

UNIVERSITY OF JYVÄSKYLÄ

Lecture 7: Linked Data

TIES4520 Semantic Technologies for Developers
Autumn 2018

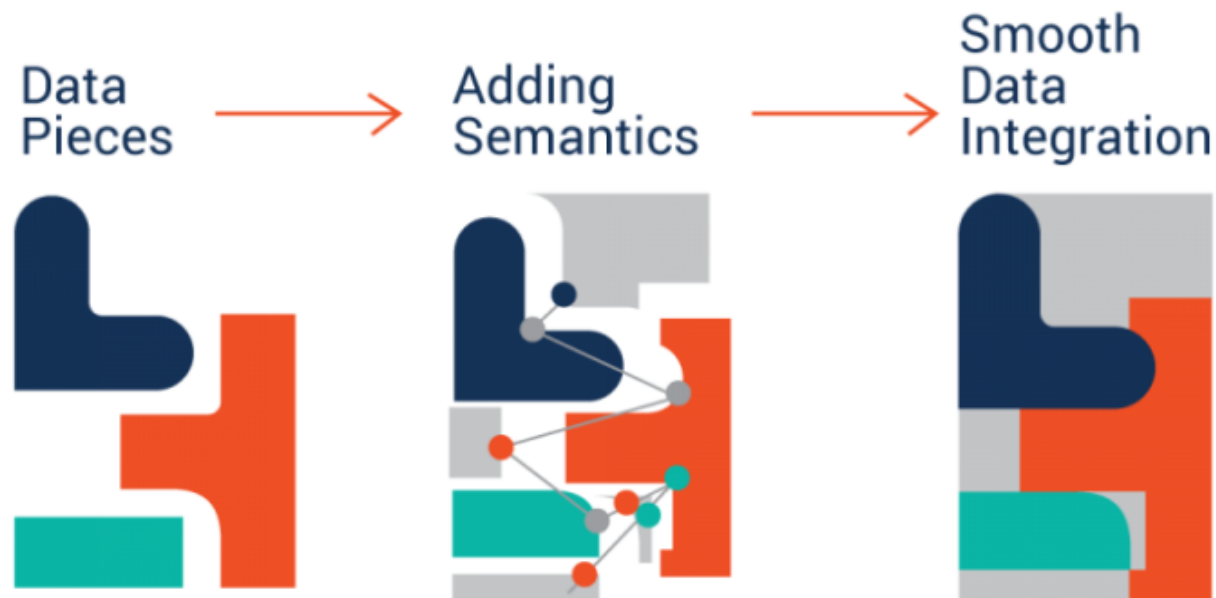


University of Jyväskylä

Khriyenko Oleksiy

Semantic Data Integration

Semantic data integration is the process of combining data from disparate sources and consolidating it into meaningful and valuable information through the use of Semantic Technology.



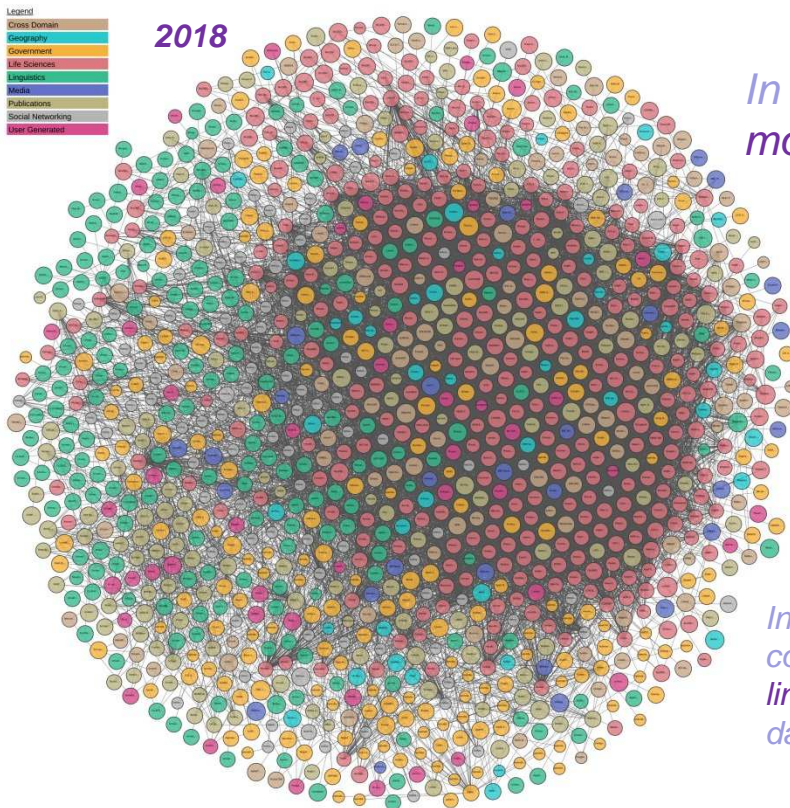
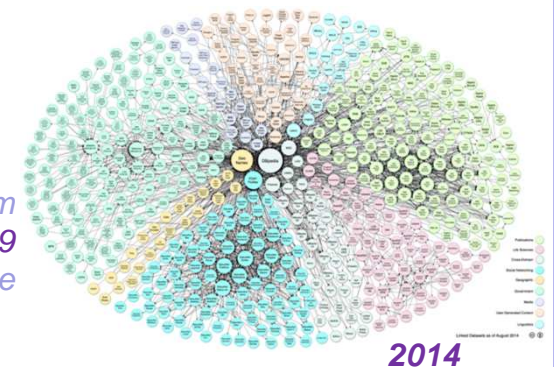
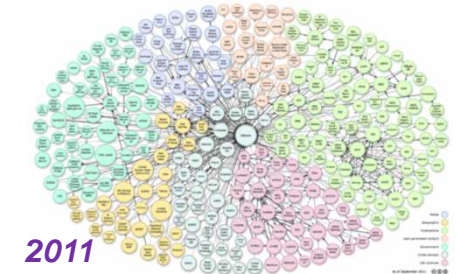
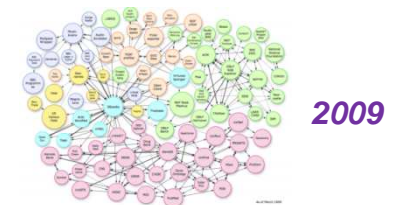
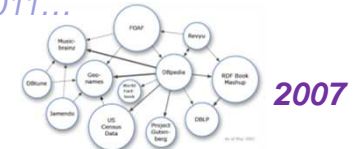
Relevant links: <https://www.ontotext.com/knowledgehub/fundamentals/semantic-data-integration/>
http://www.dit.unitn.it/~pavel/OM/articles/Cheatham_hBDT17.pdf
<https://ieeexplore.ieee.org/document/7889517>

Linked Data

Linked Data is a recommended best practice for exposing, sharing, and connecting pieces of *data*, *information*, and *knowledge* on the Semantic Web using *URIs* and *RDF* (Wikipedia). It is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods.

(<http://linkeddata.org>) (<http://linkeddatabook.com/editions/1.0/>)

volume of data has grown from around 2 billion triples in 2007 to over 30 billions in 2011...



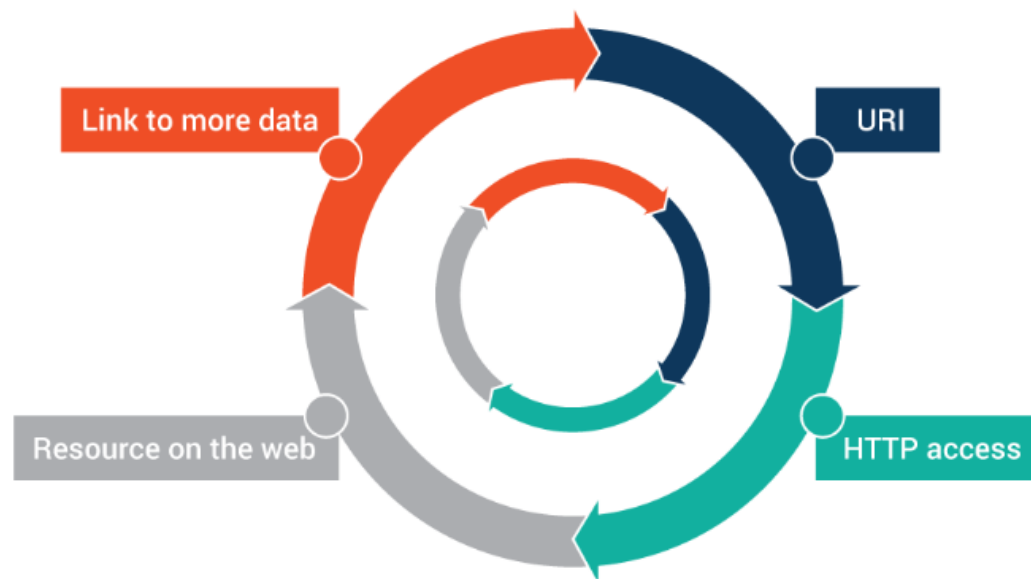
In 2018, 1229 datasets with more than 16,125 links...

In 2014, altogether, the diagram contains 570 datasets and 2909 linkage relationships between the datasets...

Linked Data

In 2006, *Tim Berners-Lee* set out four simple principles for publishing data on the web.

- Use URIs to identify things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, RDFS, SPARQL).
- Include links to other URIs, so that they can discover more things.



Relevant links: <https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/>

20/11/2018

TIES4520 - Lecture 7

Cool URIs – what's the problem?

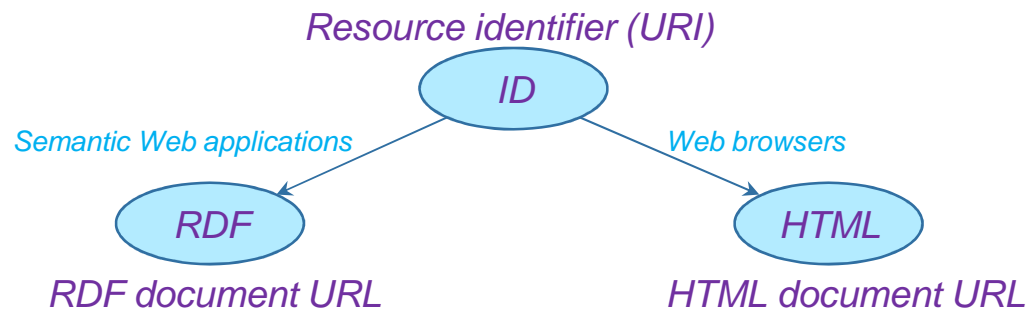
- W3C note from 2008 (<http://www.w3.org/TR/cooluris/>)
- URIs identify concepts (*real-world objects*)/(non-information resource)
- At the same time, web documents have always been addressed with URIs
- What URIs should we use in our RDF documents?
- Problem:
 - Alice is a real person that has a web page
 - What URI should represent Alice as an individual?
 - Her web page URL?
 - Her email address?

```
<URI-of-alice> a foaf:Person;  
foaf:name "Alice";  
foaf:mbox <mailto:alice@example.com>;  
foaf:homepage <http://example.com/people/alice> .
```

- Crucial concept: HTTP content negotiation

Cool URIs rules

- Be on the Web
 - Given only a URI of a resource, both machines and people will get the description of the resource
 - People will get human-readable HTML page
 - Machines will get RDF data
- Be unambiguous
 - No confusion between identifiers for Web documents and identifiers for other resources
 - One URI can't stand for both a Web document and real-world object (RWO)
- So which URI for Alice???



Cool URIs: good practice

- The URIs related to a single real-world object (non-information resource):
 - *resource identifier*
 - *HTML document URL*
 - *RDF document URL*

- Several ideas for choosing related URIs:

`http://smith-family.com/resource/alice`

- Identifier for Alice, the person

`http://smith-family.com/page/alice`

- Alice's homepage

`http://smith-family.com/data/alice`

- RDF document with description of Alice

`http://id.smith-family.com/alice`

- Identifier for Alice, the person

`http://pages.smith-family.com/alice`

- Alice's homepage

`http://data.smith-family.com/alice`

- RDF document with description of Alice

`http://smith-family.com/alice`

- Identifier for Alice, the person

`http://smith-family.com/alice.html`

- Alice's homepage

`http://smith-family.com/alice.rdf`

- RDF document with description of Alice

Solution: 303 URIs

- Use HTTP redirect status code *303 See Other*
 - to distinguish non-document resources from regular web documents
 - to point to the proper human-readable document

HTTP request:

```
GET /page/alice HTTP/1.1
Host: www.acme.com
Accept: text/html
Accept-Language: en, de
```

HTTP response (web document):

```
HTTP/1.1 200 OK
Content-Type: text/html
Content-Language: en
```

HTTP request:

```
GET /resource/alice HTTP/1.1
Host: www.acme.com
Accept: application/rdf+xml
```

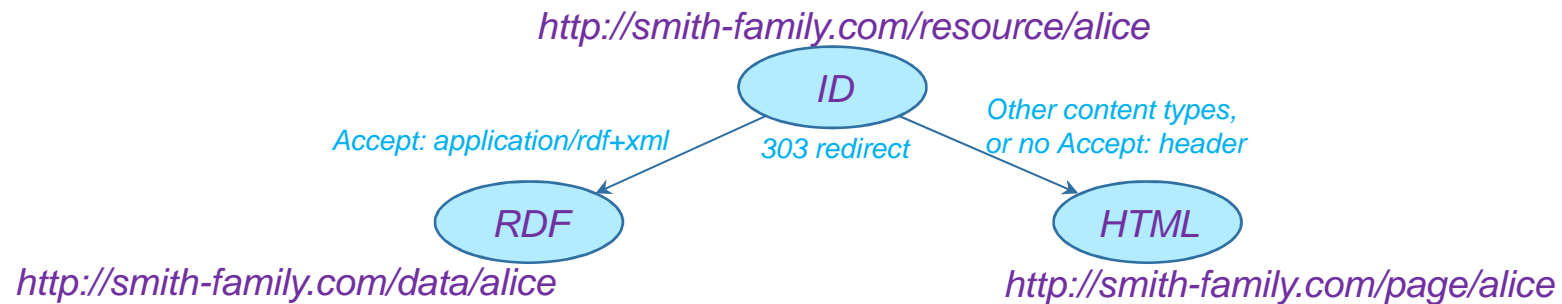
HTTP response (content negotiation):

```
HTTP/1.1 303 See Other
Location: http://www.acme.com/data/alice.rdf
Vary: Accept
```


Solution: 303 URIs

■ Alice, the person (RWO)

- Link: `http://smith-family.com/resource/alice`
- Machine access -> redirect to an RDF file <http://smith-family.com/data/alice>
- Human access -> redirect to address <http://smith-family.com/page/alice>

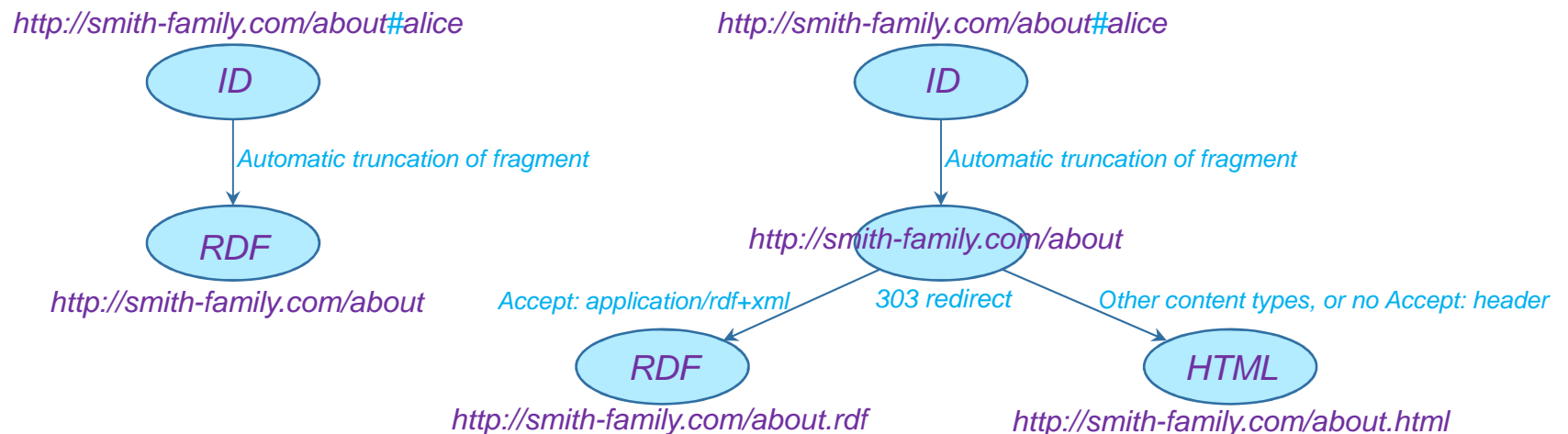


■ Alice, the document describing her (Web document)

- Link: `http://smith-family.com/page/alice`
- Human access -> returns HTML page
- Machine access -> returns RDF description of the *web page* or an error (page not found)
 - This could be a URI for the web page about Alice
 - This is not URI for Alice as a person

Solution: Hash URIs

- Use URIs with fragments (#) for non-information resources
 - URI with a hash cannot be retrieved directly (it is required to strip off the fragment part) and therefore cannot identify a web document
 - we can use them to identify other, non-information resources
- Alice, the person (RWO)
 - Link: `http://smith-family.com/about#alice`
 - Machine access -> returns `http://smith-family.com/about` as RDF file (which contains info about Alice)
 - Human access -> returns `http://smith-family.com/about` as HTML file



Hash URIs vs. 303 URIs

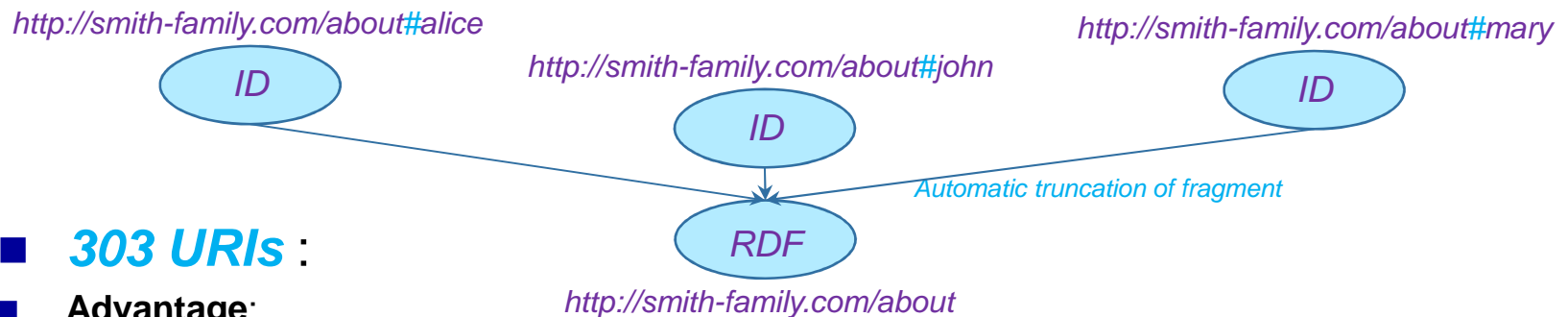
■ *Hash URIs* :

■ Advantage:

- reduced number of necessary HTTP requests
- a family of URIs can share the same non-hash part

■ Disadvantage:

- it loads other unrequested data



■ *303 URIs* :

■ Advantage:

- redirection target can be configured separately for each resource
- there could be one describing document for each resource, or one large document for all of them, or any combination in between.

■ Disadvantage:

- the large number of redirects may cause higher latency (waiting time)

Cool URIs: good practice

- All the URIs related to a single real-world object – *resource identifier*, *RDF document URL*, *HTML document URL* – should be explicitly linked with each other to help information consumers understand their relation.

`http://smith-family.com/resource/alice`

- Identifier for Alice, the person

`http://smith-family.com/page/alice`

- Alice's homepage

`http://smith-family.com/data/alice`

- RDF document with description of Alice

RDF file from <http://www.smith-family.com/data/alice>

```
...
<http://smith-family.com/resource/alice>
    foaf:page <http://smith-family.com/page/alice>;
    rdfs:isDefinedBy <http://smith-family.com/data/alice>;
    a foaf:Person;
    foaf:name "Alice";
    foaf:mbox <mailto:alice@acme.com>;
...
```

HTML file from <http://www.smith-family.com/people/alice>

```
<html lang="en">
  <head>
    <title>Alice's Homepage</title>
    <link rel="alternate" type="application/rdf+xml"
      title="RDF Version"
      href="http://smith-family.com/data/alice" />
  </head> ...
```

Linked Open Data

- In 2010, Tim Berners-Lee suggested a **5 star** deployment scheme for **Open Data** to encourage people (especially government data owners) to improve linked data.
- **Linked Open Data (LOD)** is Linked Data which is released under an open license, which does not impede its reuse for free. It denotes publicly available RDF Data in the Web, identified via URI and accessible via HTTP. **LOD2** - <http://lod2.eu/Welcome.html>

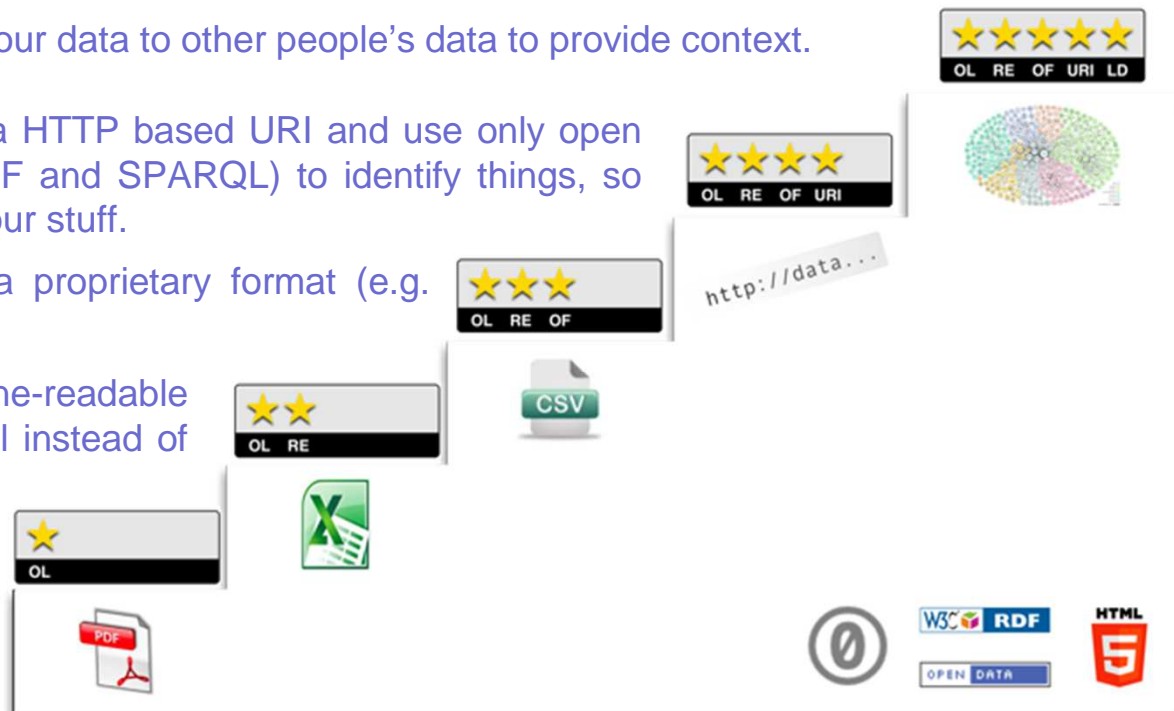
All the before, plus: Link your data to other people's data to provide context.

All the previous plus, data HTTP based URI and use only open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.

The data does not use a proprietary format (e.g. CSV instead of excel).

Available as machine-readable structured data (e.g. excel instead of image scan of a table).

Available on the web (whatever format) *but with an open license, to be Open Data.*

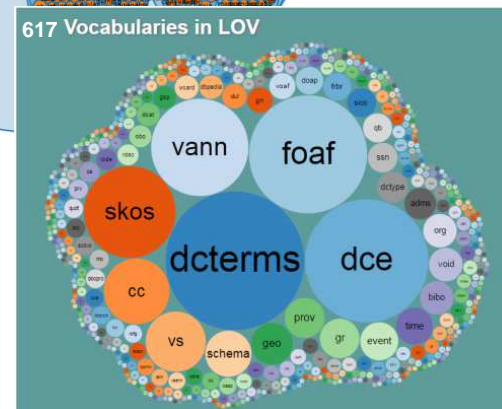
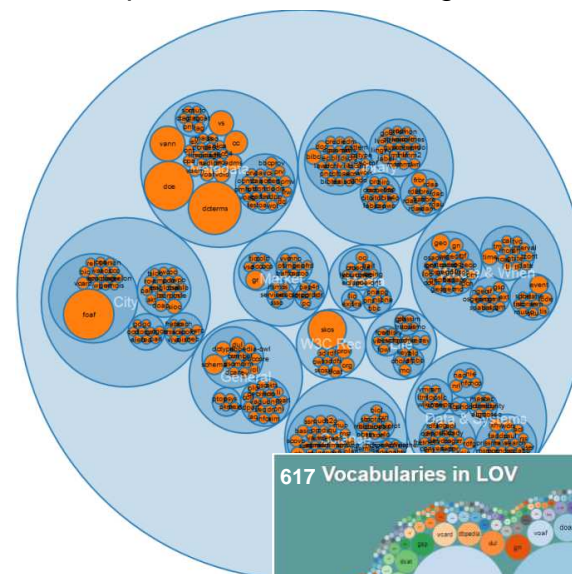


Some related materials: <http://harvardpolitics.com/online/open-data/>

Linked Data: good practice

- **Reusing existing well-known vocabularies.** In order to make it possible for client applications to process your data, you should *reuse terms from well-known vocabularies* wherever possible. You should only define new terms yourself if you can not find required terms in existing vocabularies. It is common practice to *mix terms* from different vocabularies.

- Google, Yahoo and Microsoft have agreed on vocabularies for publishing structured data on the Web. Their shared 'ontology' is maintained on schema.org.
- **Friend-of-a-Friend (FOAF)** provides terms for describing people and their social network
- **SIOC** Semantically-Interlinked Online Communities
- **DOAP** Description of a Project
- **Dublin Core** Defines general metadata attributes.
- **SKOS** Simple Knowledge Organization System
- **SKOS DataZone** list of vocabularies available in SKOS schema
- **Review Vocabulary** provides terms for representing reviews.
- **GoodRelations** provides terms for describing products and business entities.
- **Music Ontology** provides terms for describing artists, albums, tracks, but also performances, arrangements, etc.
- **Organization Ontology** for describing the structure of organizations.
- **Linking Open Description of Events (LODE)** provides terms for describing the basic properties of an event and contains a list of axioms expressing mapping relationships with other ontologies such as DOLCE, CYC, CIDOC-CRM, Event Ontology, F, and SEM.
- **MarineTLO (core) Ontology** is a top-level ontology for the marine domain (also applicable to the terrestrial domain) and **MarineTLO (imarine) Ontology** is an extension and operational version of the MarineTLO core.
- **SNOMED CT**, **Gene Ontology**, **Foundational Model of Anatomy**, **OpenGALEN**, etc. – are medical domain ontologies.
- etc.



Linked Open Vocabularies: <http://lov.okfn.org/dataset/lov/>

Well-known vocabularies: <http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies>

Linked Data: good practice

- **Reusing existing URIs.** If you need URI references for *geographic places, research areas, general topics, artists, books or CDs*, you should consider using URIs from existing data sources (for instance *Geonames, DBpedia, Musicbrainz, dbtune, RDF Book Mashup*, etc.). The two main benefits of using URIs from such data sources are:
 - The URIs are dereferenceable, meaning that a description of the concept can be retrieved from the Web.
 - The URIs are already linked to URIs from other data sources.

Well-known Data Sets:

<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets>

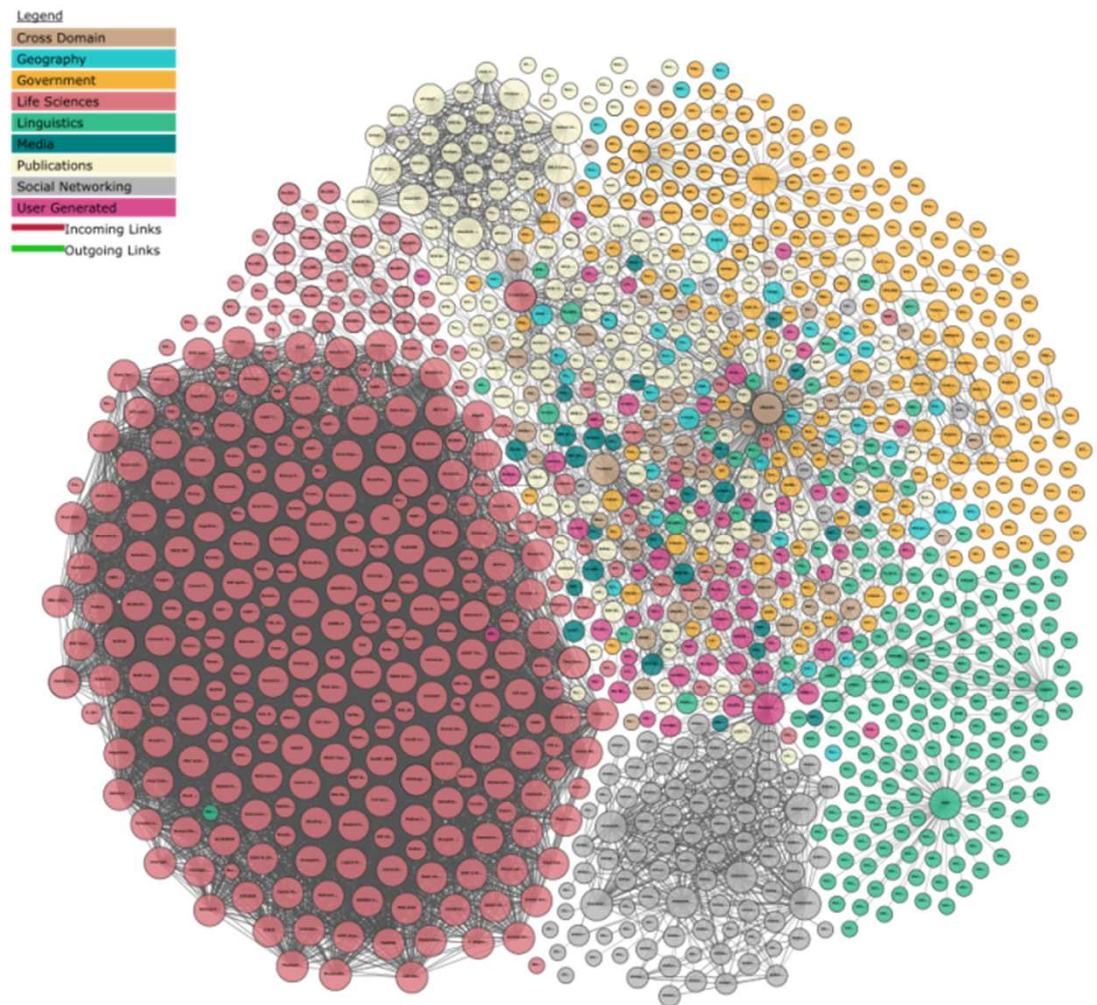
Linked Data Sets available as RDF Dumps:

<http://www.w3.org/wiki/DataSetRDFDumps>

SparqlEndpoints list and availability service:

<http://www.w3.org/wiki/SparqlEndpoints>

<http://sparql.es.ai.wu.ac.at/availability>



Linked Data: good practice

Guidance for *own term definition*:

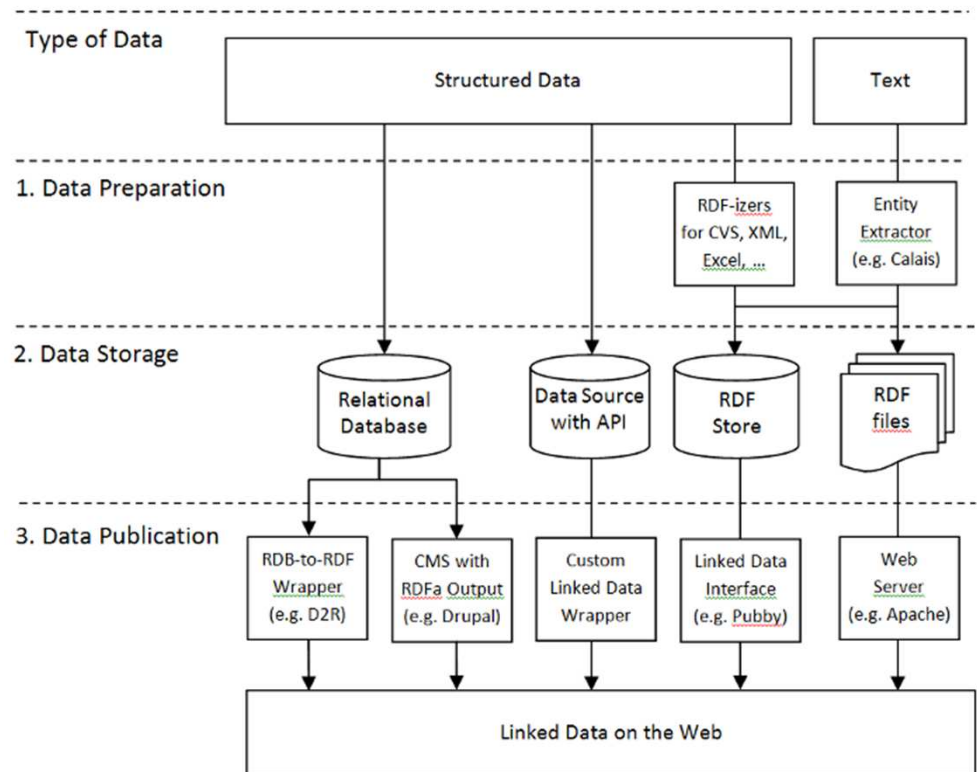
- *Do not define new vocabularies from scratch*, but complement existing vocabularies with additional terms (in your own namespace) to represent your data as required.
- *Provide for both humans and machines*. At this stage in the development of the Web of Data, more people will be coming across your code than machines, even though the Web of Data is meant for machines in the first instance. Don't forget to add prose, e.g. *rdfs:comment* for each term invented. Always provide a label for each term using the *rdfs:label* property.
- *Make term URIs dereferenceable*. It is essential that term URIs are dereferenceable so that clients can look up the definition of a term. Therefore you should make term URIs dereferenceable following the W3C Best Practice Recipes for Publishing RDF Vocabularies (<http://www.w3.org/TR/swbp-vocab-pub>).
- *Make use of other people's terms*. Using other people's terms, or providing mappings to them, helps to promote the level of data interchange on the Web of Data, in the same way that hypertext links built the traditional document Web. Common properties for providing such mappings are *rdfs:subClassOf* or *rdfs:subPropertyOf*.
- *State all important information explicitly*. For example, state all ranges and domains explicitly. Remember: humans can often do guesswork, but machines can't. Don't leave important information out!
- *Do not create over-constrained, brittle models; leave some flexibility for growth*. For instance, if you use full-featured OWL to define your vocabulary, you might state things that lead to unintended consequences and inconsistencies when somebody else references your term in a different vocabulary definition. Therefore, unless you know exactly what you are doing, use RDF-Schema to define vocabularies.

Best Practices for Publishing Linked Data: <https://www.w3.org/TR/ld-bp/>

Linked Data: publishing

Triple Stores (SPARQL Endpoints):

- *RDF4J (Sesame)*
- *GraphDB*
- *Fuseki*
- *OpenLink Virtuoso*
- *etc.*



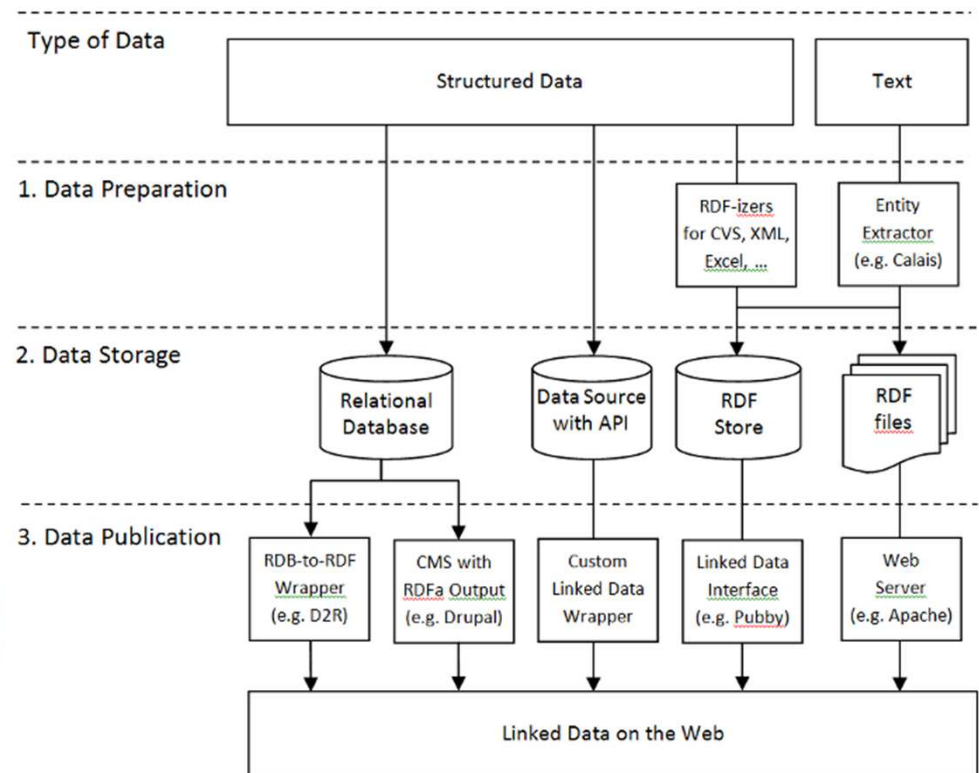
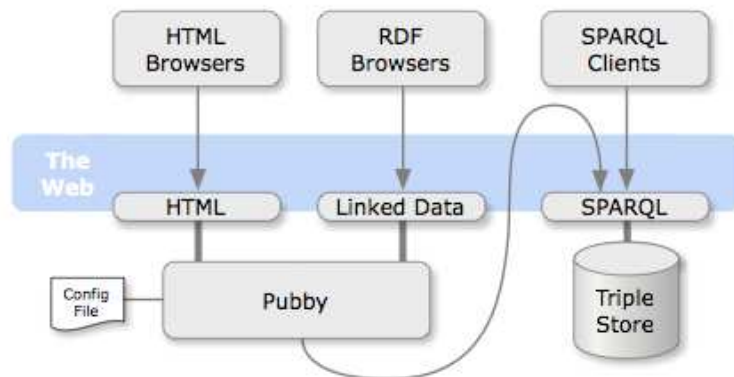
<http://linkeddatabook.com/editions/1.0/>

Linked Data: publishing

Linked Data Endpoints

Pubby is a Linked Data Frontend for SPARQL Endpoints. *Pubby* makes it easy to turn a SPARQL endpoint into a Linked Data server providing a Linked Data interface to those RDF data sources.

(<http://wifo5-03.informatik.uni-mannheim.de/pubby/>)



<http://linkedatabook.com/editions/1.0/>

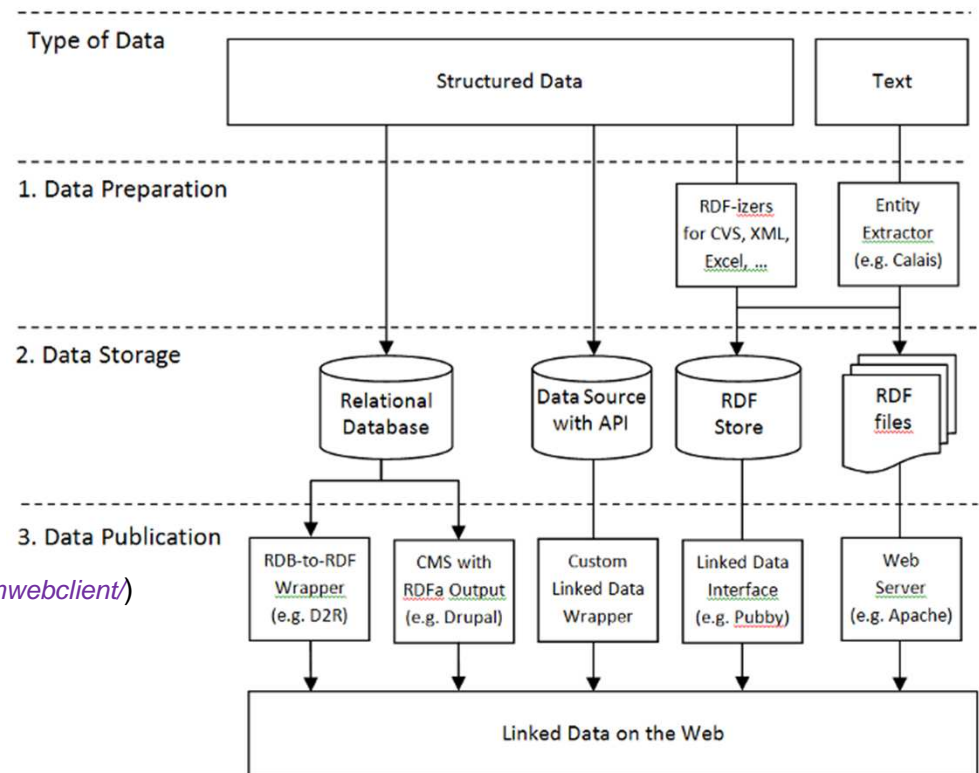
Linked Data: publishing

Linked Data Endpoints

Semantic Web Client Library

represents the complete Semantic Web as a single RDF graph. The library enables applications to query this global graph using SPARQL- and find(SPO) queries. To answer queries, the library dynamically retrieves information from the Semantic Web by dereferencing HTTP URIs, by following *rdfs:seeAlso* links, and by querying the *Sindice* search engine. The library is written in Java and is based on the Jena framework.

(<http://wifo5-03.informatik.uni-mannheim.de/bizer/ng4j/semwebclient/>)



<http://linkeddatabook.com/editions/1.0/>

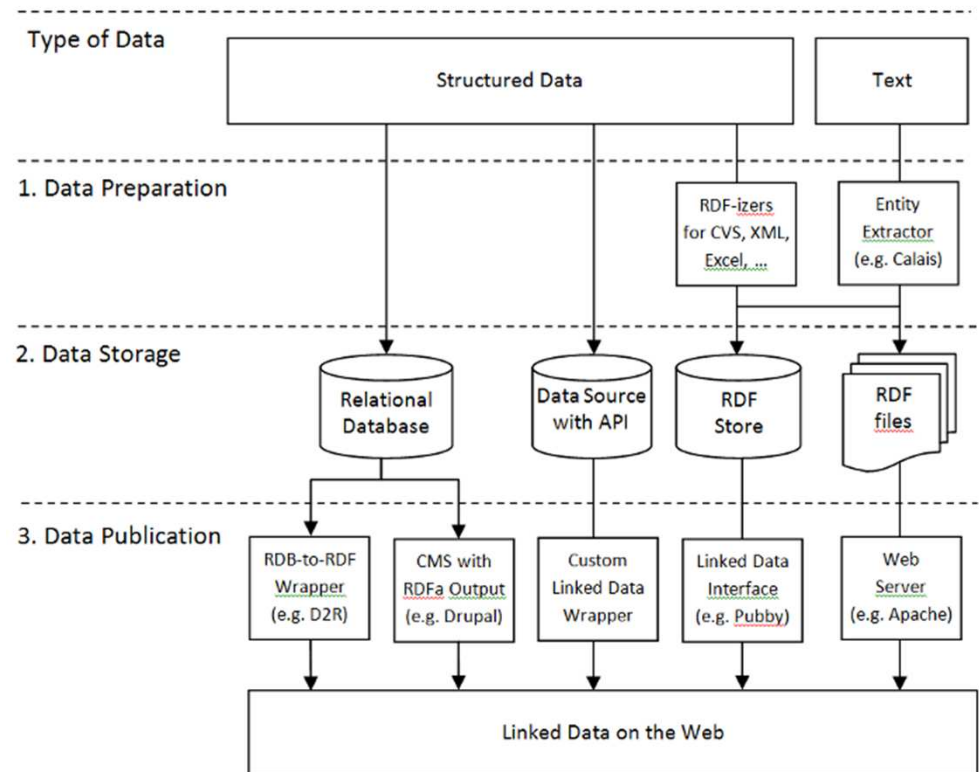
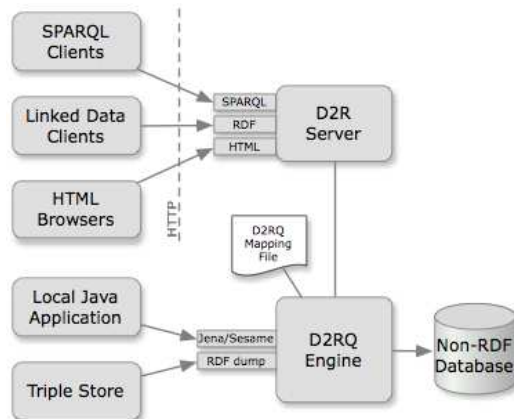
Linked Data: publishing

D2R Servers

D2RQ Platform is a system for accessing relational databases as virtual, read-only RDF graphs. It offers RDF-based access to the content of relational databases without having to replicate it into an RDF store. (<http://d2rq.org>)

Platform consists of:

- **D2RQ Mapping Language**, a declarative mapping language for describing the relation between an ontology and an relational data model.
- **D2RQ Engine**, a plug-in for the Jena Semantic Web toolkit, which uses the mappings to rewrite Jena API calls to SQL queries against the database and passes query results up to the higher layers of the frameworks.
- **D2R Server**, an HTTP server that provides a Linked Data view, a HTML view for debugging and a SPARQL Protocol endpoint over the database.



<http://linkeddatabook.com/editions/1.0/>



Linked Data: publishing

Mobi is a decentralized, federated, and distributed graph data platform for teams and communities to publish and discover data, data models, and analytics that are instantly consumable. (<https://mobi.inovexcorp.com>)

Mobi is built with *Apache Karaf* and utilizes *OWL 2* for authoring ontologies, the *SPARQL* query language for data lookup, and a pluggable backend system for processing and handling graph data modeled using the *Resource Description Framework (RDF)*.

The Mobi Solution Platform links all your data into an enterprise model, and provides the tools and APIs that empower your team to build better solutions:

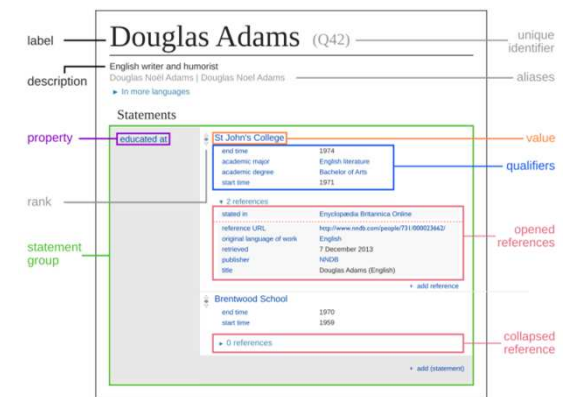
- *Collaborative Data Modeling* Bring your team or community together to create, share, and evolve data models in a modern, web-based, collaborative environment.
- *Automated Data Enhancement* Mobi makes it easy to integrate data into your models. The included mapping tool aligns your data to generate interoperable data without writing a single line of code.
- *Intuitive Data Exploration* Bundled tools make it easy to search and explore your enterprise data without the hassle of writing queries or analyzing models. Oh, and the query tools are there too.
- *Modular and Extensible Platform* Pluggable backend storage solutions, open REST and Java APIs, and a plugin framework make it easy to build solutions that extend and customize the Mobi Platform to solve your business problems.

Web of Data: knowledge base

Wikidata is the free knowledge base with **52,224,186 data items** that anyone can edit. It is a project of the *Wikimedia Foundation*: a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other Wikimedia projects, and well beyond that.

(<http://www.wikidata.org>)

- **Wikidata Toolkit** is an open source *Java library* for using data from Wikidata and other Wikibase sites. Its main goal is to make it easy for external developers to take advantage of this data in their own applications.



DBpedia is a crowd-sourced community effort to extract structured information from *Wikipedia* and make this information available on the Web. It is to improve **free and open data** and services for everyone. It is a movement towards a Public Data Infrastructure for a Large, Multilingual, Semantic Knowledge Graph. (<http://wiki.dbpedia.org/>)

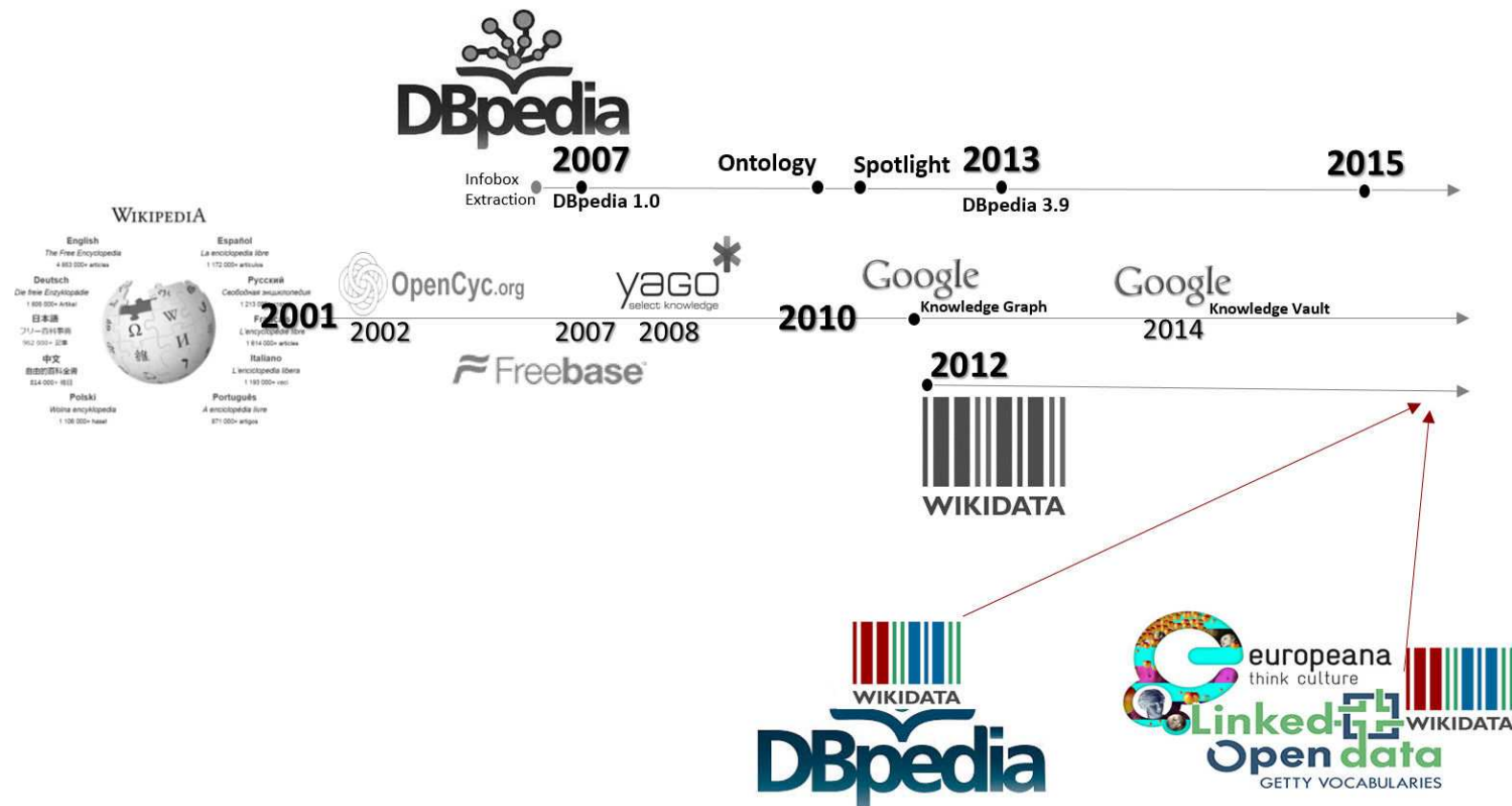
The English version of the DBpedia knowledge base describes **4.58 million things**, out of which 4.22 million are classified in a consistent *ontology*, including:

- 1,445,000 persons
- 735,000 places (including 478,000 populated places)
- 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games)
- 241,000 organizations (including 58,000 companies and 49,000 educational institutions)
- 251,000 species
- 6,000 diseases.



Web of Data: knowledge base

Knowledge Bases: Timeline



Web of Data: knowledge base

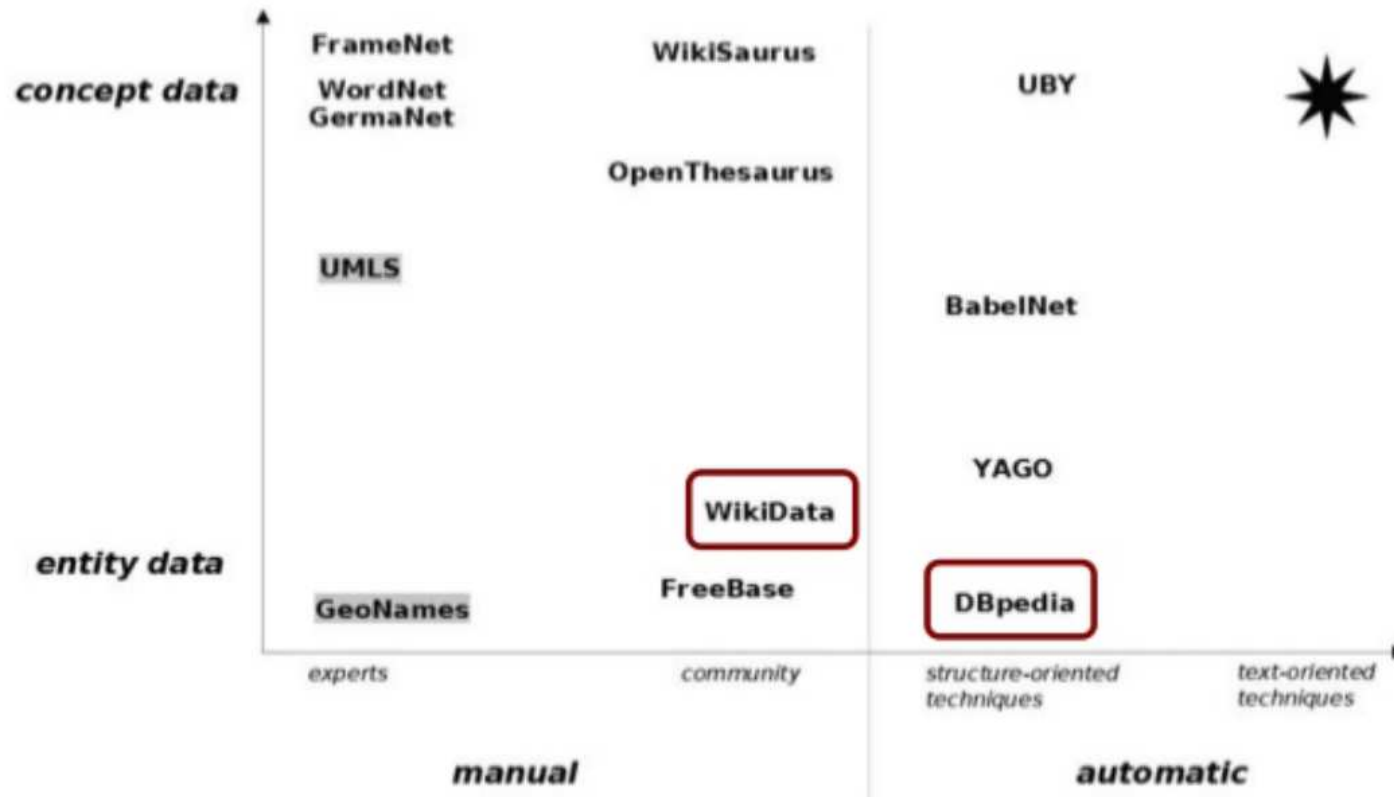


Fig. 1. Classification of selected background knowledge resources.

Arnold, P., & Rahm, E. (2015). Automatic Extraction of Semantic Relations from Wikipedia. *International Journal on Artificial Intelligence Tools*, 24(2), 1540010.

Source; Arnold, P., & Rahm, E. (2015). Automatic Extraction of Semantic Relations from Wikipedia. *International Journal on Artificial Intelligence Tools*, 24(2), 1540010.

A preliminary study on Wikipedia Dbpdeia and Wikidata (2015):

<https://www.slideshare.net/andreasinica/a-preliminary-study-on-wikipedia-dbpdeia-and-wikidata>

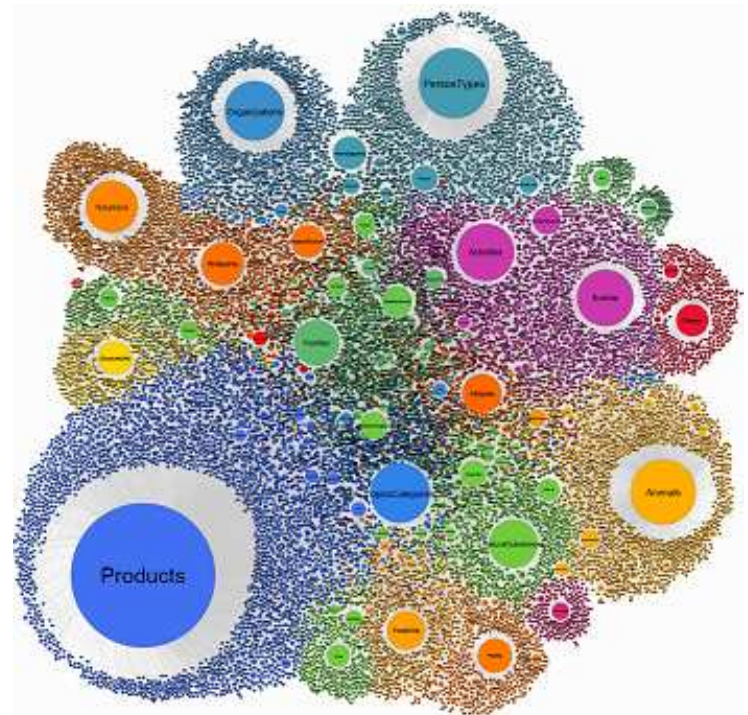


KBpedia

Web of Data: knowledge base

KBpedia is complete open source second-generation knowledge graph successor to **UMBEL** (Upper Mapping and Binding Exchange Layer – <http://umbel.org/>), and includes an upper ontology (KKO), full knowledge graph, mappings to major leading knowledge bases, and 70 logical concept groupings called typologies. (<http://kbpedia.org/>)

- is a comprehensive knowledge structure for promoting data interoperability and knowledge-based artificial intelligence. Its structure combines seven 'core' public knowledge bases: **Wikipedia**, **Wikidata**, **schema.org**, **DBpedia**, **GeoNames**, **OpenCyc**, and **UMBEL** — into an integrated whole. KBpedia's upper structure, or knowledge graph, is the KBpedia Knowledge Ontology (KKO), which is based on the universal categories and knowledge representation theories of the great 19th century American logician, polymath and scientist, Charles Sanders Peirce.
- Is written primarily in **OWL 2**, includes **55,000** reference concepts, about **30 million** entities, and **5,000** relations and properties, all organized according to about 70 modular typologies that can be readily substituted or expanded.
- exploits large-scale knowledge bases and semantic technologies for **machine learning**, **data interoperability and mapping**, and **fact extraction and tagging**. It is a flexible and computable knowledge graph that can be sliced-and-diced and configured for all sorts of machine learning tasks, including *supervised*, *unsupervised* and *deep learning*.



Relevant links: <http://fgiasson.com/blog/index.php/2016/11/07/building-and-maintaining-the-kbpedia-knowledge-graph>

Web of Data Tools

Linked Data Browsers and Mashup Applications

(<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/SemWebClients>)

RelFinder is a tool from Visual Data Web initiative for Interactive Relationship Discovery in RDF Data.

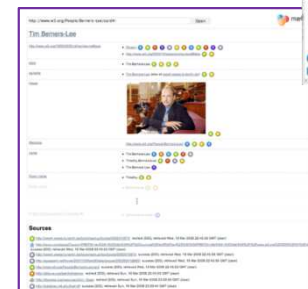
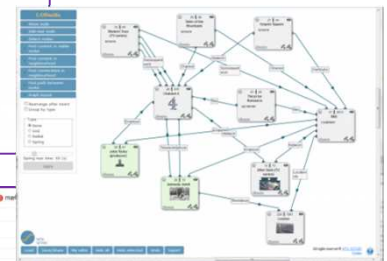
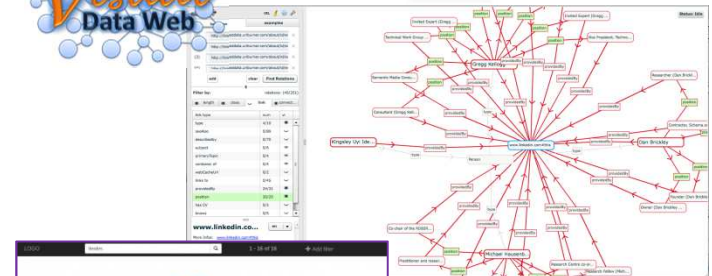
(<http://www.visualdataweb.org/relfinder.php>)

Bubble Navigator is a tool for visual navigation of semi-structured or semantic web data. (<http://wiki.dbpedia.org/projects/bubble-navigator>)

LODmilla is a DBpedia visualization service. LODmilla aims at visualizing associations in LOD graphs with special linked data functions such as: searching and exploring the neighborhood of a resource node, saving and sharing graph views, doing minor edits on triples, etc. (<http://wiki.dbpedia.org/projects/lodmilla>)

Marbles is a server-side application that formats Semantic Web content for XHTML clients using Fresnel lenses and formats. Colored dots are used to correlate the origin of displayed data with a list of data sources, hence the name. (<http://mes.github.io/marbles/>)

...



Web of Data Tools

Large-scale Graph Visualization

(<http://www.mkbergman.com/414/large-scale-rdf-graph-visualization-tools/>)

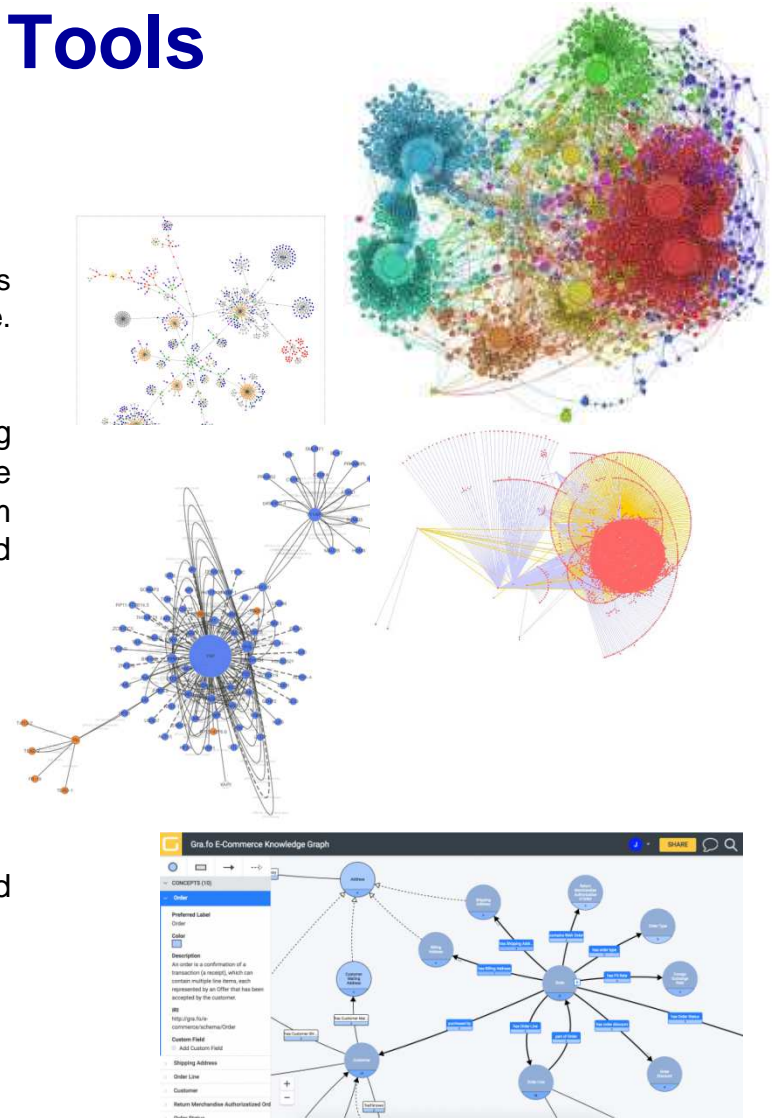
Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free. (<https://gephi.org/>)

Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. A lot of *Apps* are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web. (<https://cytoscape.org/>)

Knowledge Graph Schema Editor

Gra.fo is a visual, collaborative, real-time ontology and knowledge graph schema editor. (<https://gra.fo/>)

Relevant links: <https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154/1009>
<http://www.mkbergman.com/968/a-new-best-friend-gephi-for-large-scale-networks/>
<http://www.juansequeda.com/blog/2018/10/19/gra-fo-a-visual-collaborative-real-time-ontology-and-knowledge-graph-schema-editor/>

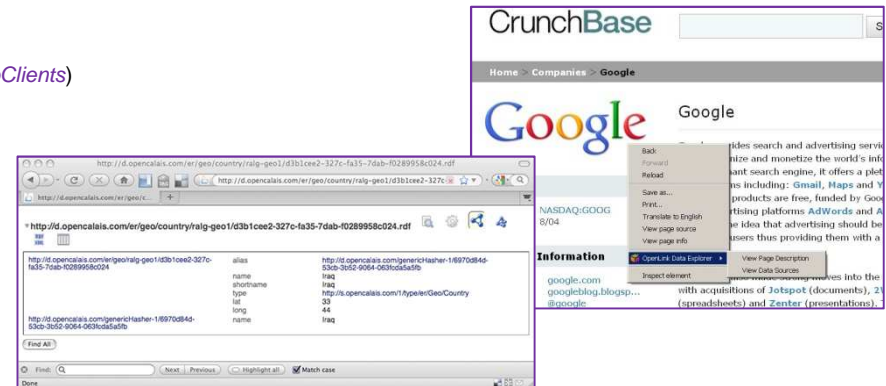


Web of Data Tools

Linked Data Browsers and Mashup Applications

(<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/SemWebClients>)

- *Tabulator*
- *DBpedia Mobile*
- *OpenLink Data Explorer*
- *Quick & Dirty RDF Browser*
- *Etc.*



Semantic Web Search Engines

(<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/SemanticWebSearchEngines>)

- *<sameAs.org>*
- *VisiNav*
- *Falcons*
- *Sindice*
- *Watson*
- *Swoogle*
- *Etc.*



Web of Data Tools

Google Knowledge Graph is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources. It provides extra information of the searched thing (object, topic, etc.) and links to other relevant things.

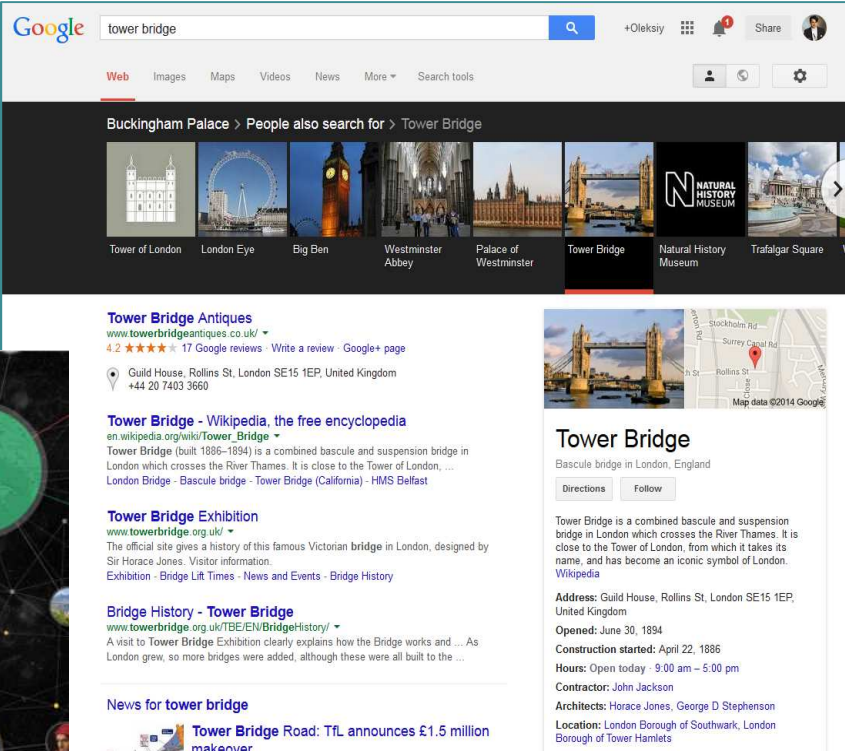
(<https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html>)

Link: https://www.youtube.com/watch?v=mg91_trV4hY

<https://www.youtube.com/watch?v=mmQl6VGvX-c>

The **Knowledge Graph Search API** lets you find entities in the **Google Knowledge Graph**. The API uses standard *schema.org* types and is compliant with the *JSON-LD* specification.

(<https://developers.google.com/knowledge-graph/>)



The screenshot shows a Google search for "tower bridge". The Knowledge Graph results are displayed, including a carousel of related images (Tower of London, London Eye, Big Ben, Westminster Abbey, Palace of Westminster, Tower Bridge, Natural History Museum, Trafalgar Square). Below the carousel, there are several information cards:

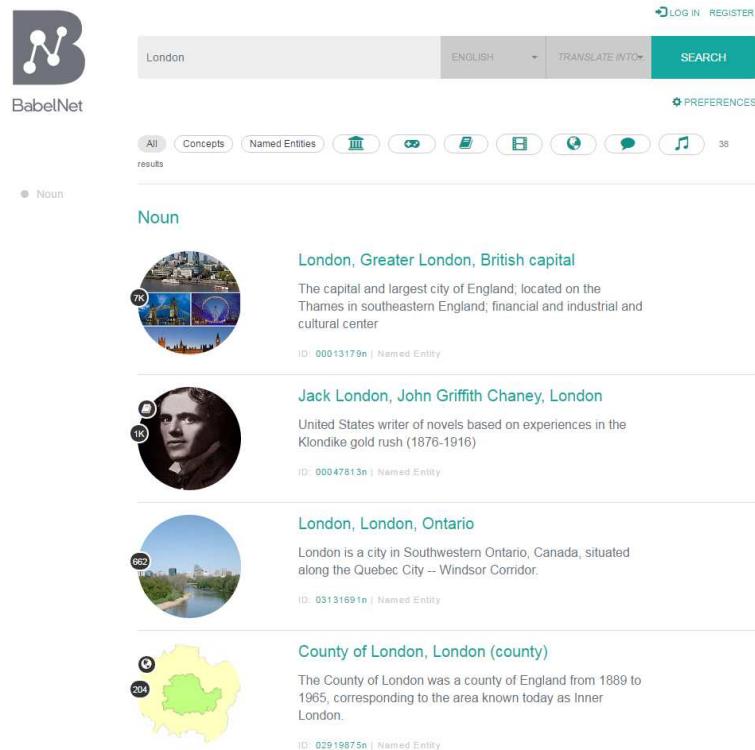
- Tower Bridge Antiques**: www.towerbridgeantiques.co.uk/, 4.2 stars, 17 Google reviews. Address: Guild House, Rollins St, London SE15 1EP, United Kingdom.
- Tower Bridge - Wikipedia, the free encyclopedia**: en.wikipedia.org/wiki/Tower_Bridge. Description: Tower Bridge (built 1886–1894) is a combined bascule and suspension bridge in London which crosses the River Thames. It is close to the Tower of London, ...
- Tower Bridge Exhibition**: www.towerbridge.org.uk/. Description: The official site gives a history of this famous Victorian bridge in London, designed by Sir Horace Jones. Visitor information. Exhibition - Bridge Lift Times - News and Events - Bridge History.
- Bridge History - Tower Bridge**: www.towerbridge.org.uk/TBE/ENBridgeHistory/. Description: A visit to Tower Bridge Exhibition clearly explains how the Bridge works and ... As London grew, so more bridges were added, although these were all built to the ...
- News for tower bridge**: **Tower Bridge Road: TfL announces £1.5 million makeover**.

On the right side, there is a detailed information card for **Tower Bridge**, including a map, directions, and follow button. The card states: "Bascule bridge in London, England", "Opened: June 30, 1894", "Construction started: April 22, 1886", "Hours: Open today - 9:00 am - 5:00 pm", "Contractor: John Jackson", "Architects: Horace Jones, George D Stephenson", "Location: London Borough of Southwark, London Borough of Tower Hamlets".



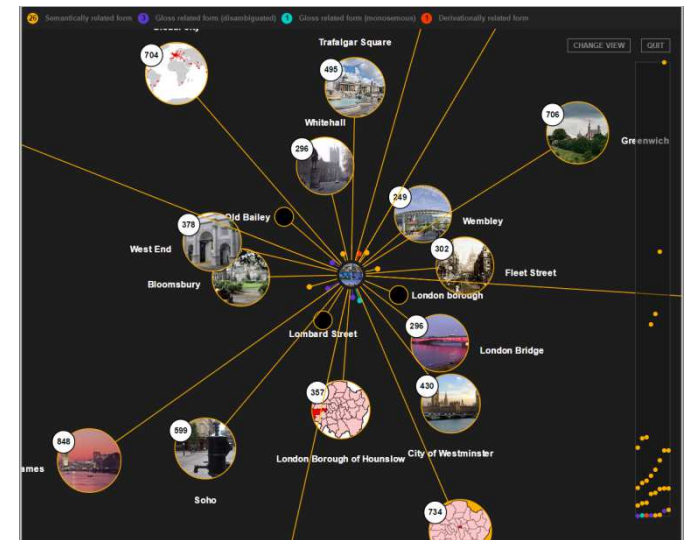
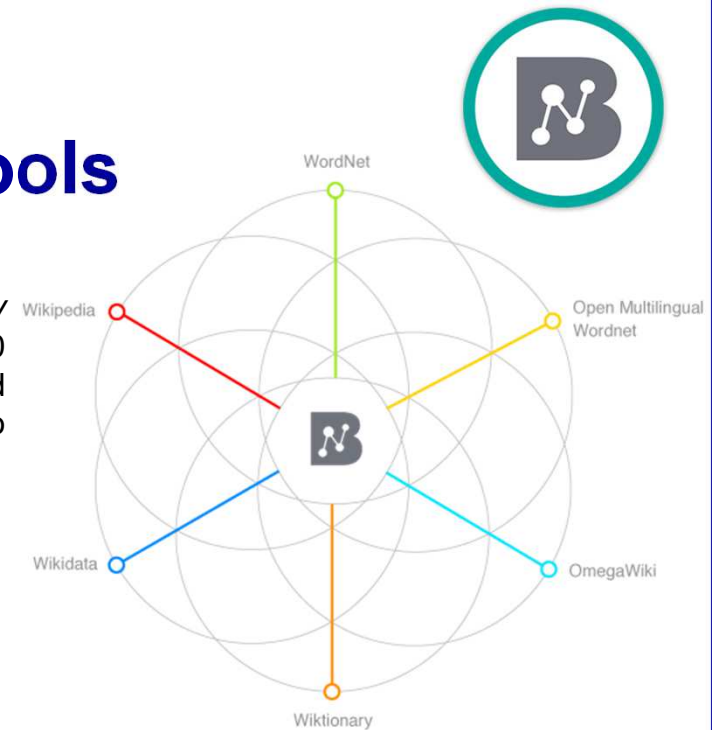
Web of Data Tools

BabelNet is both a very large *multilingual encyclopedic dictionary* (with lexicographic and encyclopedic coverage of terms in 50 languages) and a *semantic network* which connects concepts and named entities in a very large network of semantic relations, made up to more than 9 million entities. (<http://babelnet.org/>)



The screenshot shows the BabelNet search interface. At the top, there's a search bar with 'London' entered, a language dropdown set to 'ENGLISH', and a 'TRANSLATE INTO' button. Below the search bar are tabs for 'All', 'Concepts', and 'Named Entities'. The 'Named Entities' tab is selected, showing results for 'Noun'. The results list four entries for 'London':

- London, Greater London, British capital**: The capital and largest city of England, located on the Thames in southeastern England; financial and industrial and cultural center. (ID: 00013179n | Named Entity)
- Jack London, John Griffith Chaney, London**: United States writer of novels based on experiences in the Klondike gold rush (1876-1916). (ID: 00047613n | Named Entity)
- London, London, Ontario**: London is a city in Southwestern Ontario, Canada, situated along the Quebec City – Windsor Corridor. (ID: 03131691n | Named Entity)
- County of London, London (county)**: The County of London was a county of England from 1889 to 1965, corresponding to the area known today as Inner London. (ID: 02919875n | Named Entity)



Web of Data Tools



Refer is an online-recommendation system based on Linked Open Data and Semantic Web Technologies. It aims to improve the user's and author's experience while curating and navigating in blogs, multimedia platforms, and archives. (<http://refer.cx/>)

The screenshot displays the Refer web interface. At the top right, there are tabs for 'Relation Browser' and 'Timeline'. The main area shows a semantic network with nodes like 'Robert Bunsen', 'Leighlinbridge', 'Haslemere', 'County Carlow', 'Dublin', 'Marburg', 'University of Marburg', 'Tyndall effect', 'Diamagnetism', 'Atmosphere', 'Infrared', 'Physics', 'Chemistry', 'Europe', 'England', 'United Kingdom', and '1820'. Relationships are indicated by lines and labels like 'is city of' and 'is almaMater of'.

Below the network, there is a section for 'John Tyndall' with a portrait and a detailed biography. To the left of the biography, there is a list of '15 Recommended Articles'.

At the bottom, there is a text editor interface with a toolbar and a text area. The text area contains a sentence about Mickey Mouse: 'Mickey Mouse star in Walk of Fame'. The text editor also shows a 'Insert/edit entity' dialog box.

Web of Data Tools

ExConQuer a linked data publication and consumption framework consists of two tools:

- **Query Builder Tool:** Works on top of DBpedia (or other SPARQL endpoints) and enables users to construct a SPARQL query without requiring any knowledge of SPARQL or the datasets' underlying schema. Users are then able to download the data they require in a number of different formats.
- **PAM Tool:** A faceted browser that allows users to browse and re-use any queries executed within the Query Builder, and either directly download the results or otherwise re-load the query in the Query Builder and edit it accordingly.

Link: <http://wiki.dbpedia.org/projects/exconquer>

Framework Demo video: <https://vimeo.com/164145033>

Quepy is a python framework to transform natural language questions to queries in a database query language. It can be easily customized to different kinds of questions in natural language and database queries. So, with little coding you can build your own system for natural language access to your database.

Link: <http://quepy.machinalis.com/>

Similar projects:

AskNow - Question Answering (QA) system for RDF datasets

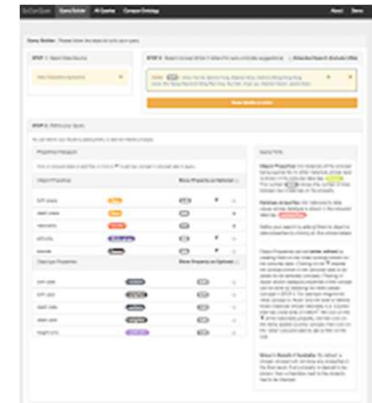
<http://sda.cs.uni-bonn.de/projects/asknow/>

LC-QuAD - Largescale Complex Question Answering Dataset

<http://lc-quad.sda.tech/>

DeFacto (Deep Fact Validation) is an algorithm for validating statements by finding confirming sources for it on the web.

<http://sda.cs.uni-bonn.de/projects/defacto/>



Try it yourself:

- 1. Ask a question**
Start by asking a question in natural language and watch the query generated.
Question:
- 2. Get a query**
This query was generated for the question "What is the capital of Finland?"

SPARQL

```

PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX quepy: <http://www.machinalis.com/quepy#>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX dbprop: <http://dbpedia.org/property/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x2 WHERE {
  ?x0 rdf:type dbpedia-owl:Country.
  ?x0 rdfs:label "Finland"@en.
  ?x0 dbpedia-owl:capital ?x1.
  ?x1 rdfs:label ?x2.
}
```

JSON

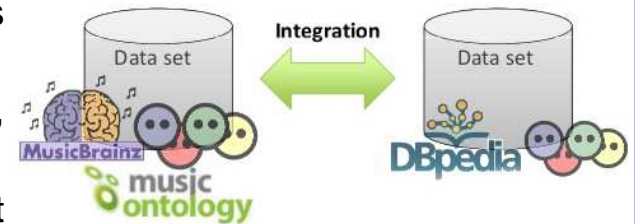
```

[[{"location/country/capital": [{"type/object/name": [{"type/object/name": "Finland", "type/object/type": "location/country"}]}]]
```
- 3. Get an answer**
Click the button to query the DBpedia database and see if there's any answer for this query.

Linked Data Integration

Linked Data applications that want to consume data from the global data space face following challenges:

- Data sources use a wide range of different RDF vocabularies to represent data about the same type of entity;
- The same real-world entity, for instance a person or a place, is identified with different URIs within different data sources;
- Data about the same real-world entity coming from different sources may contain conflicting value.



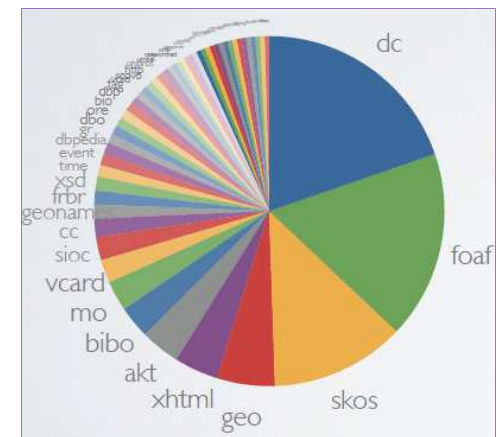
4 steps to Linked Data Integration:

Step#1: *Access linked data:* on-the-fly dereferencing, Query federation, Crawling and Caching.

Step#2: *Normalize vocabularies.* Schema Mapping can be performed based on rules or SPARQL queries.

Step#3: *Resolve identifiers.* Most LOD sources only provide *owl:sameAs* links to one other data source. Identity resolution can be done by manual merging or rule-based approaches (e.g. SILK, LINES)

Step#4: *Filter Data.* Due to the different knowledge levels, views and intents of data sources as well as wrong, inconsistent or outdated information, data can be stored and queried separately using named graphs based structure of a storage.



Linked Data supporting Tools

- **LDIF** (Linked Data Integration Framework) integrates Linked Data from multiple sources into a clean, local target representation while keeping track of data provenance (<http://ldif.wbsg.de/>)

The **LDIF integration pipeline** consists of the following steps:

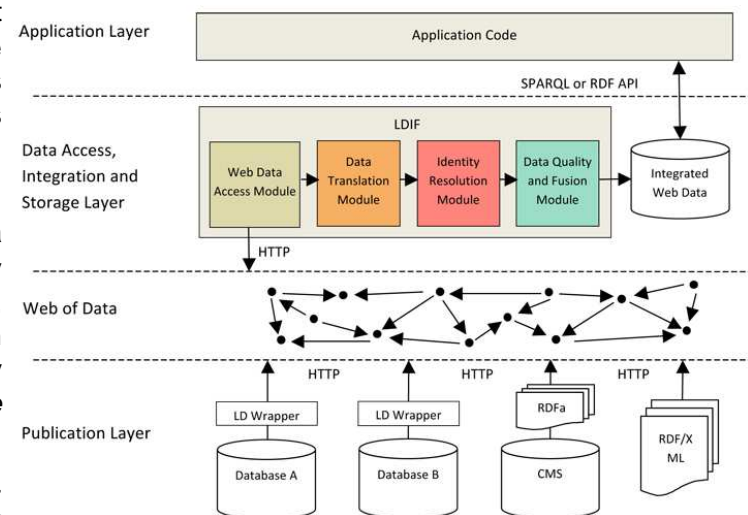
Step#1: Collect data. Import modules locally replicate data sets via file download, crawling or SPARQL. Supported data sources: RDF dumps (all common formats), SPARQL Endpoints, Crawling Linked Data via HTTP.

Step#2: Translate data (map to Schema). An expressive mapping language allows for translating data from the various vocabularies that are used on the Web into a consistent, local target vocabulary. LDIF supports simple mappings using OWL/RDFS statements (x rdfs:subClassOf y), complex mappings with SPARQL expressivity, built-in transformation function library (XPath) as well as *R2R Framework* (<http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/>).

Step#3: Resolve identifiers. An identity resolution component discovers URI aliases in the input data and replaces them with a single target URI based on user-provided matching heuristics. LDIF uses automated link creation based on SILK Link Specifications as well as supports various comparators and transformations (string similarity, basic arithmetics, time, geographical distance).

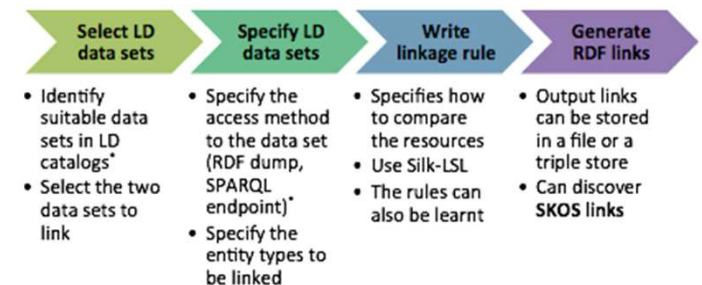
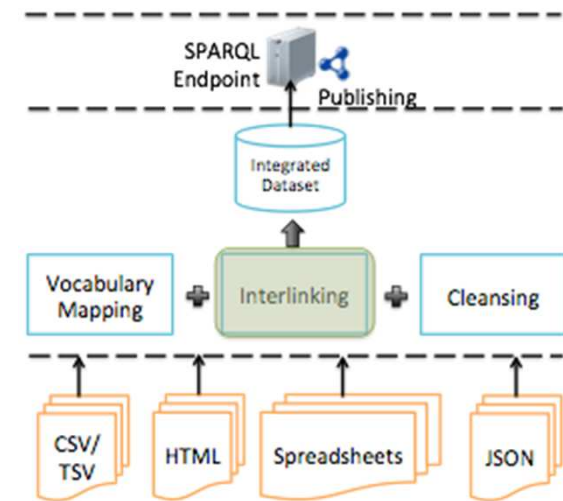
Step#4: Cleanse data. A data cleansing component filters data according to different quality assessment policies (assign quality scores to Named Graphs by time, by source preference, thresholds, etc.) and provides data fusion according to different conflict resolution methods (resolve conflicting property values according to quality scores, frequency, averages, etc.). LDIF employs *Sieve* (<http://sieve.wbsg.de/>).

Step#5: Output. LDIF outputs the integrated data in N-Quads, N-Triples or SPARQL Update Stream. For provenance tracking, LDIF employs the Named Graphs data model.



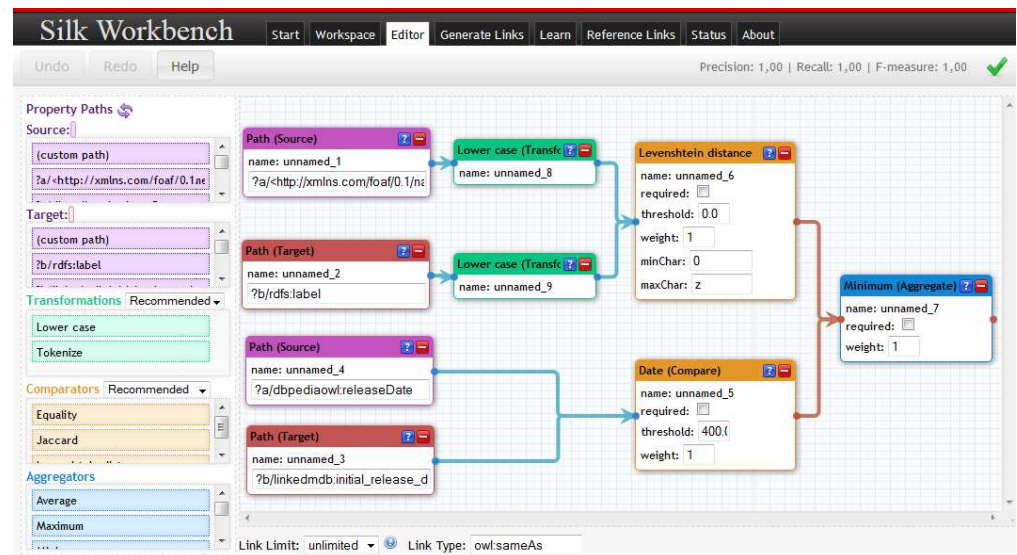
Linked Data supporting Tools

- **SILK** – a Link Discovery Framework for the Web of Data. SILK is an open source tool for discovering RDF links between data items within different Linked Data sources. (<http://silkframework.org/>)
- **Silk Link Specification Language** (*Silk-LSL*) is used to define rules for linking entities from two different datasets. For example, a rule may express that if two entities belong to specified classes and have matching labels then they should be linked by a certain property. This property could be *owl:sameAs* or some other property such as *skos:closeMatch*.
- **SILKS** can run in different variations:
 - Silk Single Machine:*
 - Generate links on a single machine
 - Local or remote data set
 - Silk MapReduce:*
 - Generate RDF links using a cluster of multiple machines
 - Based on Hadoop (Can be run on Amazon Elastic MapReduce)
 - Silk Server :*
 - Provides an HTTP API for matching instances for an incoming stream of RDF data while keeping track of known entities
 - Can be used as an identity resolution component within applications that consume Linked Data from the Web



Linked Data supporting Tools

- **SILK Workbench** is a web application built on top of *SILK* that can be used to create projects and manage the creation of links between two RDF datasets.
(<http://silkframework.org/>, https://www.assembla.com/spaces/silk/wiki/Silk_Workbench)
- The *SILK Workbench* has a graphical editor that can be used to create linkage rules. Support is also provided for the automatic learning of linkage rules.
- The *SILK Workbench* also provides an interface for examining automatically learned rules. These suggested rules can then be added to the set of linkage rules or rejected.



Linked Data supporting Tools

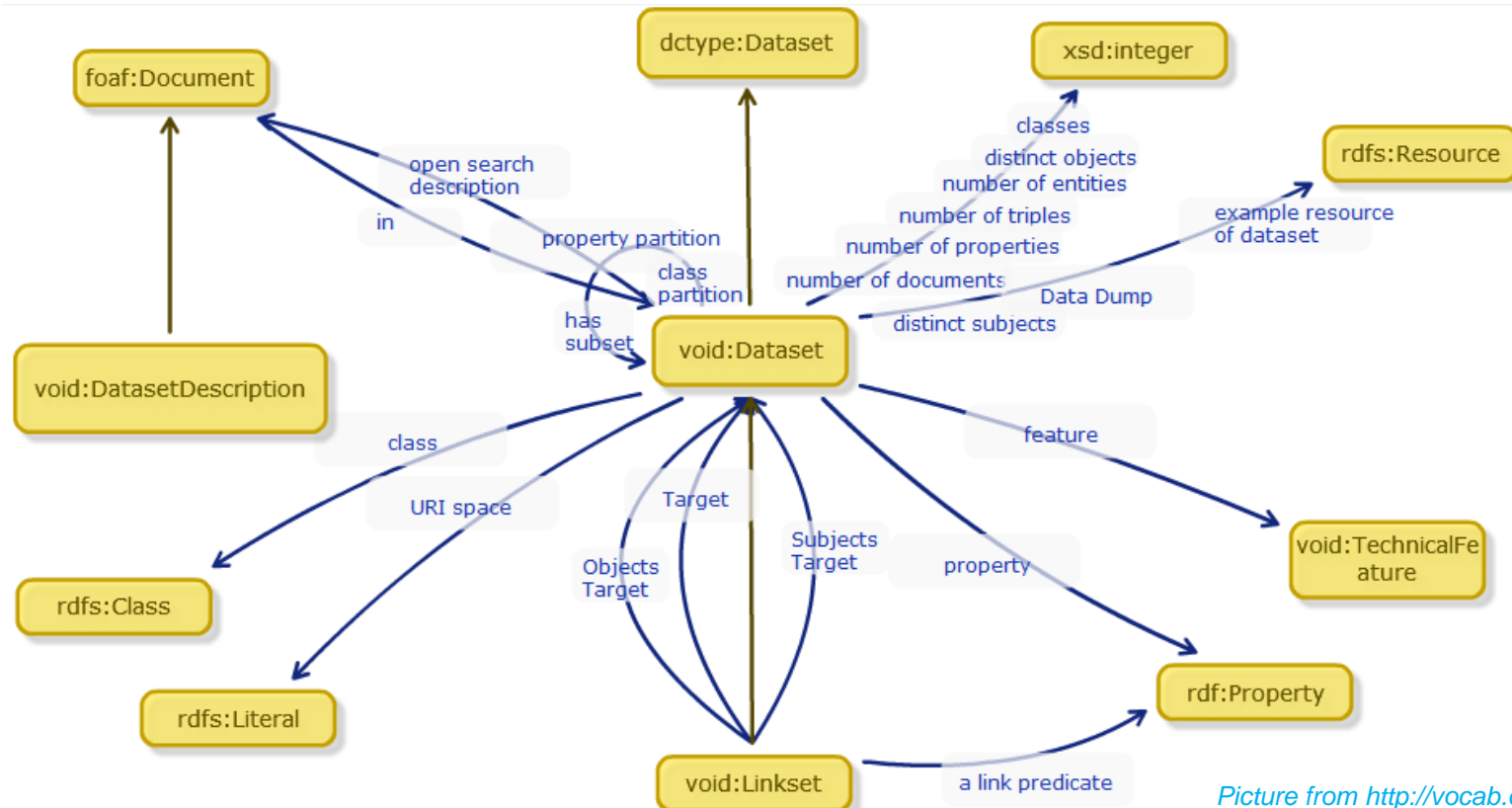
- LIMES** (Link discovery framework for MEtric Spaces) is a link discovery framework for the Web of Data. It implements time-efficient approaches for large-scale link discovery based on the characteristics of metric spaces. LIMES applies different approximation techniques to compute estimates of the similarity between instances. It is easily configurable via a web interface as well as can be user as standalone tool locally. (<http://aksw.org/Projects/LIMES.html>)

The screenshot displays the LIMES web interface, which is divided into several sections:

- Examples (toggle):** Drugbank, Vacations, Duplicate Cities.
- Download:** Manual | Distribution.
- Source and Target Configuration:**
 - Source:** Endpoint: <http://dbpedia.org/sparql>, Graph: -1, Var: ?x, Pagesize: 1000, Restriction: ?x rdf:type dbpedia-o:City, Property: dbpedia-o:populationTotal.
 - Target:** Endpoint: <http://dbpedia.org/sparql>, Graph: -1, Var: ?y, Pagesize: 1000, Restriction: ?y rdf:type dbpedia-o:City, Property: dbpedia-o:populationTotal.
- Metric:** AND(euclidean(x.dbpedia-o:populationTotal,y.dbpedia-o:populationTotal)).
- Output:** N3.
- Execution:** Linear.
- Acceptance:** Threshold: 1, Relation: owl:sameAs.
- Review:** Threshold: 0.9, Relation: owl:sameAs.
- Detected prefixes:**
 - rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> (-)
 - dbpedia-o: <http://www.w3.org/2000/01/rdf-schema#> (-)
 - owl: <http://www.w3.org/2002/07/owl#> (-)
 - dc-terms: <http://purl.org/dc/terms/> (-)
 - dbpedia-o: <http://dbpedia.org/ontology/> (-)
 - dbpedia-p: <http://dbpedia.org/property/> (-)
- Machine Learning Layer:**
 - Input Layer:** CSV, RDF(N3), SPARQL.
 - Preprocessing Layer:** Regular Expressions, String Cleaning, Conversion.
 - Data Layer:** Atomic Mappers, Mappers, Filters.
 - Processing Layer:** Tab-Separated, RDF(N3).
 - Output Layer:** Tab-Separated, RDF(N3).
- Workflow Diagram:**
 - Source Properties:** x.rdf:s.label, y.name.
 - Target Properties:** name.
 - Measures:** cosine, euclidean, exactmatch, geo_avg, geo_centroid_indexed_hausdorff_distance.
 - Operators:** add, and, diff, max, min, minus.
 - Flow:** Source and Target properties are processed by Jaccard and Trigram operators. These are combined using an 'and' operator with thresholds (jaccard threshold: 0.5, trigram threshold: 0.7). The result is then processed by an 'output' operator with an acceptance threshold of 0.8 and a verification threshold of 0.5.

VoID

- **VoID** - *Vocabulary of Interlinked Datasets* - is an RDF Schema vocabulary for expressing metadata about linked datasets.
- Documentation: <http://www.w3.org/TR/void/>
- Vocabulary: <http://vocab.deri.ie/void>



Picture from <http://vocab.deri.ie/void>

VoID: Datasets

- Definition: *void:Dataset*
- *Dataset is a collection of data which is:*
 - *published and maintained by a single provider*
 - *available as RDF*
 - *accessible, for example, through HTTP URIs or a SPARQL endpoint.*

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix void: <http://rdfs.org/ns/void#> .

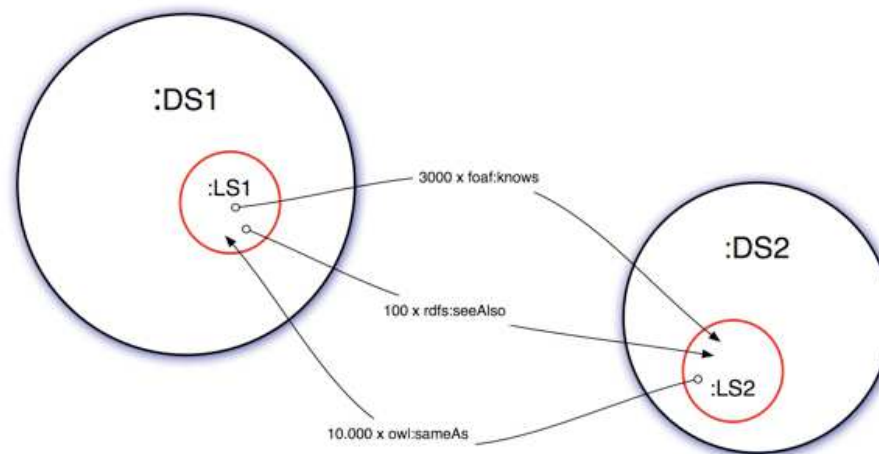
:DBpedia rdf:type void:Dataset ;
          foaf:homepage <http://dbpedia.org/> .
:DBLP rdf:type void:Dataset ;
       foaf:homepage <http://www4.wiwiwiss.fu-berlin.de/dblp/all> ;
       dcterms:subject <http://dbpedia.org/resource/Computer_science> ;
       dcterms:subject <http://dbpedia.org/resource/Journal> ;
       dcterms:subject <http://dbpedia.org/resource/Proceedings> .
```

VoID: Linksets

■ Definition: *void:Linkset*

- Is a subclass of *void:Dataset* , used for storing triples to express the interlinking relationship (e.g. owl:sameAs or foaf:knows) between datasets
- In each interlinking triple, the subject is a resource hosted in one dataset and the object is a resource hosted in another dataset

```
:DBpedia void:subset :DBpedia2DBLP .  
  
:DBpedia2DBLP rdf:type void:Linkset ;  
               void:linkPredicate owl:sameAs ;  
               void:target :DBpedia ;  
               void:target :DBLP .
```



Picture from <http://semanticweb.org/wiki/VoID>

VoID: SPARQL endpoints

- SPARQL endpoints: *void:sparqlEndpoint*

```
@prefix void: <http://rdfs.org/ns/void#> .  
  
:DBpedia a void:Dataset;  
    void:sparqlEndpoint <http://dbpedia.org/sparql> .
```

- SPARQL query for available *SPARQL endpoints*:

```
PREFIX void: <http://rdfs.org/ns/void#>  
  
SELECT DISTINCT ?endpoint  
WHERE {  
    ?ds a void:Dataset .  
    ?ds void:sparqlEndpoint ?endpoint  
}
```

- See: <http://void.rkbexplorer.com/sparql/>

VoID: URI lookup endpoints

- URI lookup endpoints: *void:uriLookupEndpoint*

```
@prefix void: <http://rdfs.org/ns/void#> .
```

```
:Sindice a void:Dataset;
```

```
void:uriLookupEndpoint <http://api.sindice.com/v2/search?qt=term&q=>.
```

- *Endpoint Lookup Service* allows a URI(s) to be submitted, and returns SPARQL endpoint(s) which may serve information about the requested resource.
 - See: <http://void.rkbexplorer.com/endpoint-search/>

VoID: Technical features

- ***void:feature*** property can be used for expressing certain technical features of a dataset (e.g. supported RDF serialization formats). The domain of the property is *void:Dataset* and its range is *void:TechnicalFeature*.

```
:DBpedia a void:Dataset;  
  void:feature <http://www.w3.org/ns/formats/RDF_XML> .
```

W3C URIs for formats are instances of class *http://www.w3.org/ns/formats/vocab-data/Format*, which is a sub-class of ***void:TechnicalFeature***.

- Customized definition of technical feature, e.g. HTTP features such as content negotiation or ETag headers...

```
:HTTPCachingETags a void:TechnicalFeature;  
  rdfs:label "HTTP ETag support";  
  rdfs:comment "the dataset supports HTTP caching using ETags";  
  rdfs:seeAlso <http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html#> .
```

VoID: Distributed location

- If an RDF dump of the dataset is available, then its location can be announced using `void:dataDump`. If the dataset is split into multiple dumps, then several values of this property can be provided.

```
:NYTimes a void:Dataset;
  void:dataDump <http://data.nytimes.com/people.rdf>;
  void:dataDump <http://data.nytimes.com/organizations.rdf>;
  void:dataDump <http://data.nytimes.com/locations.rdf>;
  void:dataDump <http://data.nytimes.com/descriptors.rdf> .
```

- `void:subset` property can be used to provide descriptions of parts of a dataset. A part of a dataset is itself a *void:Dataset*.

```
:DBpedia a void:Dataset;
  void:subset :DBpedia_shortabstracts;
  void:subset :DBpedia_infoboxes .
:DBpedia_shortabstracts a void:Dataset;
  dcterms:title "DBpedia Short Abstracts";
  dcterms:description "Short Abstracts of Wikipedia Articles";
  void:dataDump <http://downloads.dbpedia.org/3.3/en/shortabstract_en.nt.bz2> .
:DBpedia_infoboxes a void:Dataset;
  dcterms:title "DBpedia Infoboxes";
  dcterms:description "Information that has been extracted from Wikipedia infoboxes.";
  void:dataDump <http://downloads.dbpedia.org/3.3/en/infobox_en.nt.bz2> .
```

VoID: voiD Store

■ *void Store* (<http://void.rkbexplorer.com>)

- simply gathers a number of voiD documents and stores them in a repository
- makes it easy for clients and applications to query these descriptions in order to identify which datasets may be of relevance for a particular need or request
- makes it possible to find endpoints which may contain a given URI

■ Service contains:

- voiD vocabulary (<http://vocab.deri.ie/void>)
- URI to endpoint lookup
- SPARQL query engine
- voiD Editor – *ve2*
- etc.

■ *ve2* (<http://lab.linkeddata.deri.ie/ve2/>) allows to:

- generate a voiD file in RDF Turtle format and define the characteristics of your linked dataset (categories, interlinking, technical features, licensing, etc.)
- announce it to the wide world

ve² - the voiD editor

Create | Inspect | Announce

Input: dataset characteristics

Define general dataset metadata

Dataset URI

Dataset Homepage URI

Dataset Name

Dataset Description

Example Resource

Select topics

Interlink to other datasets

List used vocabularies

Tell about provenance and licensing

Specify access methods (SPARQL endpoint, etc.)

Output: void Description

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dcat: <http://purl.org/dc/terms/> .
@prefix void: <http://rdf.org/ns/void#> .
@prefix : <#> .
```