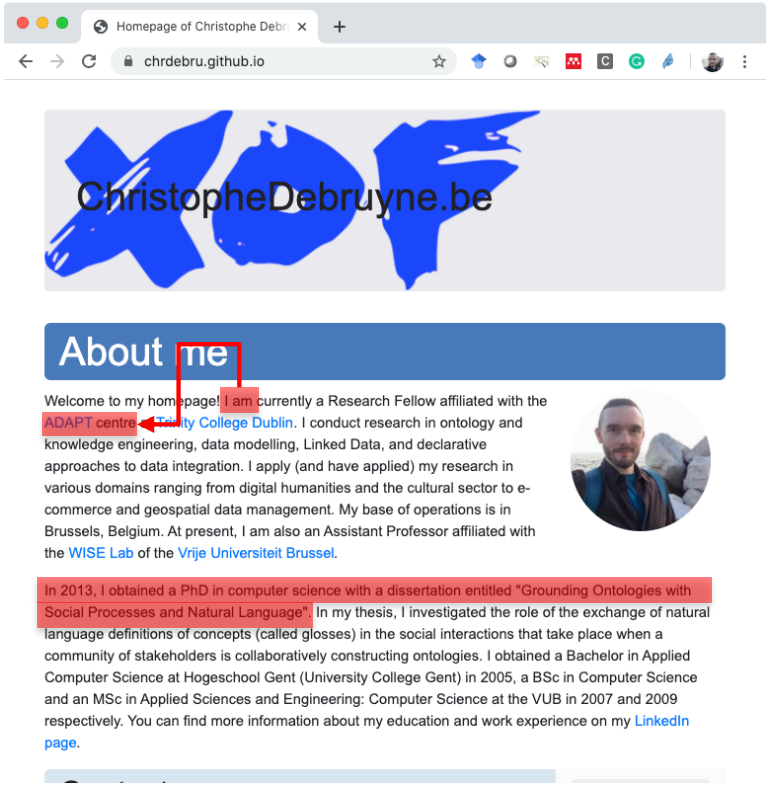
Open Information Systems
2019-2020

Lecture 6: Linked Data

Christophe Debruyne

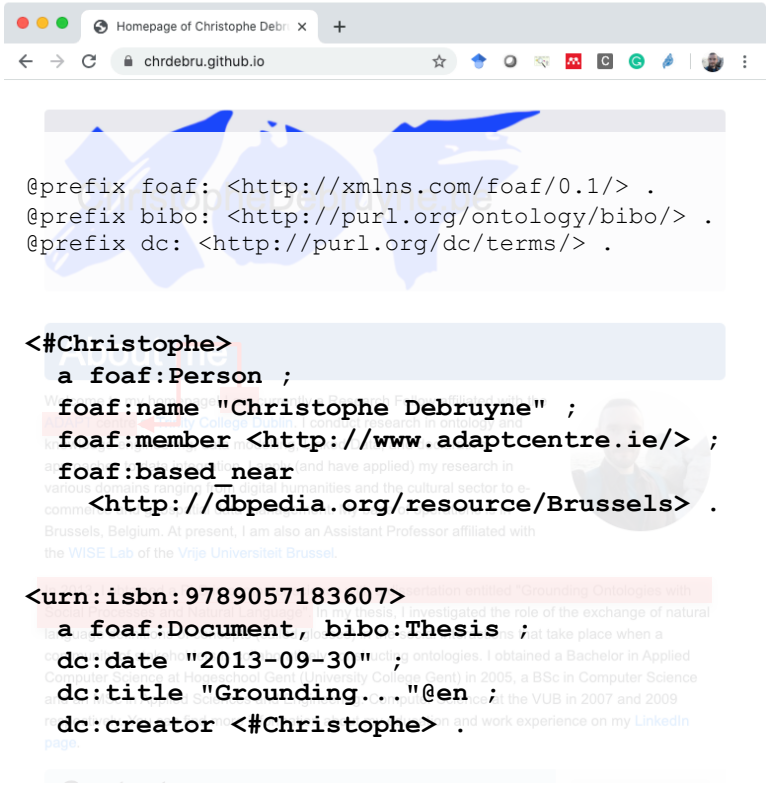
The Semantic Web: Motivation

- The Web was initially conceived to relate (pieces) of documents–a Web of Documents.
- We as humans can interpret the usually implicit relationships denoted by a hyperlinks, and facts contained in the contents.
- This is a difficult task for machines.



The Semantic Web

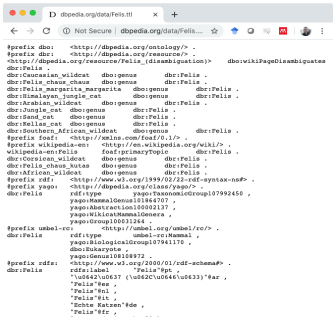
- How can we make information contained in these documents more meaningful for both humans and computers?
- "The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in co-operation." [BLHL01]
- Instead of relating (pieces) of documents, we will be relating data contained in documents with explicit meaning (i.e., semantics) – a Web of Data
- Those relations can be embedded or provided in different types of document depending on the request/client (cf., *content negotiation*).



Ontologies

An ontology is “a [formal,] explicit specification of a [shared] conceptualization" [Gru95] and extended by [Stu98]

- Explicit → externalized in a document to be shared and used by agents
- Formal → a mathematical or logic foundation to allow reasoning



$\exists\ \forall x(Cat(x) \rightarrow Animal(x))$

- Shared → for it to be meaningful
  - "Danger Noodle" vs. "Snake"

What is Linked Data?

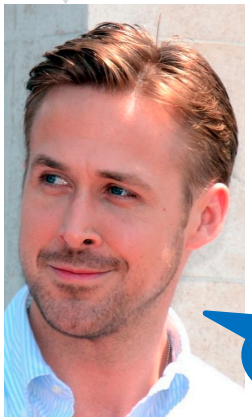
- Linked Data started off as an initiative called the Linking Open Data (LOD) project.

- Linked Data is a global initiative to publish and interlink structured data on the Web using a clever combination of simple, standardized technologies.
  - Uniform Resource Identifiers – to name things;
  - Resource Description Framework – to represent things;
  - HTTP infrastructure – to obtain those representations.

The gist with an example: http://www.DBpedia.org/

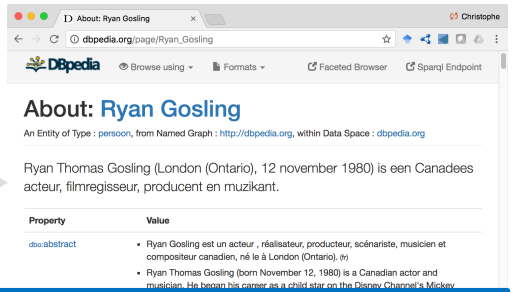
http://dbpedia.org/resource/Ryan_Gosling

refers to



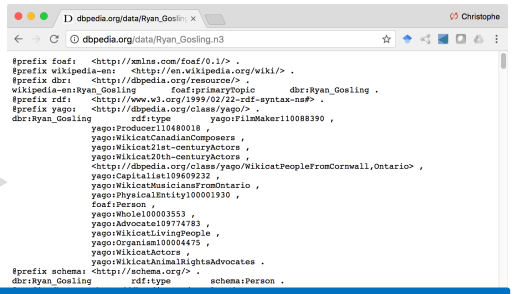
Source: Georges Biard [CC BY-SA 3.0], via Wikimedia Commons

described by



http://dbpedia.org/page/Ryan_Gosling

described by



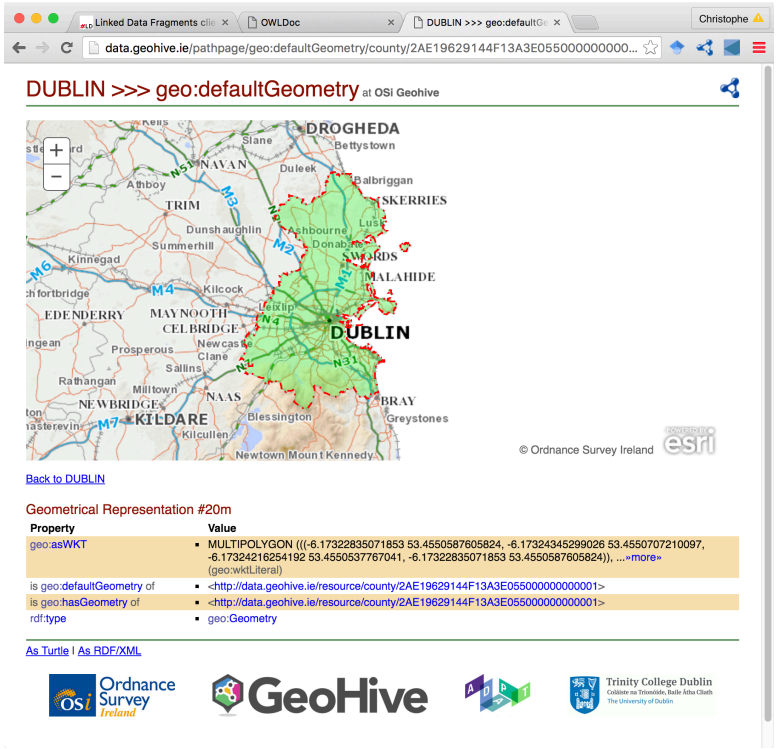
described by

http://dbpedia.org/data/Ryan_Gosling.n3

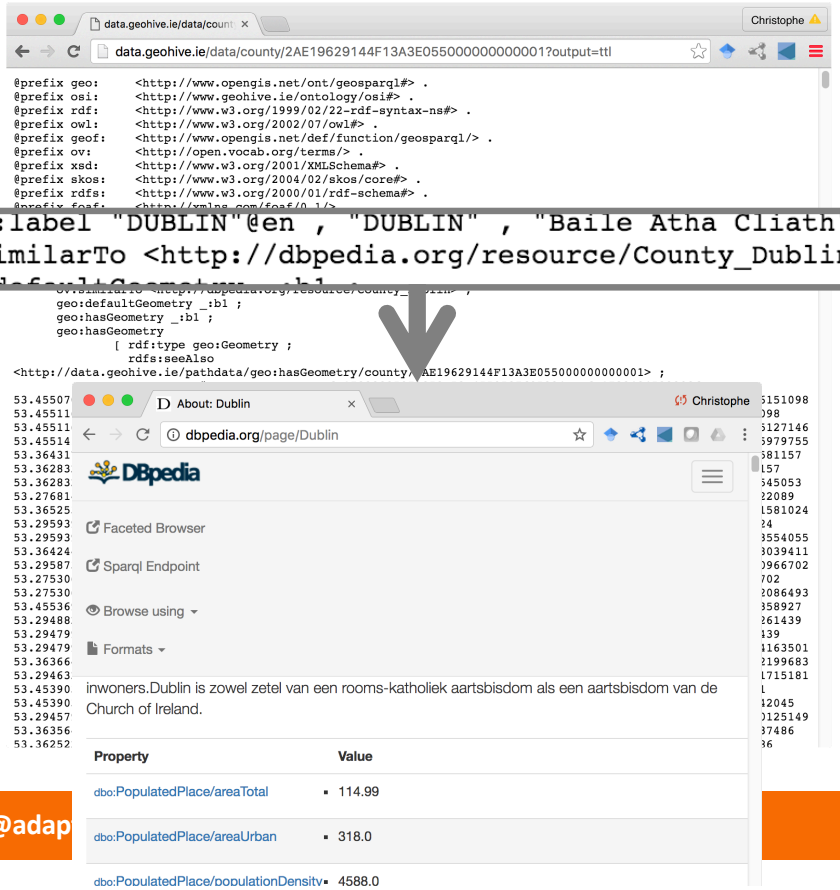
The gist with a 2nd example: data.geohive.ie

Following http://data.geohive.ie/page/county/2ae19629-144f-13a3-e055-000000000001

Redirect to /page/county



Redirect to /data/county



Web of Documents vs. Web of Data

- The Web of Documents were created by humans for humans; the links between documents bore little meaning for machines and documents provided little structured information.

- Structured information can be found on the Web
  such as XML, CSV, etc. – but, …

- How do we link data rather than documents, and create a global "database" of information?

Web of Documents vs. Web of Data

- Semantic Web is not only about data, but about making links between data instead of links between documents.
- Enabling persons and machines to explore the Web of Data.
- Links between arbitrary “things” in data are described in RDF.
- Connect distributed data across the Web → http://linkeddata.org/

Web of Documents vs. Web of Data

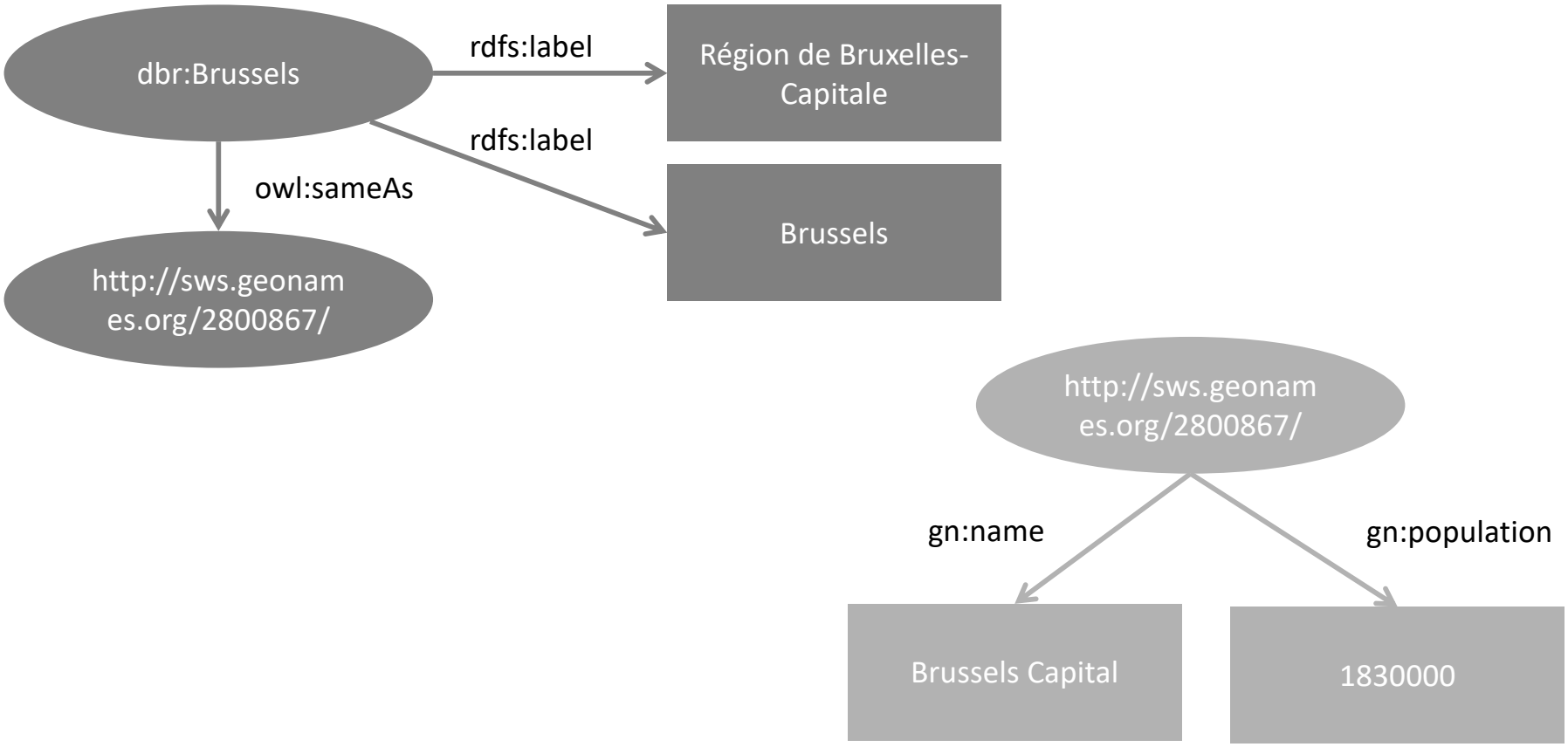
	Web of Documents	Web of Data
Analogy	Global file system	Global database
Primary objects	Documents	(Descriptions of) Things
Links between	(Parts of) Documents	Things
Degree of structure	Low	High
Semantics between links and content	Implicit	Explicit
Designed for	Human consumption	Both human and computer-based agents

Compiled from http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf

Linked Data

- Linked Data is also a community effort to publish *(open)* data sets as Linked Data on the Web (to which anyone can refer to)
- According to some “protocol” and ...
- Interlink these data sets and ...
- Develop clients that consume Linked Data from the Web

Example of linking across datasets



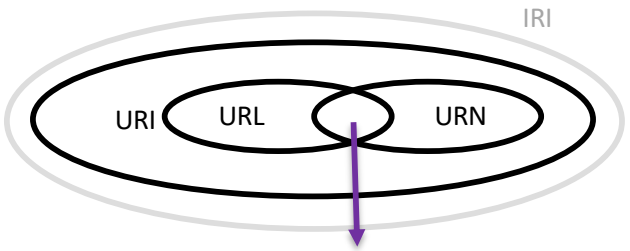
Towards a Web of Data

- We need appropriate methods (guidelines) and standards.

- Tim Berners-Lee formulated four rules for creating and publishing Linked Data on the Web.
  - Use URIs as names for things.
  - Use HTTP URIs so that people can look up those names.
  - When someone looks up a URI, provide useful information using the standards.
  - Include links to other URIs, so that they can discover more things.

Use URIs to “name” things

- Use Uniform Resource Identifiers (URIs) to name anything you *need* to describe on the Web
  - People, geographical locations, books, …
  - Events, emotion, religion, …



- Examples
  - http://dbpedia.org/resource/James_Joyce — URL (resource + location; contains name)
  - http://example.org/index.html#christophe — URL (resource + location; contains name)
  - ftp://example.org/file.txt — URL (resource + location; contains name)
  - urn:ISSN:1535-3613 — URN (resource + name; not availability)
  - data:,Christophe — URI (data)

- There are best practices and guidelines for URIs (cool vs. readable, cool vs. opaque, fragments, …), but this is out of today's session's scope.

- IRIs (International Resource Identifiers) extend URIs with non-Latin characters.

Different types of URIs

- Cool URIs do not change over time
  http://foo.bar/people.php?first=christophe&last=debruyne

- Avoid the inclusion of specific technologies and variables in your URIs.
  - Creation dates are OK…
  - But leave out information such as subject, author, status, file name extensions, etc.

http://www.w3.org/Provider/Style/URI.html

Different types of URIs

- `Foo`, `Foo.png` and `Foo.gif` are each valid resources on the Web.
- `Foo` is content-type-generic
- `Foo.png` and `Foo.gif` are content-type-specific.

http://www.w3.org/Provider/Style/URI.html

Different types of URIs

- **Cool, Readable, and Opaque**
  - http://kdeg.scss.tcd.ie/people.php?f=christophe&l=debruyne
  - http://tcd.ie/person/christophe-debruyne
  - http://tcd.ie/person/0116296s

- Readable URIs allow a "follow your nose approach" and makes sense to humans, the latter allows for the most flexibility. Opaque URIs are not affected by changes.

- Still widely debated and provides an interesting tension field.

#2) Use HTTP URIs so that people can look up those names.

- HTTP URIs allow one to reuse the existing HTTP infrastructure to return *something* when one performs an HTTP GET request.

- One can for instance put the HTTP URI in a browser's address bar and – *hopefully* – get a result.

- http://dbpedia.org/resource/James_Joyce

#3) When someone looks up a URI, provide useful Information for URI look-ups

- When entities are identified by URIs that use the `http://` scheme, these entities can be looked up simply by dereferencing the URI over the HTTP protocol.

- Simple, standardized mechanism for retrieving resources via these URIs.

- Provide information suitable for the "consumer"
  - Suitable for browsers vs. suitable for machines
  - Humans rather see HTML pages, PDFs, pictures, …
  - Machines want machine-readable formats such as RDF

- RDFa (RDF annotation) is an HTML document with embedded RDF

(Non-)Information Resources

- Information resources are documents – referred to by a URI – that describe non-information resources – named with a URI – that represent things such as cars, people, etc.

- The NIR http://dbpedia.org/resource/James_Joyce is described by the following IRs:
  - The web page http://dbpedia.org/page/James_Joyce
  - The RDF doc http://dbpedia.org/data/James_Joyce

- Either is returned depending on what you need. How?

Content Negotiation

Resource identifiers:

- HTTP URIs not only as a name, but also for a Web look-up.
- Non-information resources can have multiple representations: HTML, RDF/XML, ...

HTTP URI dereferencing:

- To dereference → "To obtain the address of a data item held in another location from a pointer"
- URI pointing to a IR returns the representation.
- URI pointing to a NIR returns a redirect to an IR describing that NIR.

Content Negotiation

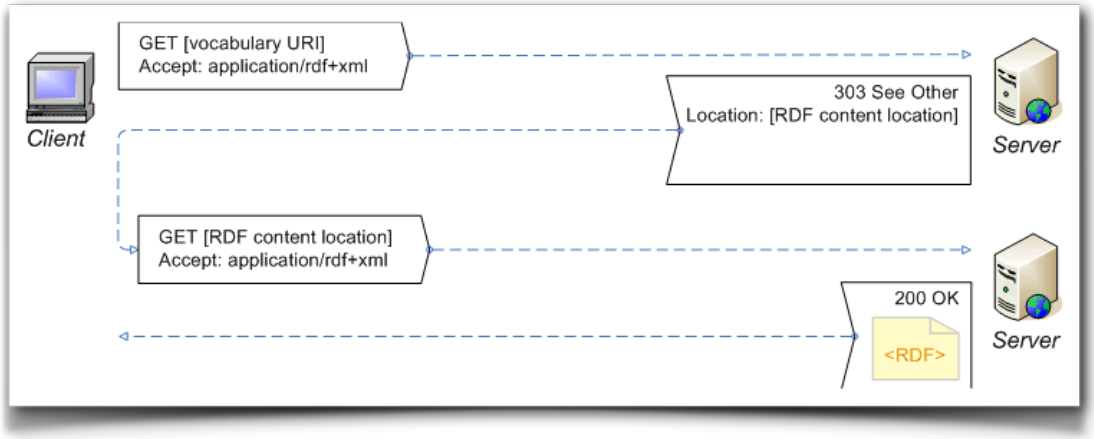


Image from http://www.w3.org/TR/swbp-vocab-pub/

What should be returned?

- **RDF should be at least be represented as RDF/XML.**
- All RDF triples with the NIR's URI as the subject in the triples. Triples where the NIR is a object are optional, but nice to have.

Which SPARQL queries can be used to obtain that information?

- DESCRIBE <*x*>
- SELECT DISTINCT * WHERE {<*x*> ?p ?o} UNION {?s ?p <*x*>}
- …

- Descriptions about related resources and metadata (e.g. publisher, creation date, etc.) should be attached to the information resource.

Basic principles

- HTTP URI dereferencing:
  - cURL is a command line tool used for transferring files

```
$ curl -L -H "Accept: text/html" –i
http://dbpedia.org/resource/Brussels
```

`-L` for accepting all redirects

`-H` "Accept: application/rdf+xml" is used to add extra header information

`-i` Include the HTTP-header in the output

```
$ curl -L -H "Accept: text/html" -i http://dbpedia.org/resource/Dublin
HTTP/1.1 303 See Other
Date: Tue, 02 May 2017 15:23:33 GMT
Content-Type: text/html; charset=UTF-8
Content-Length: 0
Connection: keep-alive
Server: Virtuoso/07.20.3217 (Linux) i686-generic-linux-glibc212-64  VDB
Location: http://dbpedia.org/page/Dublin
Expires: Tue, 09 May 2017 15:23:33 GMT
Cache-Control: max-age=604800
Access-Control-Allow-Origin: *
Access-Control-Allow-Credentials: true
Access-Control-Allow-Methods: GET, POST, OPTIONS
Access-Control-Allow-Headers: DNT,X-CustomHeader,Keep-Alive,User-Agent,X-Requested-With,If-Modified-Since,Cache-
Control,Content-Type,Accept-Encoding

HTTP/1.1 200 OK
Date: Tue, 02 May 2017 15:23:34 GMT
Content-Type: text/html; charset=UTF-8
# OMITTED FOR BREVITY

<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" "http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
```

Requesting Information Resources

- A HTTP GET request on a non-information resource with URI *x* should resolve to an information resource describing that NIR, which *can* have a different URI *y*.

- What if we HTTP GET *y*?
- We get the document referred to with URI *y*.
- But, what is the topic/subject/… of *y*? In other words, that is the resource being described?
- One common practice → include a **`foaf:primaryTopic`** statement.

Using URI Fragments

- Another approach is to use **URI fragments**
- http://foo/#bar resolves and the document referred to by http://foo/ describes the thing with id “bar”
- Clients will remove the fragment to request the document URI.

Content Negotiation vs. URI fragments

- HTTP was extended from only documents to documents and things.
  - An issue that emerged in early 00's known as HttpRange-14: What is the range of the HTTP dereference function?
  - TBL's argument that HTTP URIs (without "#") should be understood as referring to documents, not things.

- After quite some discussion, the is solved with two solutions: URI fragments and content negotiation.

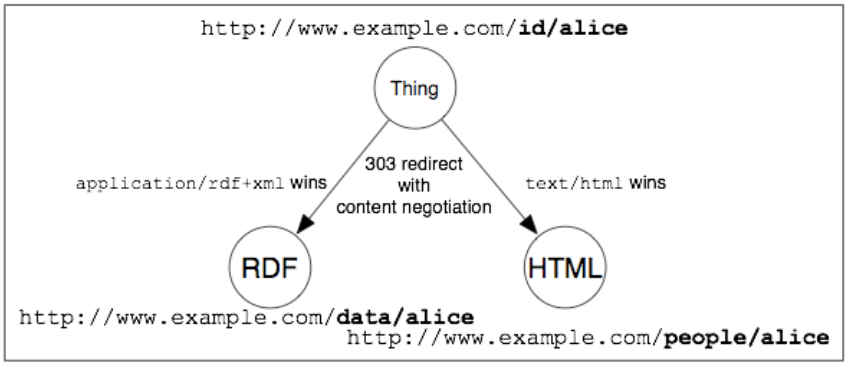
http://www.w3.org/2001/tag/group/track/issues/14

Content Negotiation vs. URI fragments

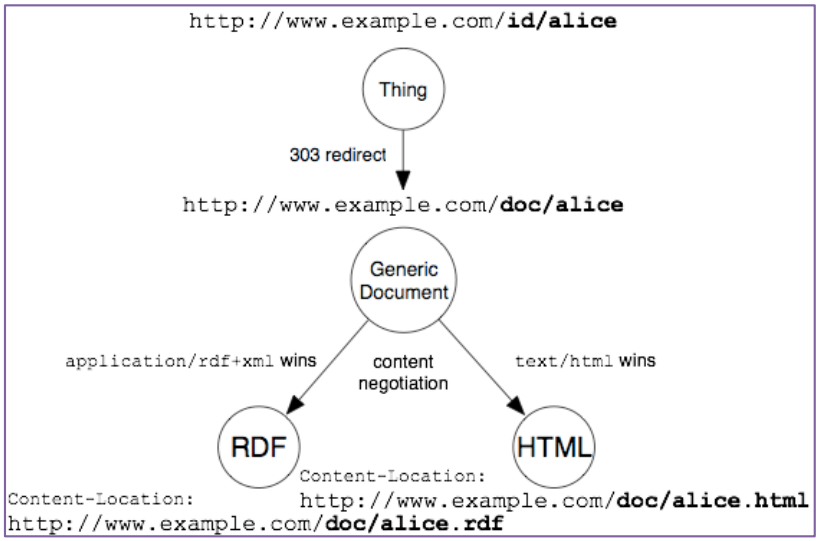
- **Content negotiation to different documents** **common for LD datasets**
  - From http://ex/resource/1 redirected to http://ex/data/1
- **Content negotiation with generic document**
  - From http://ex/resource/1 redirected to http://ex/doc/1, and then redirected to http://ex/data/1
- **URI Fragments without content negotiation**
  - From http://ex/doc#foo, the fragment is truncated and the client fetches http://ex/doc, it is up to the client to "seek" foo
- **URI Fragments with content negotiation** **common for ontologies**
  - From http://ex/doc#foo, the fragment is truncated and the client fetches http://ex/doc, the client is redirected to http://ex/data, and it is up to the client to "seek" foo

Content Negotiation vs. URI fragments

Content negotiation to different documents



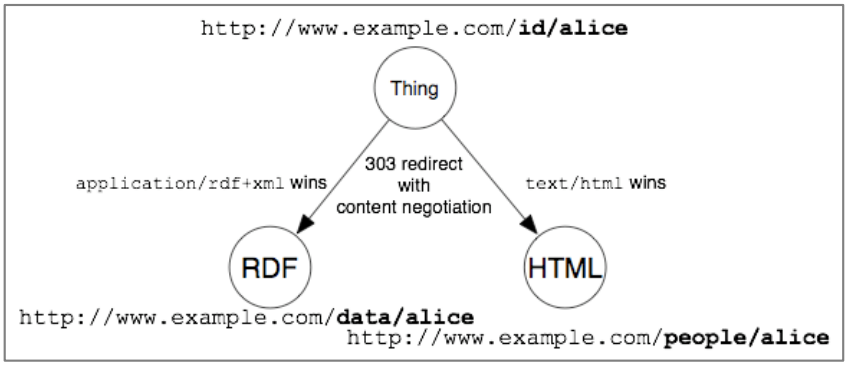
Content negotiation with generic document



http://www.w3.org/TR/cooluris/

Content Negotiation vs. URI fragments

Content negotiation to different documents



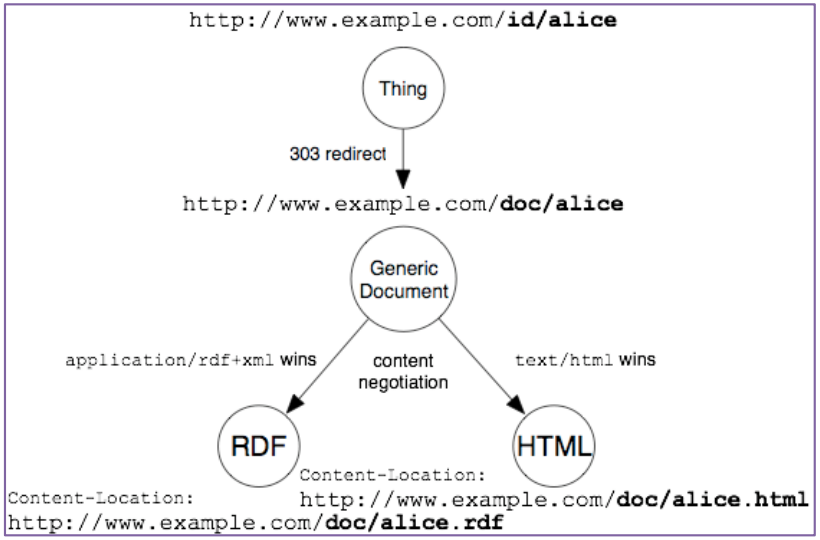
1. Redirect from thing to specific representation with content negotiation

http://www.w3.org/TR/cooluris/

Content Negotiation vs. URI fragments

1. Redirect from thing to content-generic resource
2. Then redirect to content-specific representation

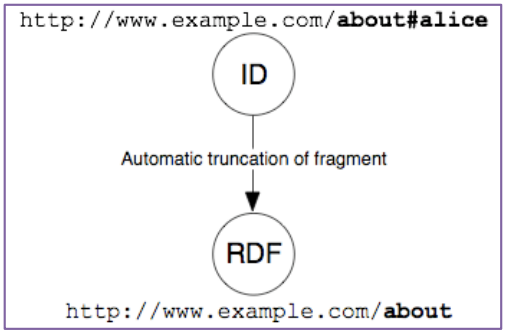
Content negotiation with generic document



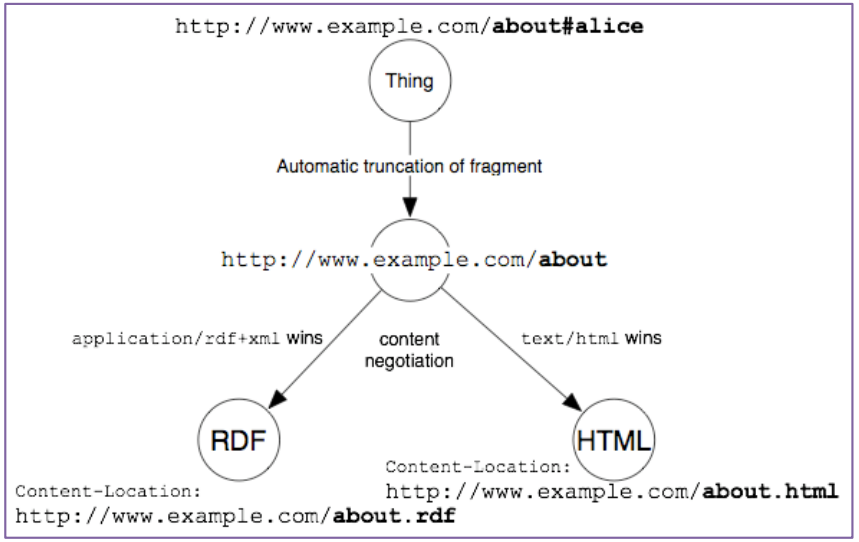
http://www.w3.org/TR/cooluris/

Content Negotiation vs. URI fragments

URI Fragments without content negotiation



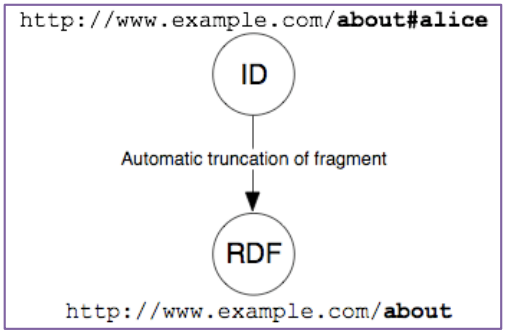
URI Fragments with content negotiation



http://www.w3.org/TR/cooluris/

Content Negotiation vs. URI fragments

URI Fragments without content negotiation



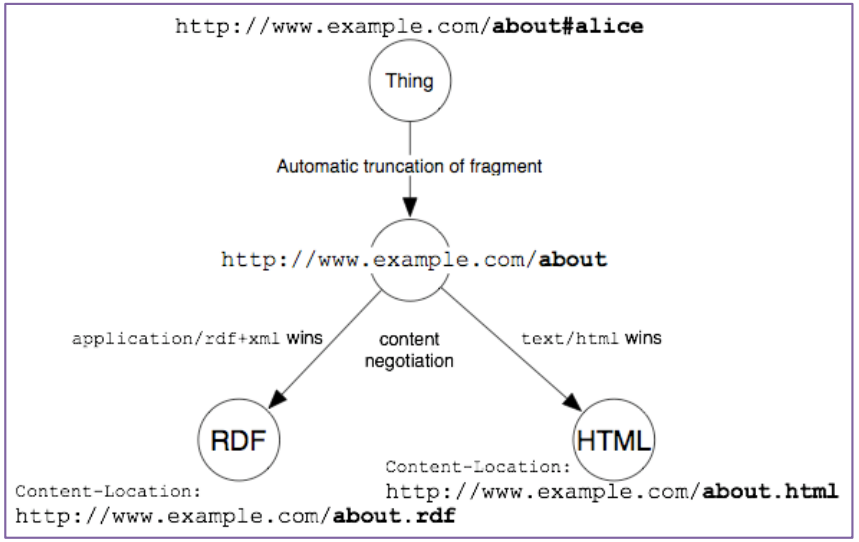
1. Client retrieves resources before fragment
2. Client looks for thing identified with fragment identifier in resource

http://www.w3.org/TR/cooluris/

Content Negotiation vs. URI fragments

URI Fragments with content negotiation

1. Client requests resource before fragment
2. Server redirects client to resource with specified content-type
3. Client fetches content-specific resource and looks for thing identified with fragment identifier



http://www.w3.org/TR/cooluris/

Choosing between content negotiation and/or URI fragments

- No consensus.
- URI fragments reduce the number of HTTP GET requests, but whole document needs to be loaded. Content negotiations is flexible and scalable and one describing document per resource possible.
- Fragments are often good for ontologies and vocabularies (because more stable). Content negotiation are often good for (large) datasets.

#4) Include links to other URIs, so that they can discover more things.

- Not only within the same dataset

```
<http://dbpedia.org/resource/James_Joyce>
dbpedia-owl:birthPlace
<http://dbpedia.org/resource/Dublin> .
```

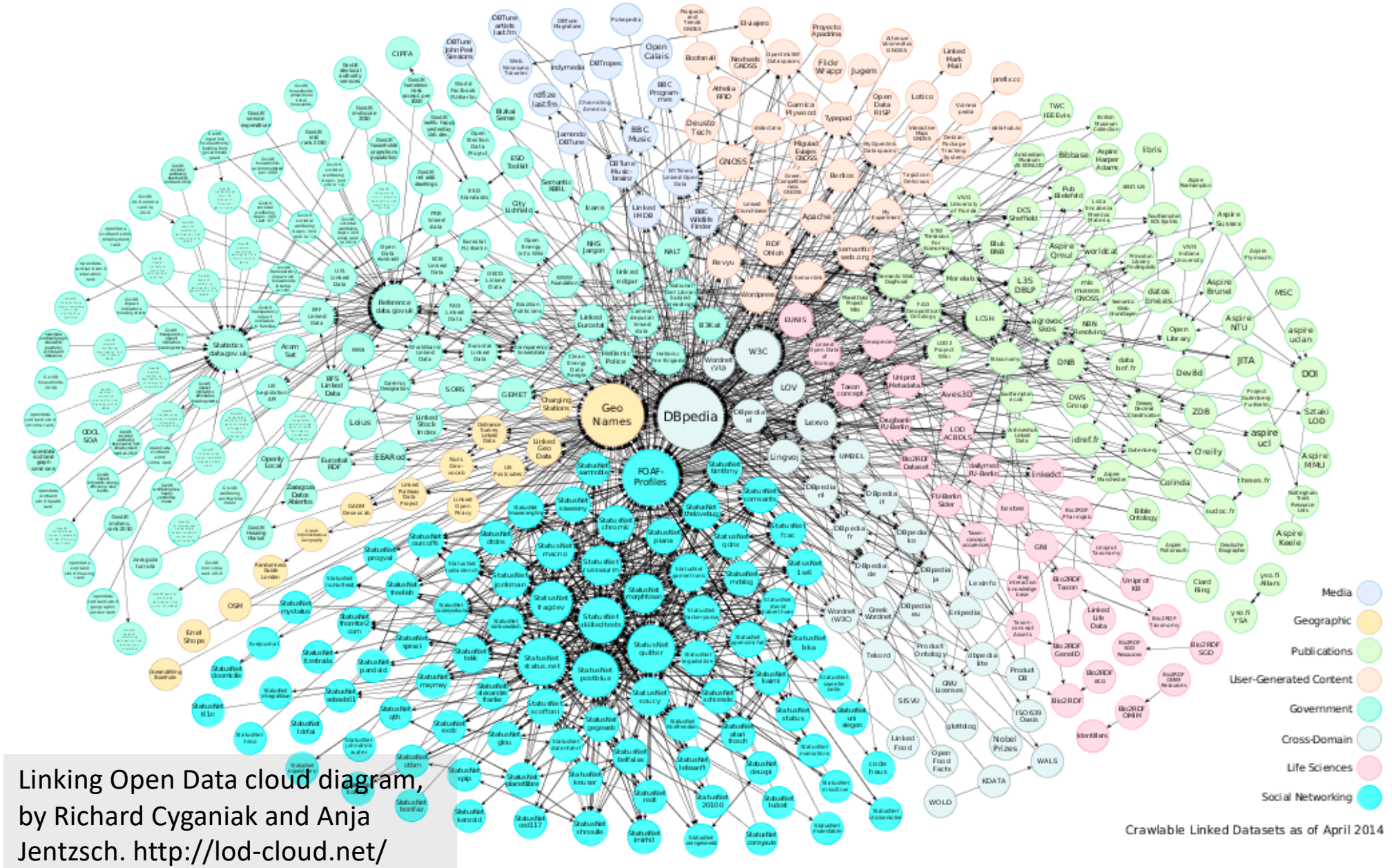
- But also across datasets

```
<http://dbpedia.org/page/Dublin>
owl:sameAs
<http://sws.geonames.org/7778677/> .
```

URI Aliases

- URI Aliases
  - Different URIs (aliases) are about the same NIR
  - Agreeing on one URI for a NIR is not realistic
  - Linking these aliases with, for instance, owl:sameAs to tackle this problem.
- Dereferencing a URI pointing to a NIR results in a IR describing that NIR.
- In Linked Data, you have not only different URI aliases that refer to the same NIR, but these aliases also dereference to different IRs!

Opinion: be careful with owl:sameAs, other predicates such as voc:similarTo might be more appropriate.



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

LOD Cloud 2014 Statistics

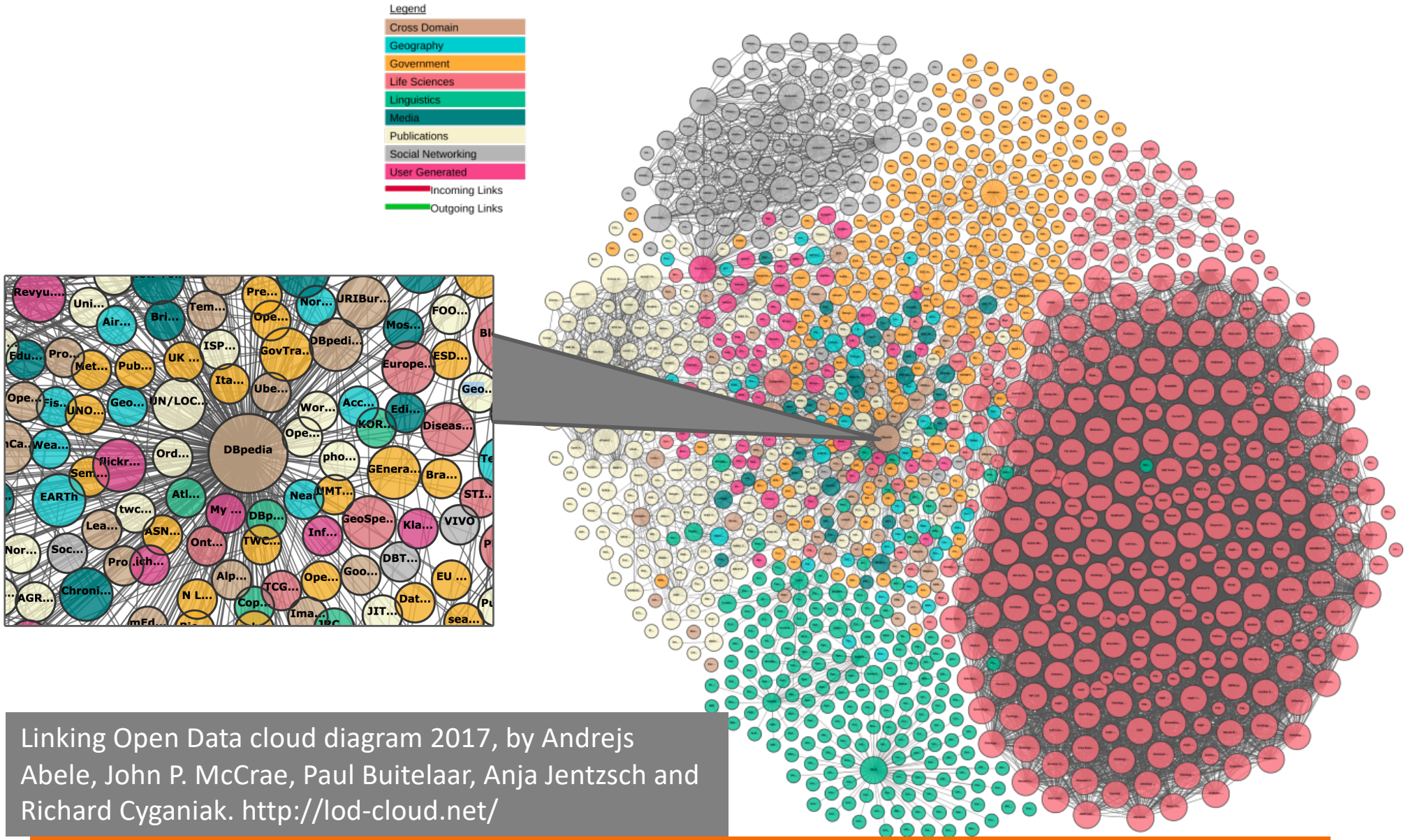
By Max Schmachtenberg, Christian Bizer, and Heiko Paulheim in the context of the EU PlanetData project.

Also provides information about vocabularies used as well as popular predicates for interlinking.

No statistics gathered for the 2017 diagram.

Datasets by topical domain.		
Topic	Datasets	%
Government	183	18.05%
Publications	96	9.47%
Life sciences	83	8.19%
User-generated content	48	4.73%
Cross-domain	41	4.04%
Media	22	2.17%
Geographic	21	2.07%
Social web	520	51.28%
Total	**1014**	

Categorization by number of linked datasets	
Number of linked datasets	Number of datasets
more than 10	79 (7.79%)
6 to 10	81 (7.99%)
5	31 (3.06%)
4	42 (4.14%)
3	54 (5.33%)
2	106 (10.45%)
1	176 (17.36%)
0	445 (43.89%)



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Result

Be applying these four principles we have, instead of relying on different interfaces and formats a ...

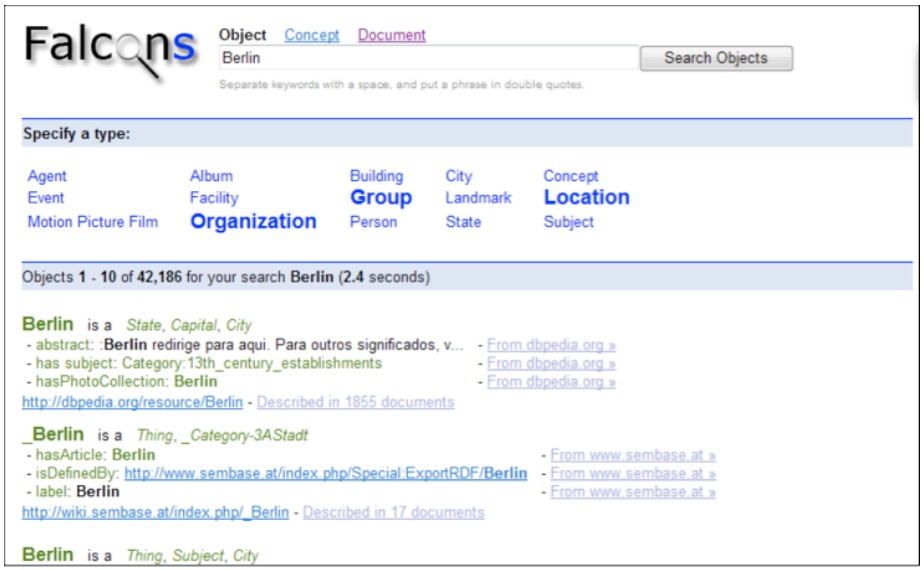
- Single, standardized access mechanism;
- That allows different data sources to be interlinked, more easily crawled, and accessed by a generic data browser.

Benefits of RDF

- Benefits of using the RDF Data Model in the Linked Data Context:
  - URI Look ups
  - Links between different resources (merging)
  - Different schemata in a single model
  - Combined use of structured and semi-structured data

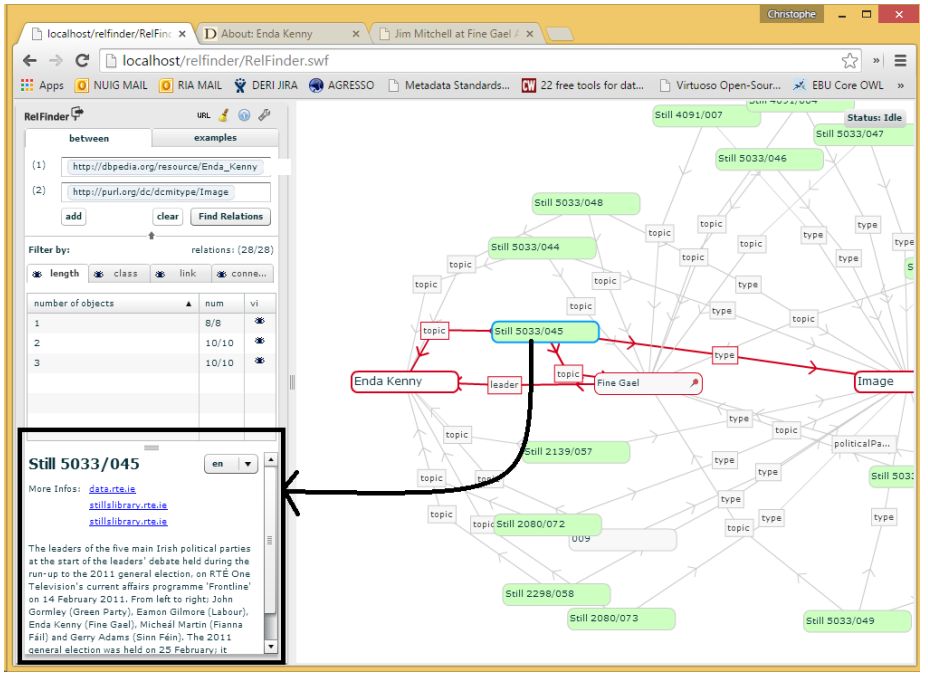
Applications

Falcons (search engine)



http://iws.seu.edu.cn/services/falcons/

RelFinder (application)



http://www.visualdataweb.org/relfinder.php

Linked Data vs. Linked Open Data

It's on the Web	
It's on the Web with an under an open license (e.g., PDF)	
It is available as some structured data that machines can process (e.g., Excel)	
Using open (i.e., non-propriety) formats such as CSV, XML,…	
And you use URI and URLs for identifying and locating things (RDF)	
Create links across datasets	

Linked Data is about the best practices and guidelines for humans and computer-based agents to engage with data on the Web. Linked Open Data is about best practices for publishing Open Data on the Web.

There is a such thing as Linked "Closed" Data.

References

- Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (5) (2001) pages 35–43
- Berners-Lee. Design Issues: Linked Data http://www.w3.org/DesignIssues/LinkedData.html
- Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. Int. J. Semantic Web Inf. Syst. 5(3), 1–22 (2009)
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2): 167-195 (2015)
- HttpRange-14 http://lists.w3.org/Archives/Public/www-tag/2005Jun/0039.html