COMP3220:
Document Processing and Semantic Technologies
DBpedia and Wikidata

Rolf Schwitter

Rolf.Schwitter@mq.edu.au

Today's Agenda

- Linked Data
- DBpedia
- Querying DBpedia
- Wikidata
- Querying Wikidata

Linked Data

- Linked Data describes a method for publishing structured data.
- Based on four principles (Tim Berners-Lee):
  1. Use URIs as names for things.
  2. Use HTTP URIs so that people can look up those things.
  3. Provide useful information about the thing when its URI is looked up, using standards such as RDF and SPARQL.
  4. Add links to other related things (using their URIs) when publishing data on the Web.

See: http://www.w3.org/DesignIssues/LinkedData.html

What is DBpedia?

- DBpedia is an effort to
  - extract structured information from Wikipedia
  - make this information available on the Web.
- DBpedia allows users to
  - query its relationships and properties, including links to other datasets.
  - find answers to questions where the information is spread across many different Wikipedia articles.

Wikipedia

- Wikipedia articles consist mainly of free text.
- But they include also structured information:
  - infoboxes
  - categorisation information
  - geo-coordinates
  - links to external web pages.
- Currently only free-text search capacities.
- The structured information can be extracted and represented as RDF which can then be queried.

Structured Information in Wikipedia

https://en.wikipedia.org/wiki/Nicole_Kidman

Infobox

Nicole Kidman
AC

Kidman at the 2017 Cannes Film Festival

Born	Nicole Mary Kidman 20 June 1967 (age 54) Honolulu, Hawaii, U.S.
Citizenship	Australia · United States
Occupation	Actress · producer
Years active	1983–present
Works	Full list
Spouse(s)	Tom Cruise (m. 1990; div. 2001) Keith Urban (m. 2006)
Children	4
Parent(s)	Antony Kidman (father)
Relatives	Antonia Kidman (sister)
Awards	Full list
Website	nicolekidmanofficial.com

DBpedia Knowledge Base

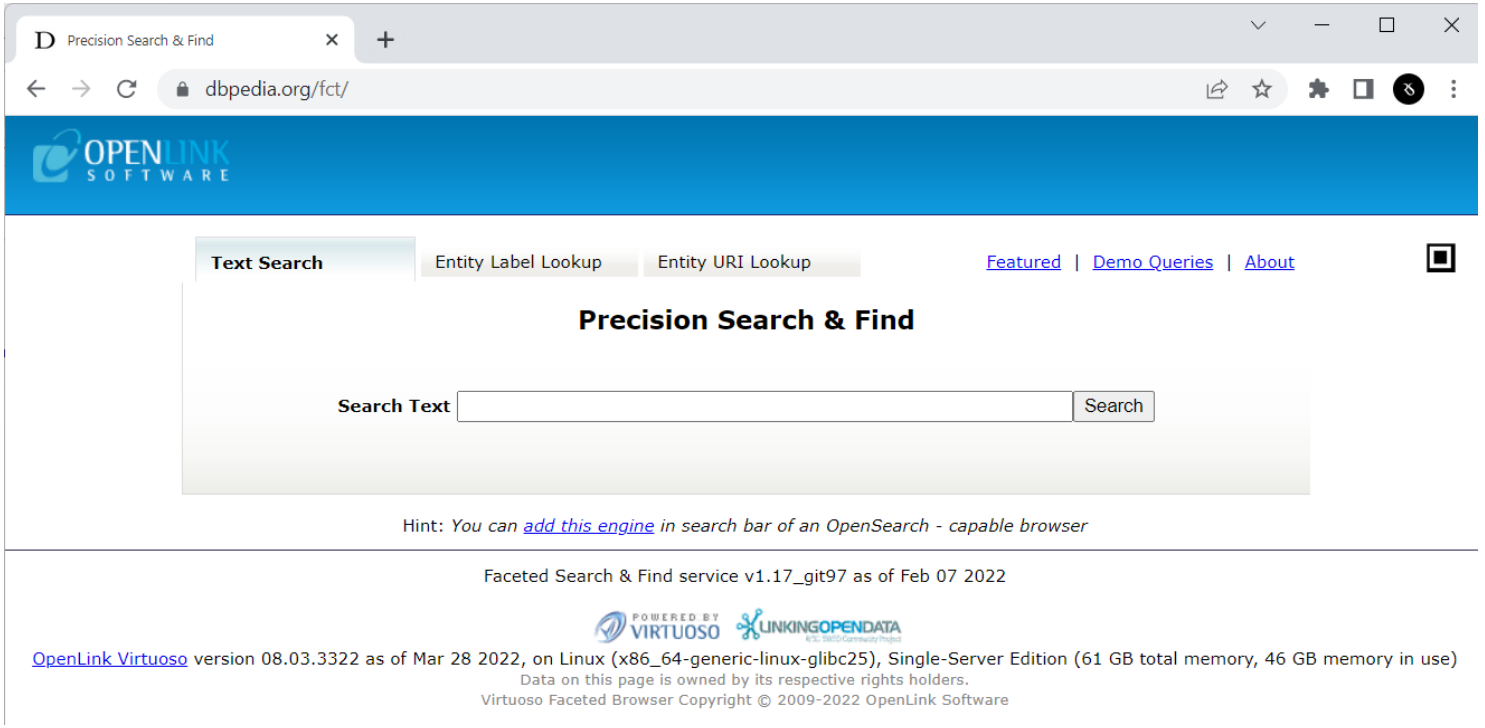
- The current DBpedia release contains
  - more than 850 millions facts (triples)
  - utilises a total of 55,000 properties
  - 1372 of these properties are defined the DBpedia ontology.
- The 768 classes of the ontology include:
  - Persons: 1,682,299
  - Places: 992,381
  - Works: 593,689 (e.g., music albums, films, video games)
  - Organisations: 341,609 (e.g., companies, educational institutions)
  - Species, Plants, Diseases, etc.

https://www.dbpedia.org/blog/snapshot-2021-06-release/

DBpedia Knowledge Base

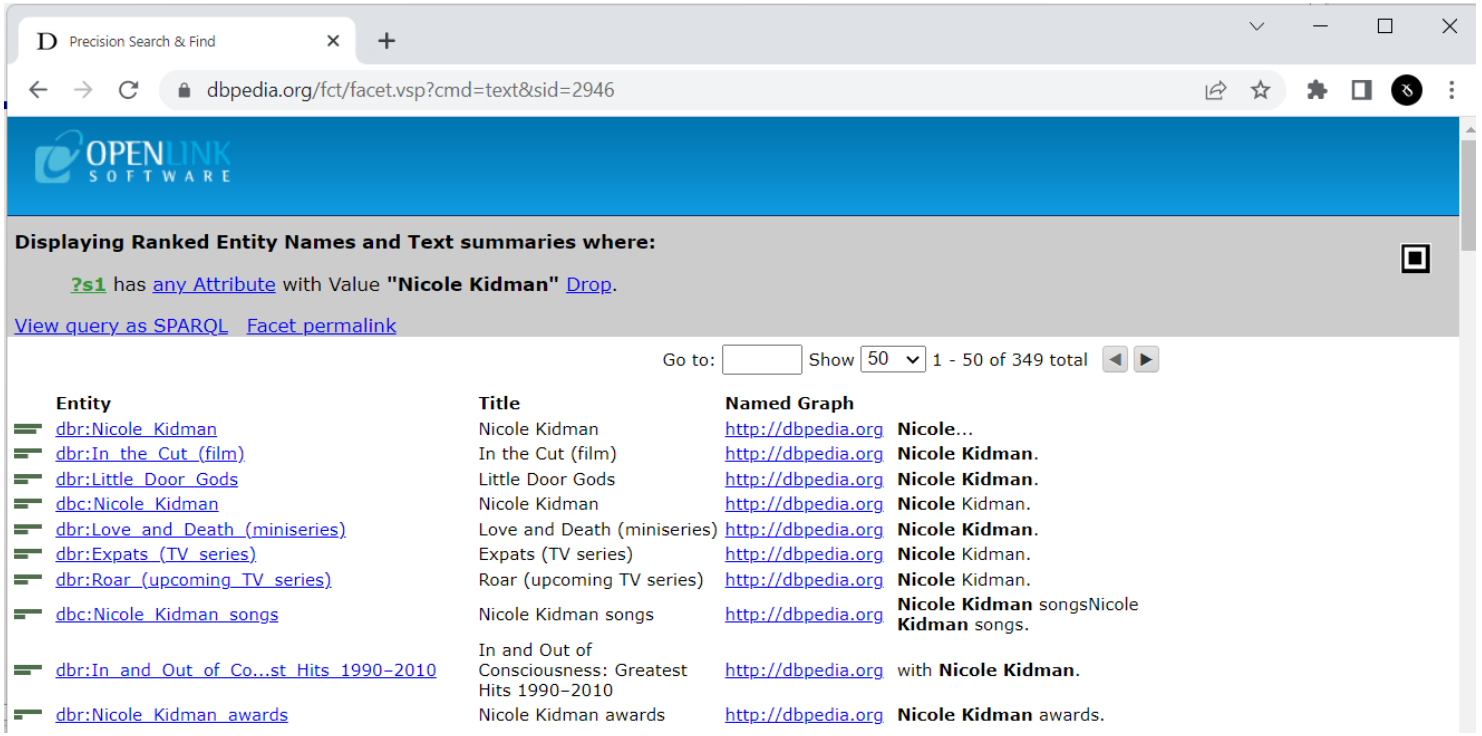
- DBpedia advantages:
  - it covers many domains
  - it represents real community agreement
  - it automatically evolves as Wikipedia changes
  - it is multilingual (localised version in 125 languages).

Browsing DBpedia

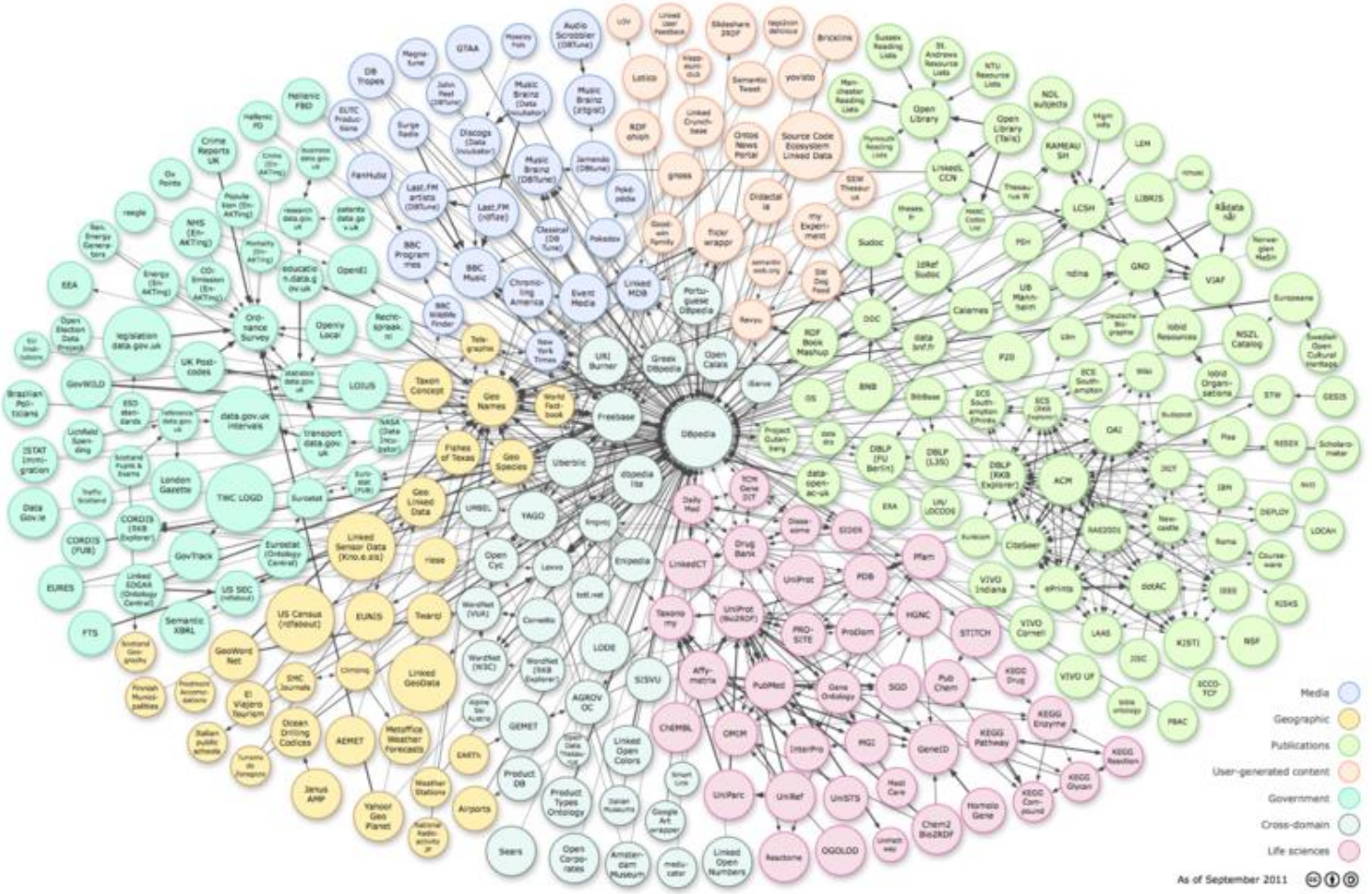


http://dbpedia.org/fct/

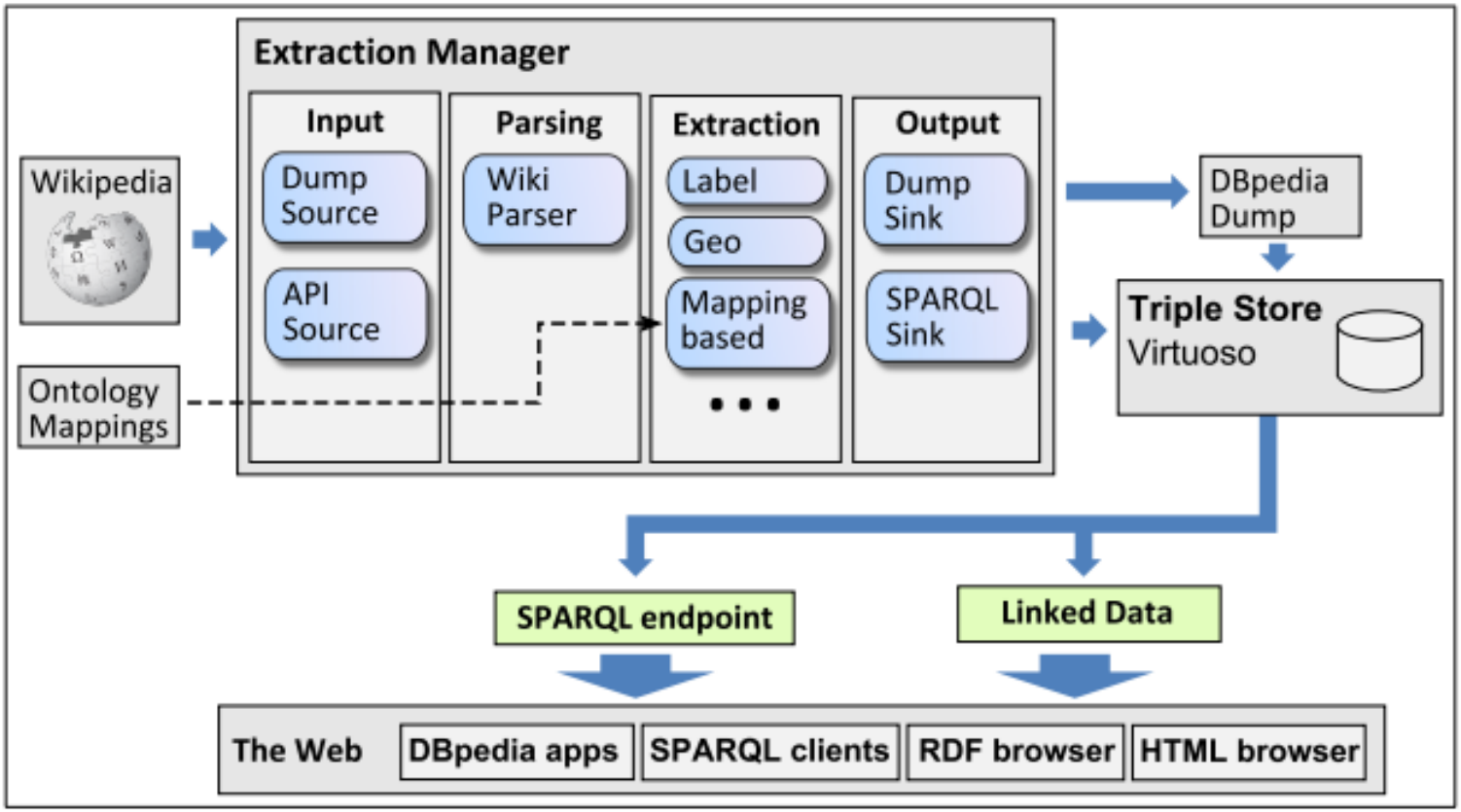
Browsing DBpedia



DBpedia as Nucleus for the Web of Data



DBpedia Extraction Framework



General Architecture

- Input
  - Wiki pages are read from an external source.
- Parsing
  - Each Wiki page is parsed by the Wiki parser and transformed into an Abstract Syntax Tree.
- Extraction
  - Abstract Syntax Tree is forwarded to the extractors.
  - Each extractor consumes an Abstract Syntax Tree and yields a set of RDF statements.
- Output
  - The collected RDF statements are written to a sink.

Extractors

- Extractors can be divided into four categories:
  1. Mapping-based infobox extraction

     uses manually written mappings that relate Wiki infoboxes to terms in the DBpedia ontology.
  2. Raw infobox extraction

     direct mappings from infoboxes to RDF (lower quality).
  3. Feature extraction

     extractors for features such as labels or geographic coordinates.
  4. Statistical extraction

     extractors for topic signatures, grammatical gender, lexicalisations, thematic concepts.

Overview of DBpedia Extractors

Name	*Description*	*Example*
abstract	Extracts the first lines of the Wikipedia article.	`dbr:Berlin dbo:abstract "Berlin is the capital city of (...)".`
article categories	Extracts the categorization of the article.	`dbr:Oliver_Twist dc:subject dbr:Category:English_novels.`
category label	Extracts labels for categories.	`dbr:Category:English_novels rdfs:label "English novels".`
category hierarchy	Extracts information about which concept is a category and how categories are related using the SKOS Vocabulary.	`dbr:Category:World_War_II skos:broader dbr:Category:Modern_history.`
disambiguation	Extracts disambiguation links.	`dbr:Alien dbo:wikiPageDisambiguates dbr:Alien_(film).`
external links	Extracts links to external web pages related to the concept.	`dbr:Animal_Farm dbo:wikiPageExternalLink <http://books.google.com/?id=RBGmrDnBs8UC>.`
geo coordinates	Extracts geo-coordinates.	`dbr:Berlin georss:point "52.5006 13.3989".`
grammatical gender	Extracts grammatical genders for persons.	`dbr:Abraham_Lincoln foaf:gender "male".`
homepage	Extracts links to the official homepage of an instance.	`dbr:Alabama foaf:homepage <http://alabama.gov/>.`
image	Extracts the first image of a Wikipedia page.	`dbr:Berlin foaf:depiction <http://.../Overview_Berlin.jpg>.`
infobox	Extracts all properties from all infoboxes.	`dbr:Animal_Farm dbo:date "March 2010".`

Overview of DBpedia Extractors

interlanguage	Extracts interwiki links.	`dbr:Albedo dbo:wikiPageInterLanguageLink dbr-de:Albedo .`
label	Extracts the article title as label.	`dbr:Berlin rdfs:label "Berlin" .`

lexicalizations	Extracts information about surface forms and their association with concepts (only N-Quad format).	`dbr:Pine sptl:lexicalization lx:pine_tree ls:Pine_pine_tree .`
		`lx:pine_tree rdfs:label "pine tree" .`
		`ls:Pine_pine_tree sptl:pUriGivenSf "0.941" .`

mappings	Extraction based on mappings of Wikipedia infoboxes to the DBpedia ontology.	`dbr:Berlin dbo:country dbr:Germany .`
page ID	Extracts page ids of articles.	`dbr:Autism dbo:wikiPageID "25" .`
page links	Extracts all links between Wikipedia articles.	`dbr:Autism dbo:wikiPageWikiLink dbr:Human_brain .`
persondata	Extracts information about persons represented using the PersonData template.	`dbr:Andre_Agassi foaf:birthDate "1970-04-29" .`
PND	Extracts PND (Personennamendatei) data about a person.	`dbr:William_Shakespeare dbo:individualisedPnd "118613723" .`
redirects	Extracts redirect links between articles in Wikipedia.	`dbr:ArtificialLanguages dbo:wikiPageRedirects` `dbr:Constructed_language .`
revision ID	Extracts the revision ID of the Wikipedia article.	`dbr:Autism <http://www.w3.org/ns/prov#wasDerivedFrom>` `<http://en.wikipedia.org/wiki/Autism?oldid=495234324> .`
thematic concept	Extracts 'thematic' concepts, the centres of discussion for categories.	`dbr:Category:Music skos:subject dbr:Music .`
topic signatures	Extracts topic signatures.	`dbr:Alkane sptl:topicSignature "carbon alkanes atoms" .`
wiki page	Extracts links to corresponding articles in Wikipedia.	`dbr:AnAmericanInParis foaf:isPrimaryTopicOf` `<http://en.wikipedia.org/wiki/AnAmericanInParis> .`

Mapping-Based Infobox Extraction

- Mappings are used to normalise the extracted information before the RDF representation is generated.
- Mapping example: `infobox_actor`

```
{{TemplateMapping
| mapToClass = Actor
| mappings =
   {{ PropertyMapping | templateProperty = name | ontologyProperty = foaf:name }}
   {{ PropertyMapping | templateProperty = birth_place | ontologyProperty = birthPlace }}
}}
```

Mapping-Based Infobox Extraction

- This particular mapping extracts the following information:
  1. type information (Actor)
  2. name of the actor
  3. actor's place of birth.
- For each infobox_actor three RDF triples are extracted:

```
dbpedia:Vince_Vaughn  rdf:type  dbpedia-owl:Actor .
dbpedia:Vince_Vaughn  foaf:name "Vince Vaughn"@en .
dbpedia:Vince_Vaughn  dbbpedia-owl:birthPlace  dbpedia:Minneapolis .
```

DBpedia Ontology

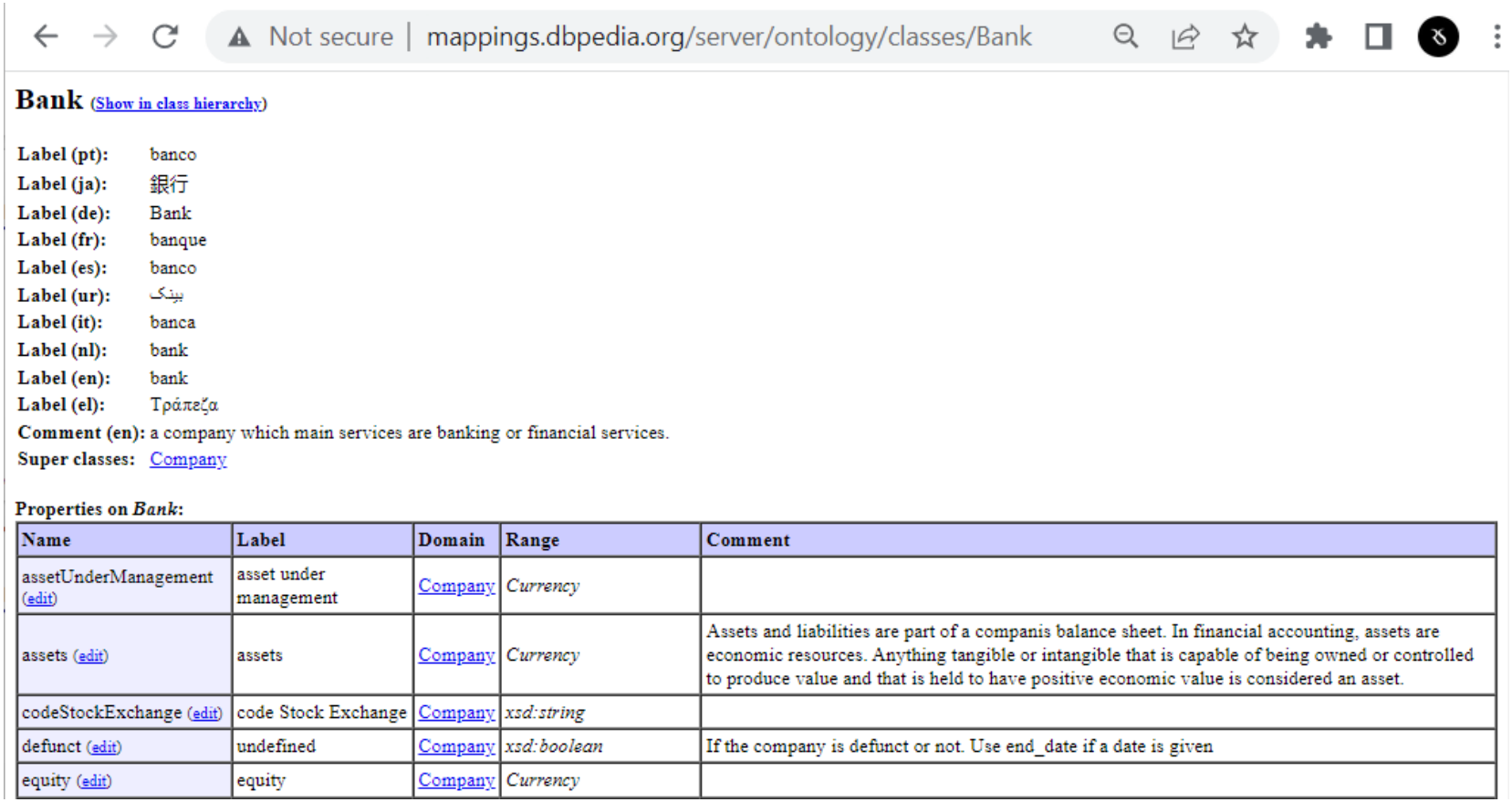
- DBpedia uses a shallow (manually created) ontology.
- The ontology covers 768 classes.
- Classes may have multiple superclasses (ontology is a directed acyclic graph).
- A taxonomy can still be constructed by ignoring all superclasses except the most important one.
- These classes are described by 3000 different properties.
- The DBpedia ontology contains about 4,828,000 instances.

https://www.dbpedia.org/resources/ontology/

DBpedia Ontology Classes

http://mappings.dbpedia.org/server/ontology/classes/

DBpedia Ontology: Bank



Bank (Show in class hierarchy)

Label (pt): banco
Label (ja): 銀行
Label (de): Bank
Label (fr): banque
Label (es): banco
Label (ur): بینک
Label (it): banca
Label (nl): bank
Label (en): bank
Label (el): Τράπεζα
Comment (en): a company which main services are banking or financial services.
Super classes: Company

Properties on *Bank*:

Name	Label	Domain	Range	Comment
assetUnderManagement (edit)	asset under management	Company	*Currency*	
assets (edit)	assets	Company	*Currency*	Assets and liabilities are part of a companis balance sheet. In financial accounting, assets are economic resources. Anything tangible or intangible that is capable of being owned or controlled to produce value and that is held to have positive economic value is considered an asset.
codeStockExchange (edit)	code Stock Exchange	Company	*xsd:string*	
defunct (edit)	undefined	Company	*xsd:boolean*	If the company is defunct or not. Use end_date if a date is given
equity (edit)	equity	Company	*Currency*	

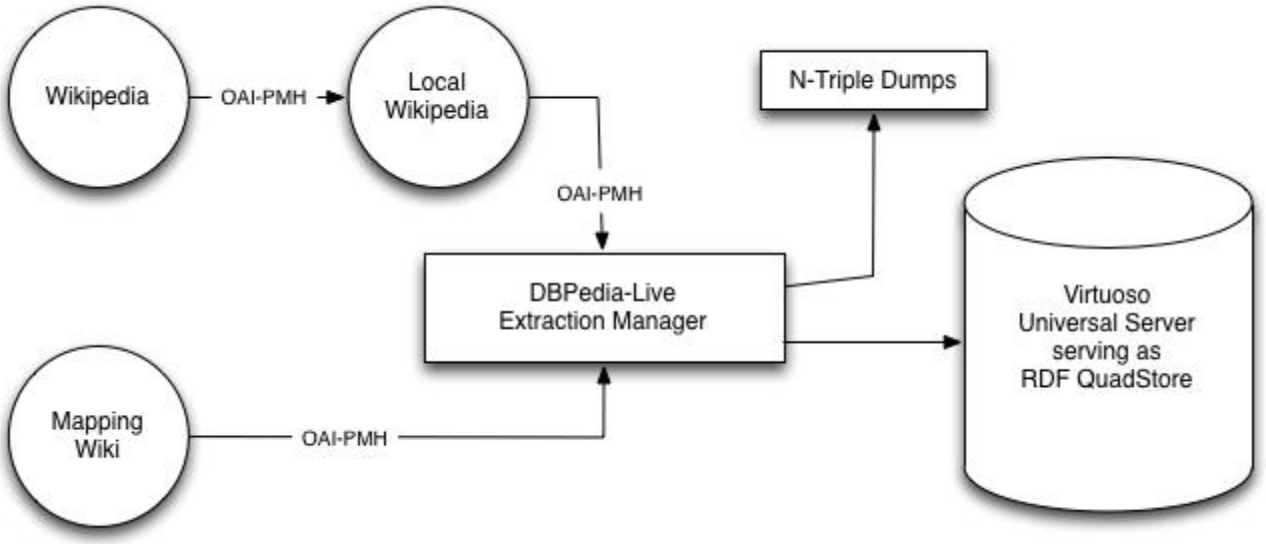
DBpedia Live Synchronisation

- Wikipedia articles are continuously revised.
- The DBpedia Live Sync API extracts the most recent data from around 300,000 updated Wikipedia articles each day.
- The aim of DBpedia Live is to keep DBpedia synchronised with Wikipedia.
- DBpedia Live allows data to stay up-to-date with a small delay of at most a few minutes.
- A special protocol is used to facilitate this process.

DBpedia Live System Architecture

- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is used to:
  - update a local version of Wikipedia
  - feed updates into the DBpedia-Live Extraction Manager
  - get a stream of updates from Mappings Wiki.
- Extracted triples are written into the triple store (Virtuoso).
- Triples are also written as N-Triples files and compressed.
- Other applications (mirrors) can use these N-Triples files.

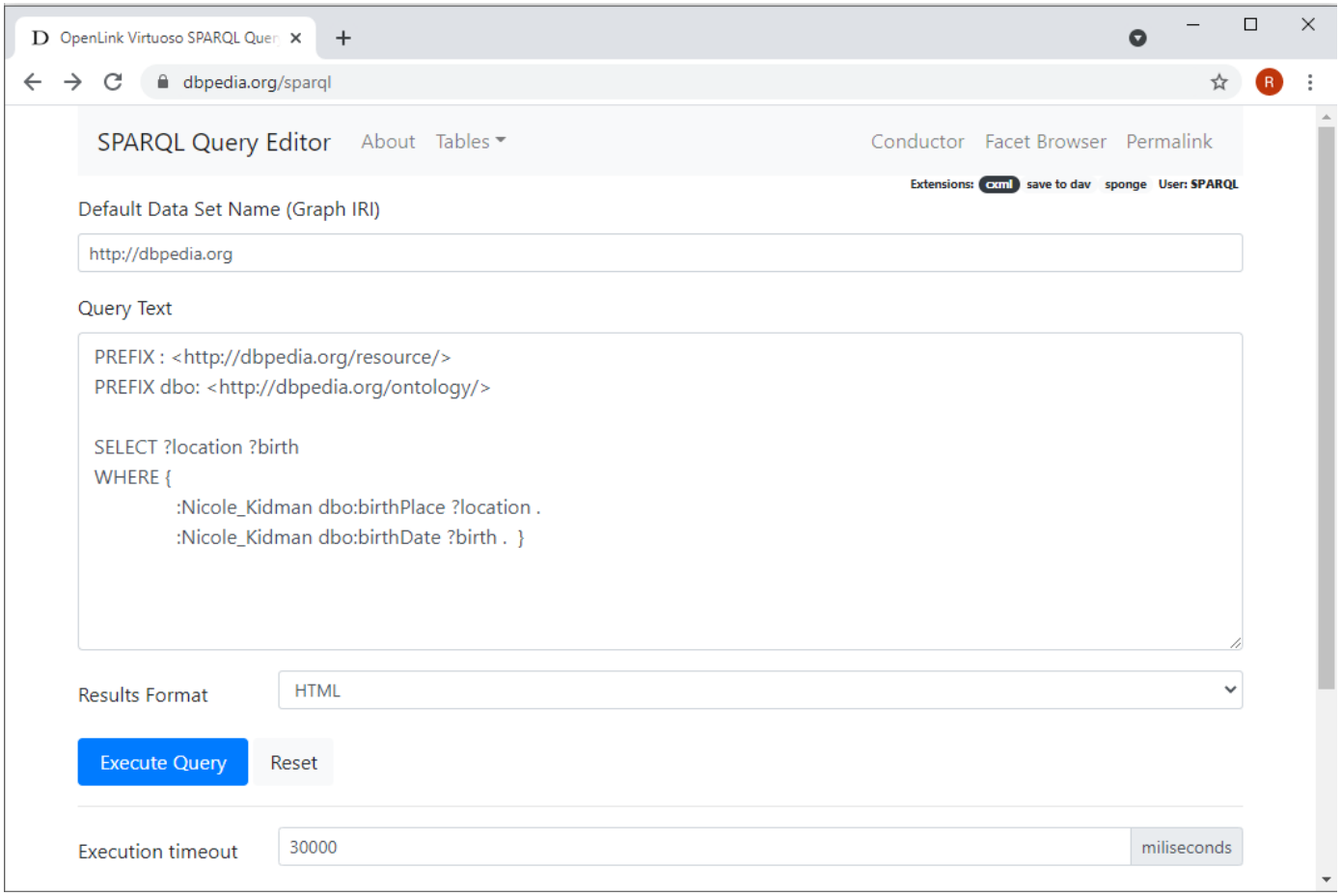
DBpedia Live System Architecture



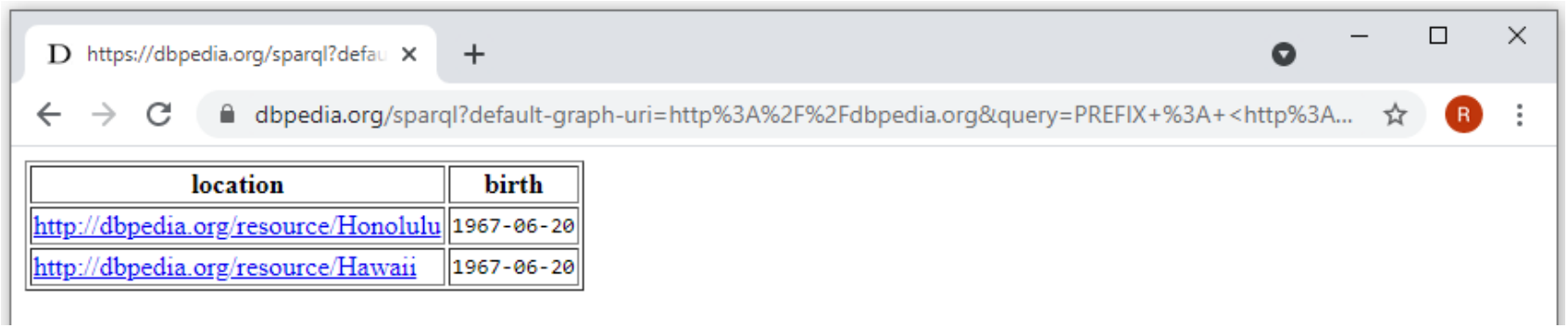
Querying DBpedia

- DBpedia can be queried via a Web interface:
  - http://dbpedia.org/sparql
- The endpoint is provided using OpenLink Virtuoso:
  - http://virtuoso.openlinksw.com/
- OpenLink Virtuoso is an ORDBMS that facilitates RDF-triple storage and the use of SPARQL.
- The cluster setup provides parallelisation of query execution.
- A maximum number of requests as well as a bandwidth limit per request are enforced.

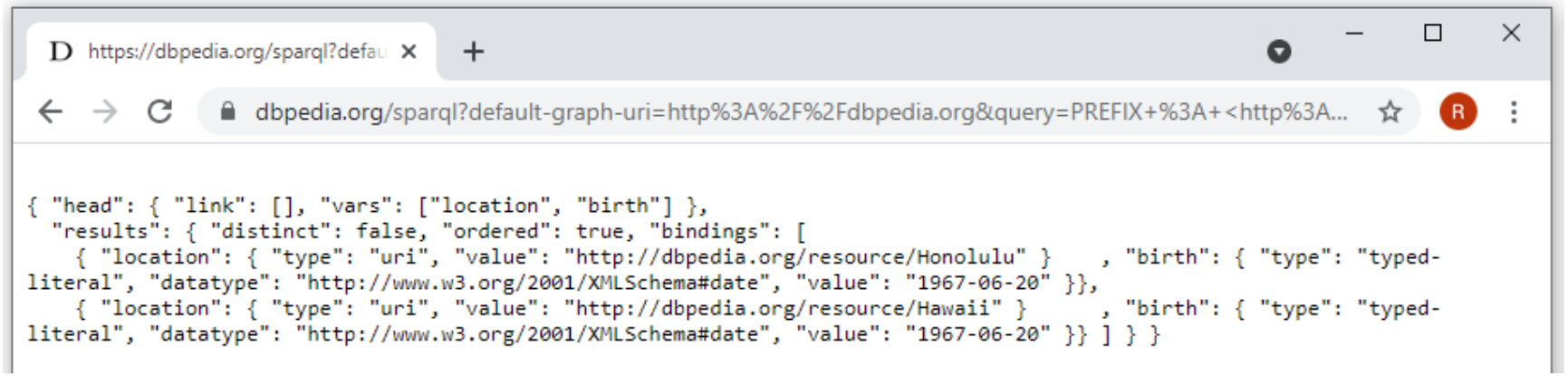
Querying DBpedia



Result in HTML



Result in JSON



Processing RDF Data

- One can get data from DBpedia in two ways:
  - either as an RDF serialization over HTTP
  - or by using a SPARQL endpoint.
- Data inside an RDF graph can be encoded in different ways in RDF/XML.
- Parsing RDF/XML using XML tools like DOM or SAX parsers may break.
- Use a specialised tool like Python's rdflib instead:
  - https://pypi.python.org/pypi/rdflib/

Processing DBpedia Data in Python

```
import rdflib

graph = rdflib.Graph()
graph.parse('http://dbpedia.org/resource/Nicole_Kidman')

print(len(graph),'\n')
print(list(graph)[300:301],'\n')

g = list(graph)[300:301]

for s, p, o in g:
    print("Subject:   ", s)
    print("Predicate: ", p)
    print("Object:    ", o)
    print()
```

Result

```
2588

[(rdflib.term.URIRef('http://dbpedia.org/resource/Cinema_of_the_United_States'),
 rdflib.term.URIRef('http://dbpedia.org/ontology/wikiPageWikiLink'),
 rdflib.term.URIRef('http://dbpedia.org/resource/Nicole_Kidman'))]

Subject:    http://dbpedia.org/resource/Cinema_of_the_United_States
Predicate:  http://dbpedia.org/ontology/wikiPageWikiLink
Object:     http://dbpedia.org/resource/Nicole_Kidman

# Note: there is no fixed order of triples; if you query in this way.
# You might get different results at position [300:301].
# Better to use a specific SPARQL query.
```

Remote SPARQL Query for DBpedia

```
from SPARQLWrapper import SPARQLWrapper, JSON

sparql = SPARQLWrapper("http://dbpedia.org/sparql")
sparql.setQuery("""
                PREFIX : <http://dbpedia.org/resource/>
                PREFIX dbo: <http://dbpedia.org/ontology/>

                SELECT ?location ?birth
                WHERE {
                       :Nicole_Kidman dbo:birthPlace ?location .
                       :Nicole_Kidman dbo:birthDate ?birth . }
                """)
```

Remote SPARQL Query for DBpedia

```
sparql.setReturnFormat(JSON)
results = sparql.query().convert()

for result in results["results"]["bindings"]:
    print(result["location"]["value"])
    print(result["birth"]["value"])
```

Result

```
http://dbpedia.org/resource/Honolulu
1967-06-20
http://dbpedia.org/resource/Hawaii
1967-06-20
```

Remote SPARQL Query for DBpedia

```
from SPARQLWrapper import SPARQLWrapper, JSON

sparql = SPARQLWrapper("http://dbpedia.org/sparql")
sparql.setQuery("""
                SELECT ?label
                WHERE { <http://dbpedia.org/resource/Semantic_Web>
                        rdfs:label ?label . }
                """)
sparql.setReturnFormat(JSON)
results = sparql.query().convert()
# print(results) # Shows entire JSON object
for result in results["results"]["bindings"]:
    print(result["label"]["value"])
```

Result

Semantic Web
ويب دلالي
Web semàntic
Sémantický web
Semantic Web
Σημασιολογικός Ιστός
Semantika Reto
Web semántica
Web semantiko
Web sémantique
Web semantik
Web semantico
セマンティック・ウェブ
…

Wikidata

- Wikidata is a collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation.

  https://www.wikidata.org/wiki/Wikidata:Main_Page
- It focuses on data items (93.5 million), which represent any kind of topics, concepts, or objects.
- Each item is allocated a unique, persistent identifier.
- Such a persistent identifier is known as a "QID".
- Examples of items include: 1988 Summer Olympics (Q8470), love (Q316), Elvis Presley (Q303), and Gorilla (Q36611).
- Check: https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual

Wikidata: SPARQL Query Service

Cats, with pictures

The following query uses these:

- Items: house cat (Q146)
- Properties: instance of (P31), image (P18)

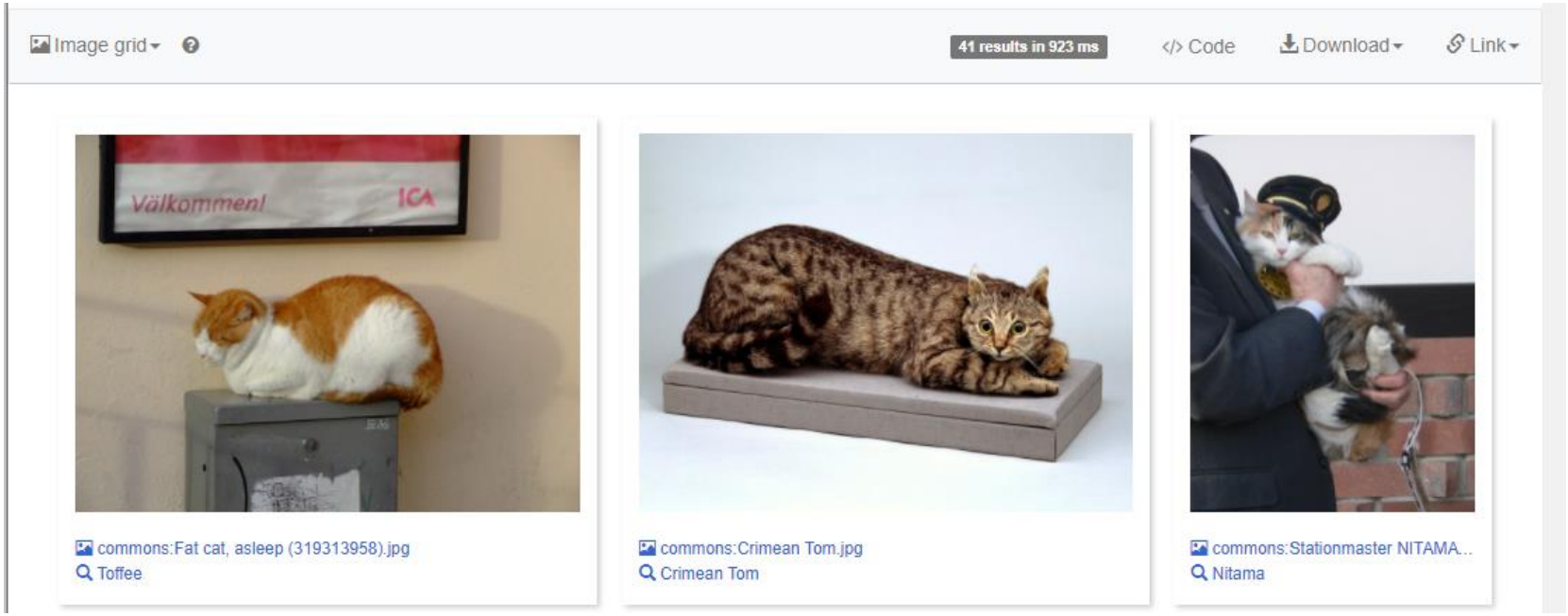
Features: ImageGrid (Q24515278)

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

https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples

Wikidata: SPARQL Query Service

Wikidata: SPARQL Query Service



Remote SPARQL Query for Wikidata

```
from SPARQLWrapper import SPARQLWrapper, JSON

sparql = SPARQLWrapper("https://query.wikidata.org/sparql")

sparql.setQuery("""
  SELECT ?item ?itemLabel ?pic
  WHERE
  {
    ?item wdt:P31 wd:Q146 .
    ?item wdt:P18 ?pic
    SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" }
  } """)

sparql.setReturnFormat(JSON)
results = sparql.query().convert()

print(results)
```

DBpedia versus Wikidata

- DBpedia mainly extracts structured data from info-boxes in Wikipedia, and publishes them in RDF and a few other formats.
- Wikidata turns the extraction process of DBpedia on its head: instead of extracting structured data from infoboxes, it allows (among other things) infoboxes to be created from structured data.
- DBpedia and Wikidata complement each other.

Take-Home Messages

- Linked data describes a method for publishing structured data building on URI/IRI, HTTP, and RDF.
- DBpedia is a knowledge base that represents structured content from Wikipedia pages.
- DBpedia can be seen as the nucleus of the web of data.
- DBpedia data can be processed via an RDF serialization over HTTP or by using a SPARQL endpoint.
- Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.