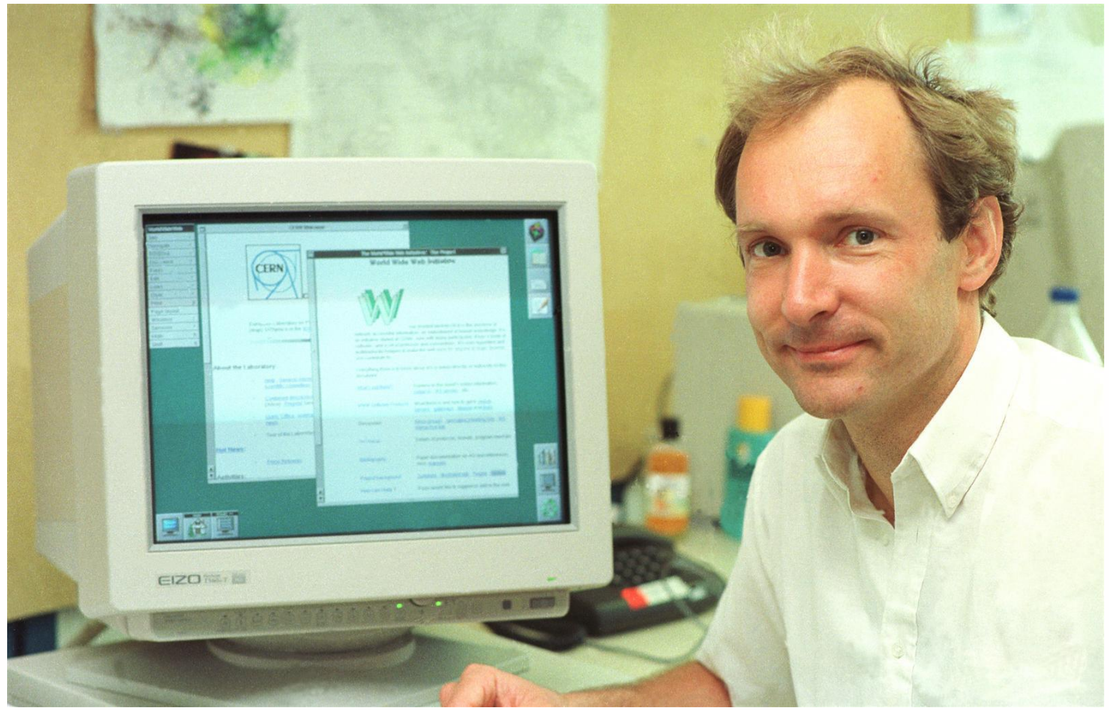
1989 — Tim Berners-Lee invents the World Wide Web

1993 — the World Wide Web released into the public domain, for general use, at no cost

1994 — the World Wide Web Consortium (W3C) is founded

1999 — Berners-Lee publishes his vision of the Semantic Web

2006 — the term Linked Data is coined in a design note about the Semantic Web project

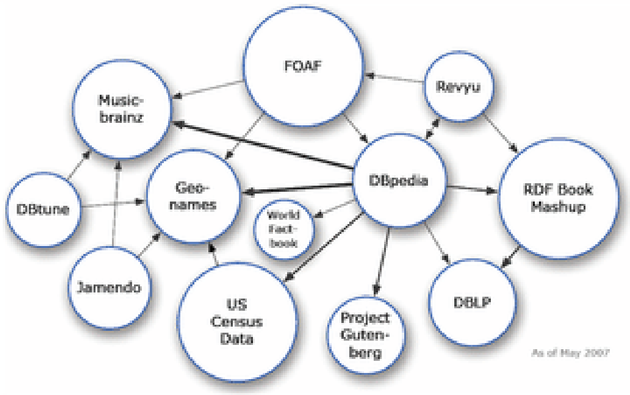


W3C Semantic Web

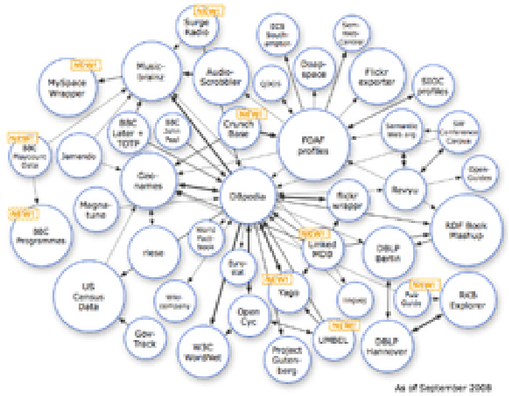
RDF

Image source: W3C

How It's Going

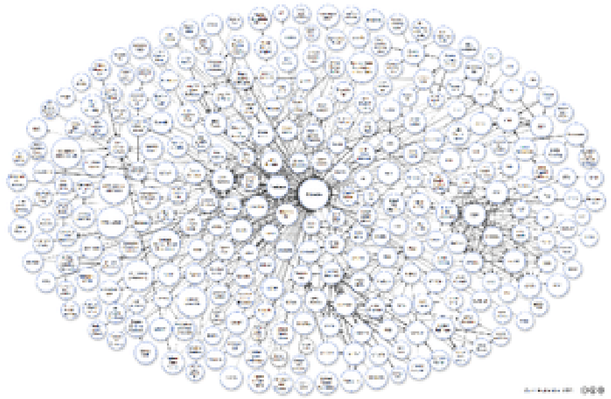


2007

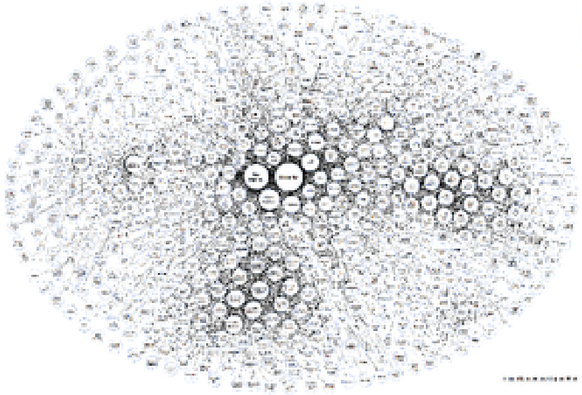


2008

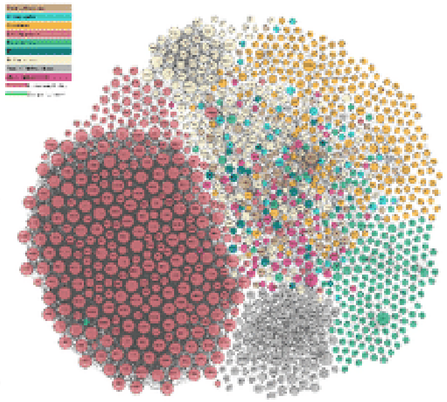
2009



2011



2014



2017

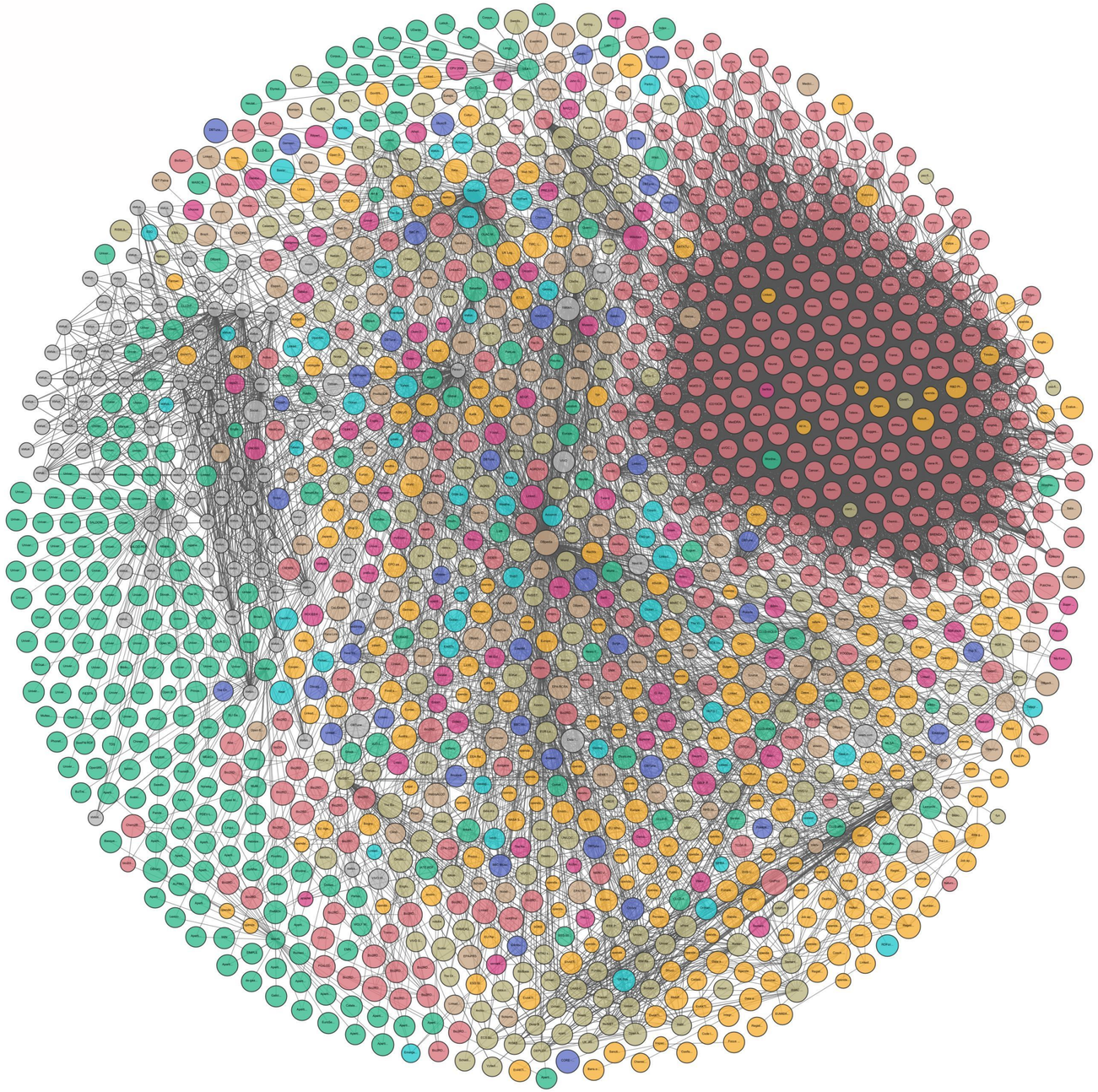
Image source: [ResearchGate]

How It’s Going

Linked Open Data cloud
https://lod-cloud.net/

As of March 26th, 2025, the LOD cloud contains **1,354 datasets** (nodes)

Provides dowloadable metadata for all listed datasets + original download links and endpoints



The Linked Open Data Cloud from lod-cloud.net

Datasets

- DBPedia — community effort to extract structured data from Wikipedia; contains 9.5 billion structured pieces of information (triples)
- WikiData — free knowledge database project hosted by Wikimedia and edited by volunteers; 12.5 billion triples
- GeoNames — user-editable geographical database, 93,896,732 triples
- Europeana — metadata on 2.4 million texts, images, videos and sounds, 117 million triples
- Government datasets
  - data.europa.eu
  - data.gov.uk
  - data.gov.ie
  - data.gov

LOD Design Principles

Ollscoil na Gaillimhe
University of Galway

1. **Uniform Resource Identifiers (URIs)** should be used as names for things
2. HTTP URIs should be used so that people can look up those names.
3. When someone looks up a URI, useful information should be provided through open standards such as **RDF**, **SPARQL**, etc.
4. Other things should be referred to with their URI-based names.

https://www.w3.org/DesignIssues/LinkedData.html

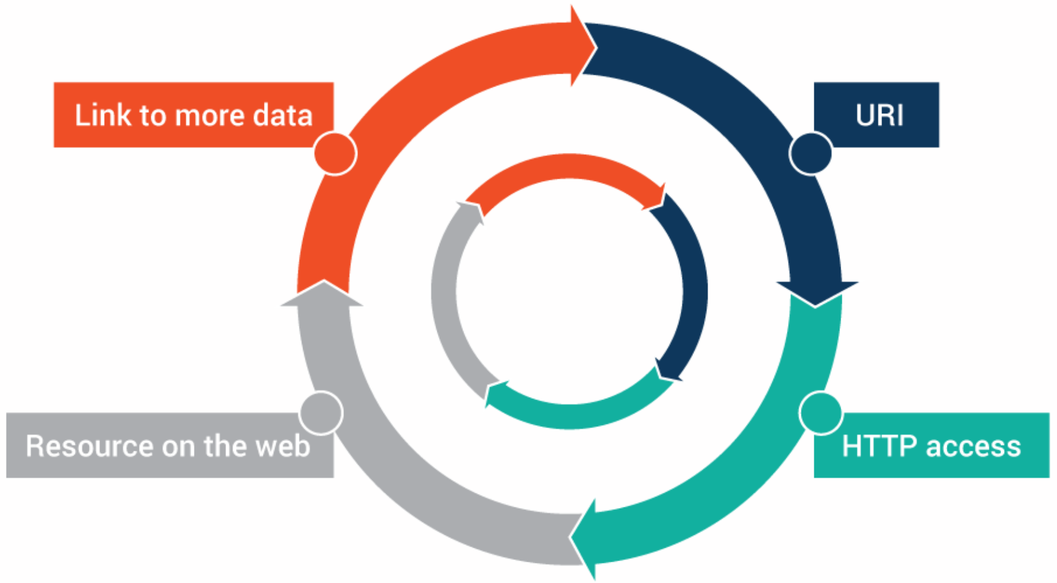


Image source: Ontotext

Uniform Resource Identifier (URI)

Uniform Resource Identifier (URI) — a compact sequence of characters that identifies an abstract or physical resource (RFC 3986 Uniform Resource Identifier (URI): Generic Syntax)

Uniform Resource Locator (URL) — a type of URI that refers to a resource on the web

Internationalised resource identifier (IRI) — a URI that allows Unicode characters

- ftp://ftp.is.co.za/rfc/rfc1808.txt
- http://www.ietf.org/rfc/rfc2396.txt
- http://publications.europa.eu/resource/authority/country/BEL
- mailto:John.Doe@example.com
- tel:+1-816-555-1212
- telnet://192.0.2.16:80
- file://localhost/etc/config

Ollscoil na Gaillimhe
University of Galway

HTTP URIs

scheme: specifies type of URI (and the associated protocol)

authority: hostname and optional port number

path: reference to content on the host server; can be interpreted in a hierarchical fashion (like a directory tree on a file system)

query: contains data that doesn't match tree structure (variables)

fragment: addresses local part of a document

IANAs registered schemes:
http://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml

RDF & OWL

Resource Description Framework (RDF) is a standard used to describe and exchange linked data.

In RDF, pieces of information are represented as **triples** that consist of a subject, an object and a predicate. A predicate is a directed link from subject to object.

RDF Schema (RDFS) and **Web Ontology Language (OWL)** are ontology languages used to describe RDF data.

An **ontology** is a formal linked data model that describes the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and properties can be used together (axioms).

A collection of triples is a **knowledge graph (KG).**

Linked Data Glossary: https://www.w3.org/TR/ld-glossary/

RDF triples

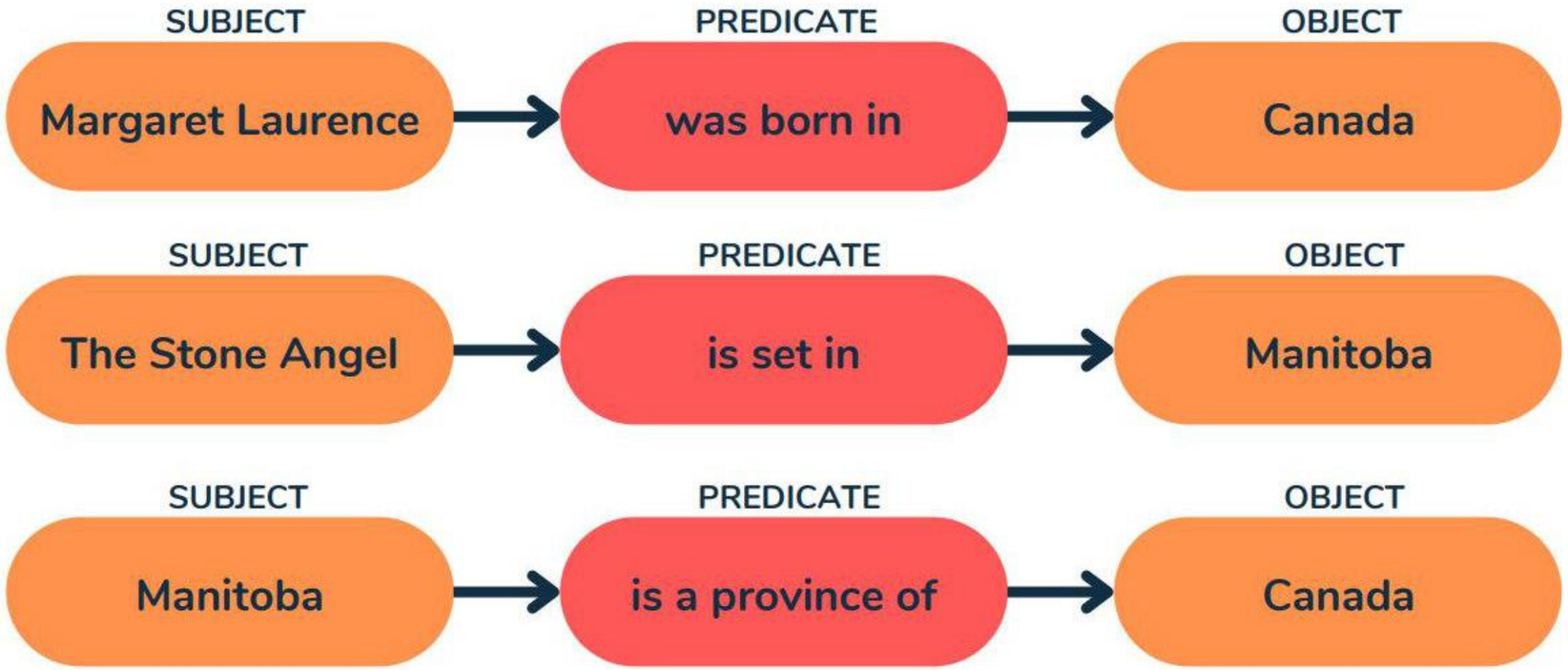
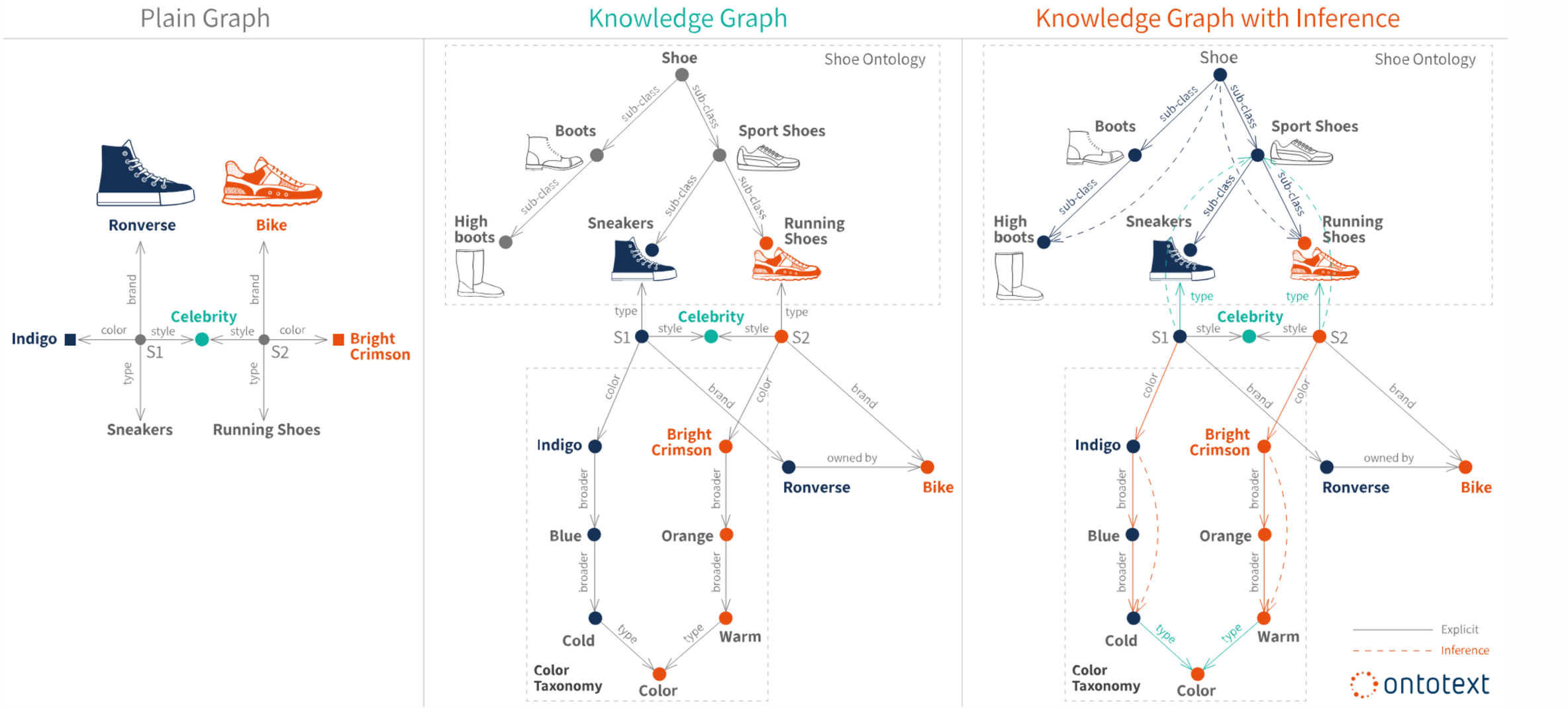


Image source: LINCS Portal

Knowledge Graphs



Linking Resources

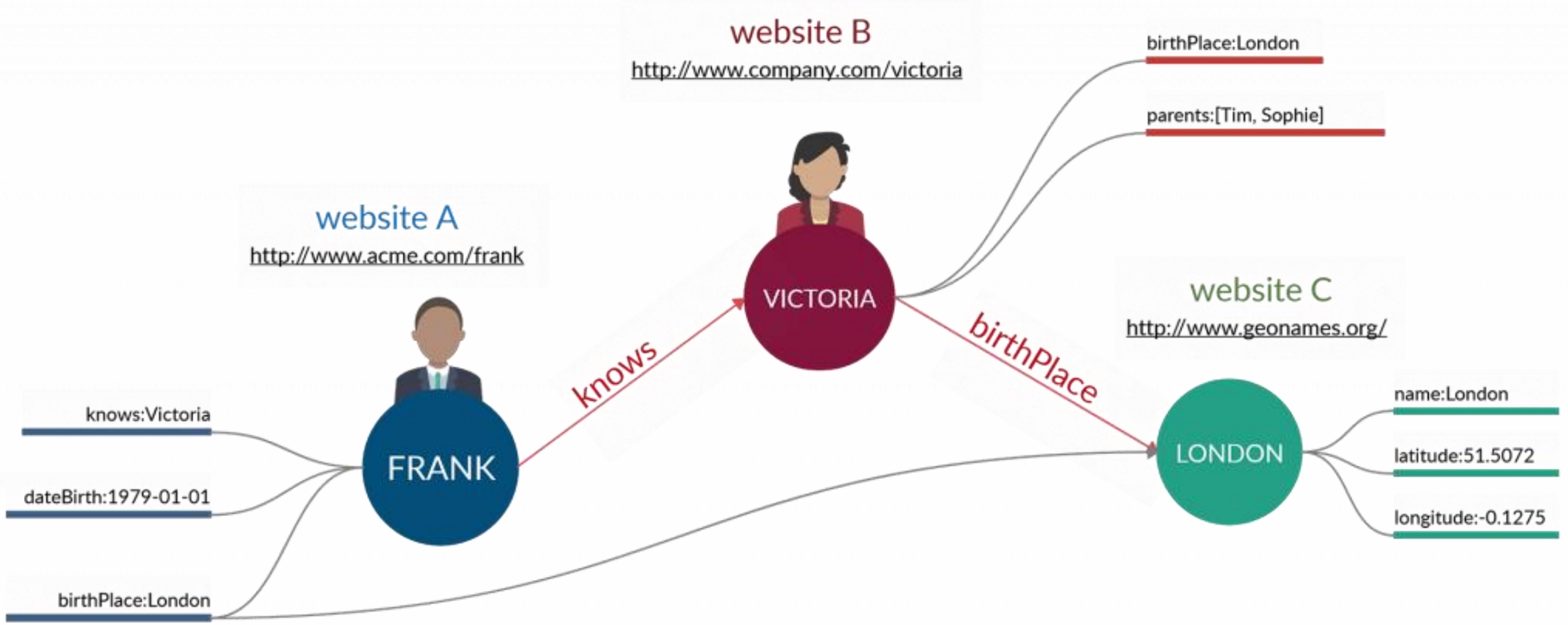


Image source: Fontistoriche

SPARQL

Ollscoil na Gaillimhe
University of Galway

SPARQL Protocol and RDF Query Language, or just **SPARQL** is a query language for RDF graphs.

SPARQL queries consist of triple patterns, conjunctions, disjunctions, and optional elements. It can be used to retrieve information from **triplestores** (a type of **graph database**) and to populate them.

Syntax description and examples by W3C: https://www.w3.org/TR/rdf-sparql-query/

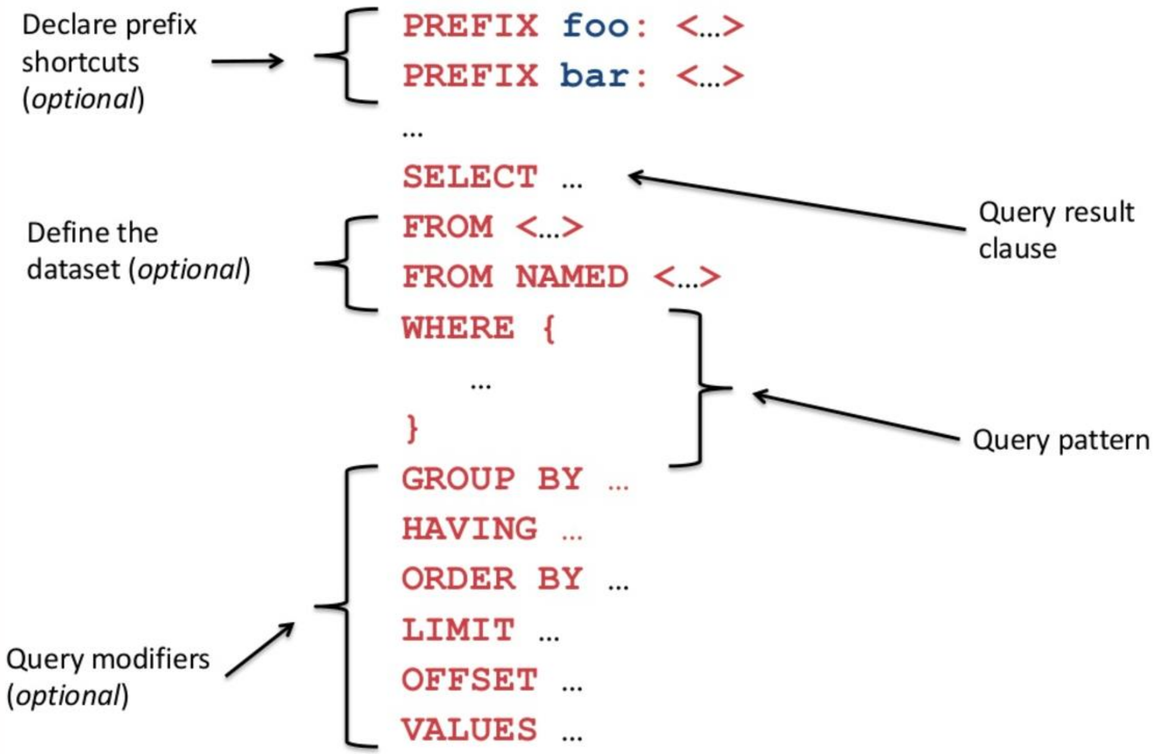
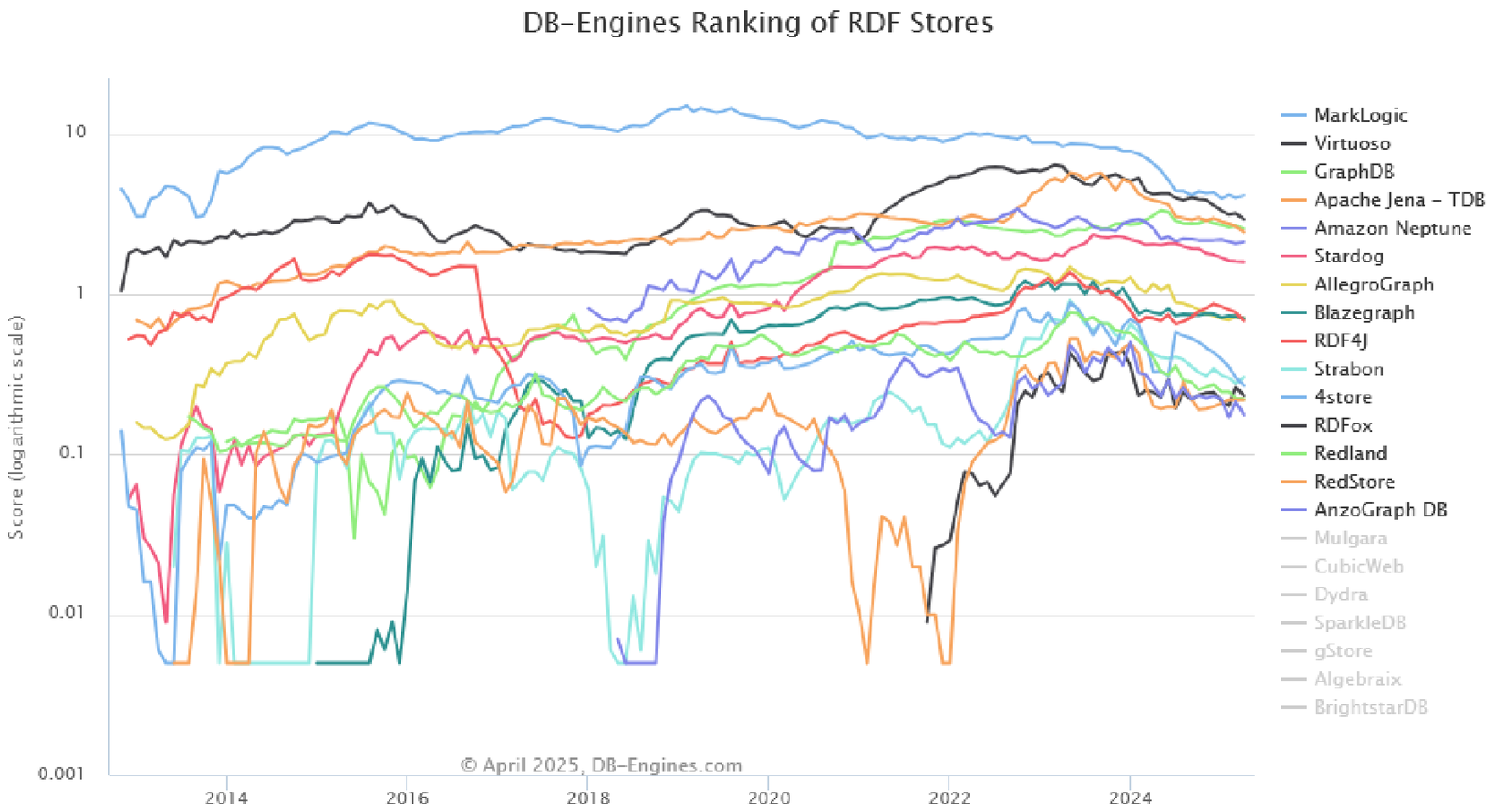


Image source: Wordlift

RDF stores

There are many software solutions for triple (RDF) stores and graph databases rated by DB-Engines



SPARQL Endpoints

Many LOD datasets and repositories have a **SPARQL endpoint**, for example https://query.wikidata.org/

Wikidata

Wikidata Query Service | Examples | Help | More tools | Query Builder

```
#Cats, with pictures
#title: Cats, with pictures
#defaultView:ImageGrid
SELECT ?item ?itemLabel ?pic WHERE {
  ?item wdt:P31 wd:Q146;
    wdt:P18 ?pic.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],mul,en". }
}
```

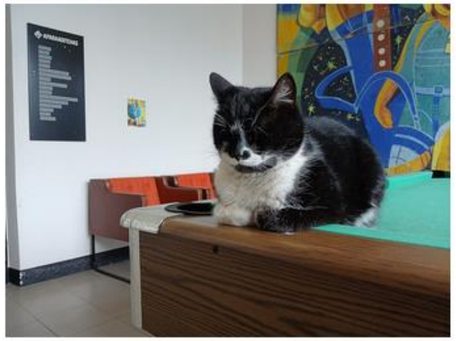
Code | Download

- JSON file
- JSON file (verbose)
- TSV file
- TSV file (verbose)
- CSV file
- HTML table

Image grid | 52 results in 484 ms | Code | Download | Link

Cats, with pictures

commons:Prabowo terima kunjungan Dubes China (18-2-2024).jpg
Category:Bobby Kertanegara



commons:Johannes Gutenberg (kass).jpg
Johannes Gutenberg

commons:Brünnhilde.tif
Brünnhilde

commons:Cat at Ayasofya.jpg
Gli

commons:Fat cat, asleep (319313958).jpg
Toffee

Wikidata

Wikidata Query Service | Examples | Help | More tools | Query Builder

```
#Cats, with pictures
#title: Cats, with pictures
#defaultView:ImageGrid
SELECT ?item ?itemLabel ?pic WHERE {
  ?item wdt:P31 wd:Q146;
    wdt:P18 ?pic.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],mul,
}
```

URL | HTML | Wikilink | PHP | JavaScript (jQuery) | JavaScript (modern) | Java | Perl | Python

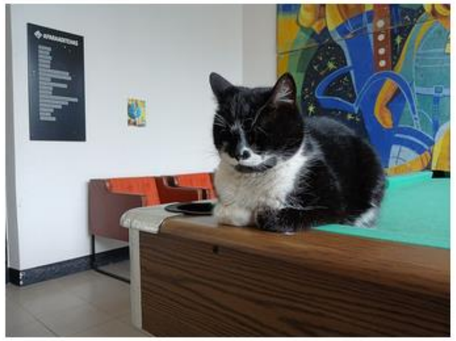
Python (Pywikibot) | Ruby | R | Matlab | listeria | mapframe

```
# pip install sparqlwrapper
# https://rdflib.github.io/sparqlwrapper/

import sys
from SPARQLWrapper import SPARQLWrapper, JSON

endpoint_url = "https://query.wikidata.org/sparql"

query = """#Cats, with pictures
#title: Cats, with pictures
#defaultView:ImageGrid
SELECT ?item ?itemLabel ?pic WHERE {
  ?item wdt:P31 wd:Q146;
    wdt:P18 ?pic.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],mul,en". }
}"""


def get_results(endpoint_url, query):
    user_agent = "WDQS-example Python/%s.%s" % (sys.version_info[0], sys.version_info[1])
    # TODO adjust user agent; see https://w.wiki/CX6
    sparql = SPARQLWrapper(endpoint_url, agent=user_agent)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()


results = get_results(endpoint_url, query)

for result in results["results"]["bindings"]:
    print(result)

```

Image grid | 52 results in 484 ms | Code | Download | Link

Cats, with pictures

commons:Prabowo terima kunjungan Dubes China (18-2-2024).jpg
Category:Bobby Kertanegara



commons:Johannes Gutenberg (kass).jpg
Johannes Gutenberg

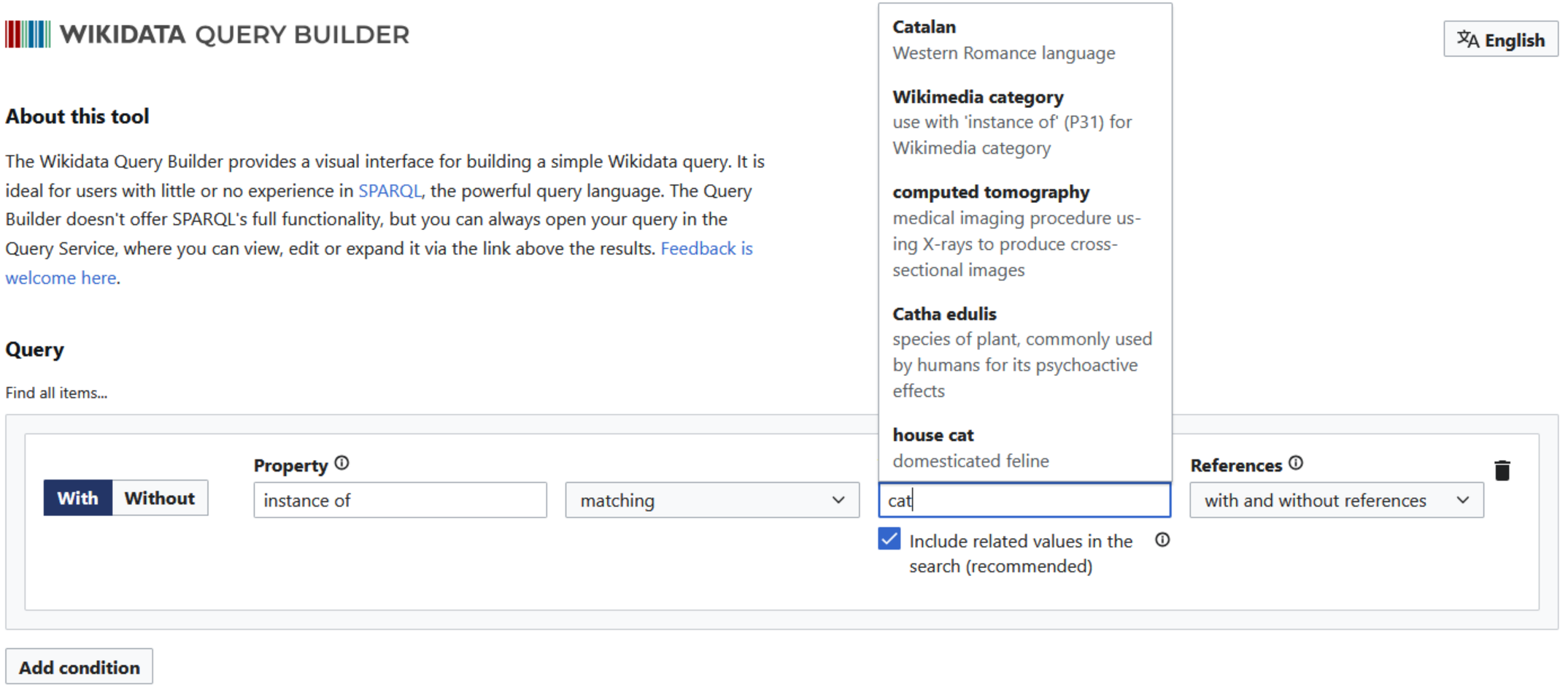
commons:Brünnhilde.tif
Brünnhilde

commons:Cat at Ayasofya.jpg
Gli

commons:Fat cat, asleep (319313958).jpg
Toffee

Query builder

Build queries to Wikidata without SPARQL: https://query.wikidata.org/querybuilder/



Linked Data Formats

Ollscoil na Gaillimhe
University of Galway

Linked data are usually serialised (recorded) in text-based (N-Triples, Turtle) or in XML-based formats (RDF/XML).

Turtle — **PREFIXES** — **RDF/XML**

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix ex: <http://example.org/stuff/1.0/> .

<http://www.w3.org/TR/rdf-syntax-grammar>
  dc:title "RDF/XML Syntax Specification (Revised)" ;
  ex:editor [
    ex:fullname "Dave Beckett";
    ex:homePage <http://purl.org/net/dajobe/>
  ] .
```

```
<rdf:RDF xmlns="http://xmlns.com/foaf/0.1/"
         xmlns:dc="http://purl.org/dc/terms/"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <Document rdf:about="http://www.w3.org/2001/sw/RDFCore/ntriples/">
    <dc:title xml:lang="en-US">N-Triples</dc:title>
    <maker>
      <Person rdf:nodeID="art">
        <name>Art Barstow</name>
      </Person>
    </maker>
    <maker>
      <Person rdf:nodeID="dave">
        <name>Dave Beckett</name>
      </Person>
    </maker>
  </Document>
</rdf:RDF>
```

Namespaces & Vocabularies

Ollscoil na Gaillimhe
University of Galway

RDF infrastructure

vocabulary	scope	prefix	namespace URI
RDF	Basic RDF elements	`rdf:`	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDF Schema	RDF Schema elements	`rdfs:`	http://www.w3.org/2000/01/rdf-schema#
Web Ontology Language (OWL)	OWL elements	`owl:`	http://www.w3.org/2002/07/owl#
SKOS	SKOS elements	`skos:`	http://www.w3.org/2004/02/skos/core#
SHACL	SHACL elements	`sh:`	http://www.w3.org/ns/shacl#

Simple Knowledge Organization System (SKOS) is a W3C recommendation for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary.

Shapes Constraint Language (SHACL) is a language for validating RDF graphs against a set of conditions

List of trusted namespaces by Australian national science agency CSIRO

Vocabulary	Scope	Prefix	Namespace URI
XML Schema: Datatypes	Datatypes used in XML schemas	**xsd:**	`http://www.w3.org/2001/XMLSchema#`
DBpedia	DBpedia classes & properties	**dbo:**	`http://dbpedia.org/ontology/`
Wikidata	Wikidata entities	**wd:**	`http://www.wikidata.org/entity/`
Wikidata	Wikidata properties	**wdt:**	`http://www.wikidata.org/prop/direct/`
Friend of a friend	Agents, persons, organisations (the original social-network ontology)	**foaf:**	`http://xmlns.com/foaf/0.1/`
Organization Ontology	Organisational structures and affiliations	**org:**	`http://www.w3.org/ns/org#`
Geonames	Geospatial semantic information (placenames & coordinates)	**gn:**	`http://www.geonames.org/ontology#`
OWL-Time	Temporal entities and relationships	**time:**	`http://www.w3.org/2006/time#`
Dublin Core	General purpose metadata	**dcterms:** **dct:**	`http://purl.org/dc/terms/`
Schema.org	General purpose metadata, used mostly for web-pages and services that they describe	**schema:** **sdo:**	`https://schema.org/`

LOD roadmap

Step 1: Select data

Step 2: Prepare the data

Step 3: Model the data

Step 4: Define a naming scheme

Step 5: Convert the data

Step 6: Organize Governance

Step 7: Add metadata

Step 8: Publish the data

Step 9: Link the data

Platform Linked Data Nederland

Check for existing vocabularies and ontologies:

- Linked Open Vocabularies (LOV)
- W3C list of common vocabularies & ontologies
- EU controlled vocabularies
- Controlled vocabularies list by CSIRO

Examples of domain-specific vocabularies & ontologies:

- Conservation Controlled Vocabularies
- Cultural Heritage Vocabularies
- Semantic Web for Earth and Environmental Terminology (SWEET) Music Ontology

LOD roadmap

Step 1: Select data

Step 2: Prepare the data

Step 3: Model the data

Step 4: Define a naming scheme

Step 5: Convert the data

Step 6: Organize Governance

Step 7: Add metadata

Step 8: Publish the data

Step 9: Link the data

Platform Linked Data Nederland

Convert

- **LODRefine**, an extension of OpenRefine
- **RDF Translator**
- **Spyder**
- **Python RDFLib**
- W3C Guidelines on **Generating RDF from Tabular Data**

Validate

- **Themis**
- **Shapes Constraint Language (SHACL)**
- **W3C RDF Validator**
- **XML validator** (one of many)

Five Star LOD Framework

★ Make available on the Web (whatever format) with an open license, to be Open Data

★★ Make available as machine-readable structured data (e.g. Excel instead of image scan of a table)

★★★ Use non-proprietary format (e.g. CSV instead of Excel)

★★★★ Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ Link your data to other people's data to provide context

Image source: W3C

Ollscoil na Gaillimhe
University of Galway

Hands-on session

University *of* Galway.ie

Protégé

A free open-source ontology editor, based on OWL

- Download: https://protege.stanford.edu/software.php
- Documentation: https://protegeproject.github.io/protege/
- Protégé wiki: https://protegewiki.stanford.edu/wiki/Main_Page
- Pizza Ontology Tutorial: https://protegewiki.stanford.edu/wiki/Protege4Pizzas10Minutes

Image source: University of Edinburgh

Pizza Ontology

Ollscoil na Gaillimhe
University of Galway

http://protege.stanford.edu/ontologies/pizza/pizza.owl

Classes, Properties & Individuals

Classes describe concepts in the domain and include individuals that share common characteristics

- Class description: https://protegeproject.github.io/protege/views/class-description/

Individuals (instances) are specific entities, representatives of a class

Object properties represent relationships between two individuals

- Object property description: https://protegeproject.github.io/protege/views/object-property-description/
- Object property characteristics: https://protegeproject.github.io/protege/views/object-property-characteristics/

Data properties connect individuals to literals (a value like a string, number, date, boolean)

Visualisation: OntoGraf

Ollscoil na Gaillimhe
University of Galway

Window > Tabs > OntoGraf

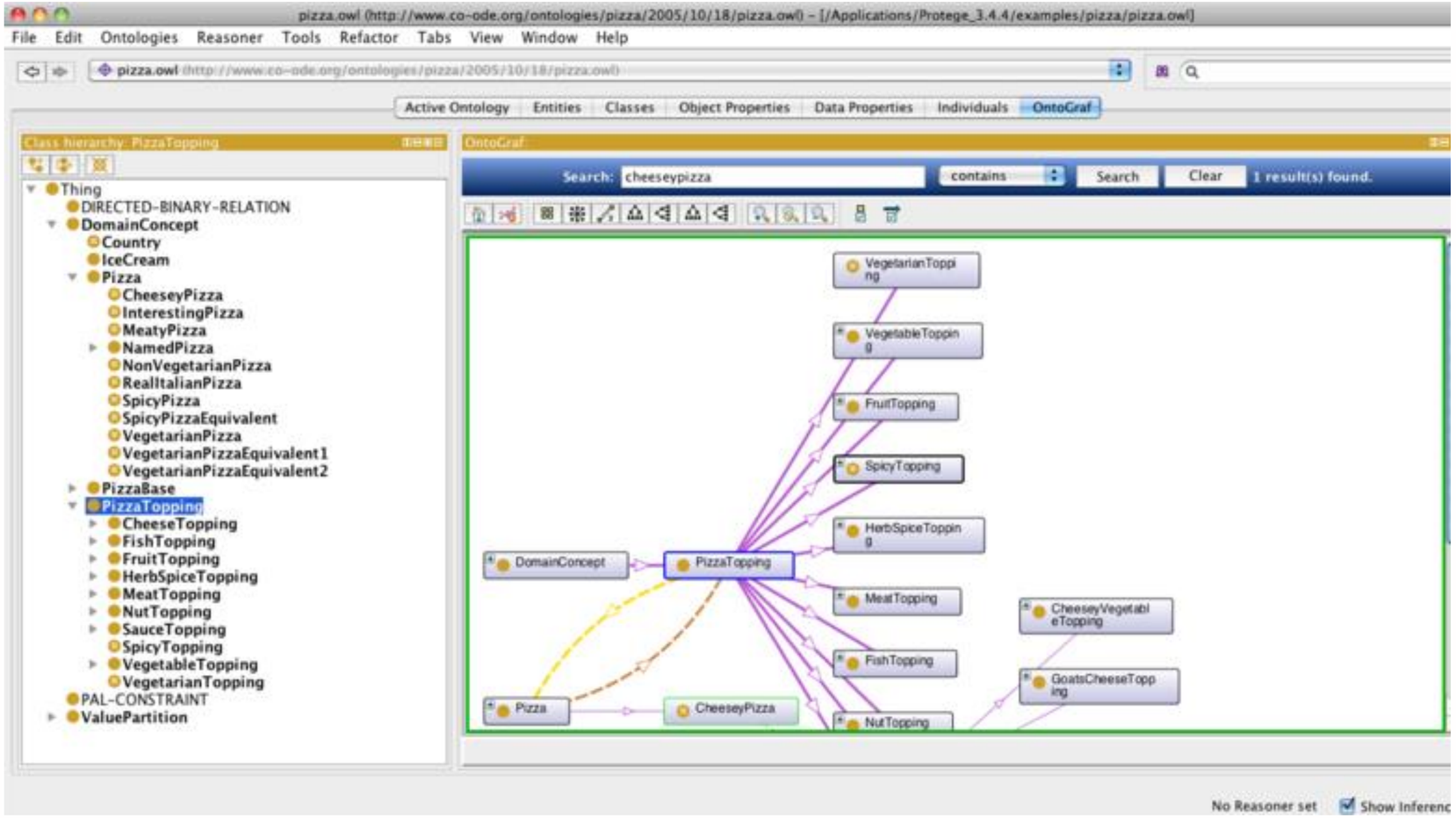


Image source: Protégé Wiki

Further Reading

1. W3C Wiki: Linked Data
2. Jonathan Blaney (2017). Introduction to the Principles of Linked Open Data. Programming Historian 6, https://doi.org/10.46430/phen0068
3. Matthew Lincoln (2015). Using SPARQL to access Linked Open Data, Programming Historian 4, https://doi.org/10.46430/phen0047
4. Michael DeBellis (2021). A Practical Guide to Building OWL Ontologies Using Protégé 5.5 and Plugins. Edition 3.2. ResearchGate.
5. Natalya F. Noy and Deborah L. McGuinness (2001). Ontology development 101: A guide to creating your first ontology. Stanford.
6. Vindula Jayawardana (2017). Ontology Generation and Visualization with Protégé. Medium.
7. Fundamentals of data and knowledge management by Ontotext (illustrated explanations of concepts and technologies)
    - Relational vs Property Graphs vs RDF Databases at a Glance
    - What Is a Knowledge Graph?
    - What is SPARQL?

Further Reading

8. Angus Addlesee on Medium
   - Understanding Linked Data Formats (2018)
   - Creating Linked Data (2018) — *with OpenRefine*
   - Using OntoRefine to Transform Tabular Data into Linked Data (2019)
   - Constructing SPARQL Queries (2019)
   - Constructing More Advanced SPARQL Queries (2019)
   - Comparing Linked Data Triplestores (2018)
   - Where to Find Linked Open Data for Your Home Projects (2020)
9. Sanaz Saki Norouzi, Adrita Barua, Antrea Christou, Nikita Gautam, Andrew Eells, Pascal Hitzler, Cogan Shimizu (2024). Ontology Population using LLMs. arXiv:2411.01612
10. Emma Griffiths (2020). Annotating Data Using Ontologies [Video]. Youtube.
11. Elissa Gilbert (2016). Triplestores 101: Storing Data for Efficient Inferencing. Dataversity.
12. LargeTripleStores (2015). W3C Wiki.
13. Gavin Mendel-Gleason (2015). Ontology Consistency and Instance Checking for Real World Linked Data [Video]. Videolectures.net