

Linked Open Data

Sam Coppens
MMLab - IBBT - UGent

- Overview:
 - Linked Open Data: Principles
 - Interlinking Data
 - LOD Server Tools

Linked Open Data: Principles

Term “*Linked Data*” was first coined by Tim Berners Lee in his note on Linked Data Web Architecture in 2006.

<http://www.w3.org/DesignIssues/LinkedData.html>

“Share structured data on the Web as easily as sharing documents today.”

Linked Open Data: Principles

Tim Berners Lee's note on Linked Data Web Architecture:

"Like the web of hypertext, the web of data is constructed with documents on the web."

Linked Open Data: Principles

Web of Hypertext

- Documents: HTML
- Access: HTML browser
- Links: HTML

Web of Data

- Data: RDF
- Access: RDF browser
- Links: RDF

Linked Open Data: Principles

RDF: Resource Description Framework

XML problem:

```
<author>
  <uri>page</uri>
  <name>Ora</name>
</author>

<document href="page">
  <author>Ora</author>
</document>

<document>
  <details>
    <uri>href="page"</uri>
    <author>
      <name>Ora</name>
    </author>
  </details>
</document>
```

Linked Open Data: Principles

RDF: Resource Description Framework

A description of a resource is represented as a number of *"triples"*.

Triples		
Subject	Predicate	Object
Chris	Has the email address	Chris@Bizer.de

Linked Open Data: Principles

RDF: Resource Description Framework

```
<?xml version="1.0"?>
```

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
```

```
<contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
```

```
<contact:fullName>Eric Miller</contact:fullName>
```

```
<contact:mailbox rdf:resource="mailto:em@w3.org"/>
```

```
<contact:personalTitle>Dr.</contact:personalTitle>
```

```
</contact:Person>
```

```
</rdf:RDF>
```



Linked Open Data: Principles

Tim Berners Lee's note on Linked Data Web Architecture (2):

"But for HTML or RDF, the same expectations apply to make the web grow:

- 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.*
- Include links to other URIs. so that they can discover more things.*

In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps."

Linked Open Data: Principles

Linked Data:

- a style of publishing and interlinking structured data on the Web.
- using the Web to create typed links between data from different sources.

2 Tenets:

- use the RDF data model to publish structured data on the Web
- use RDF links to interlink data from different data sources

Linked Open Data: Principles

Architecture:

- Resource: identify the items of interest
 - Information: Documents, images, and other media files
 - Non-Information: People, places, physical products, etc.
- Resource Identifiers
 - HTTP URI's (\leftrightarrow URN, DOI, etc)
- Representation
 - Information resources have bitstreams in certain format (HTML, RDF/XML, JPEG)

Linked Open Data: Principles

Architecture (2):

- Content Negotiation:

- HTML browsers: Show raw RDF code / download RDF file ☹

→ Serve HTML representation in addition of RDF representation by content negotiation.

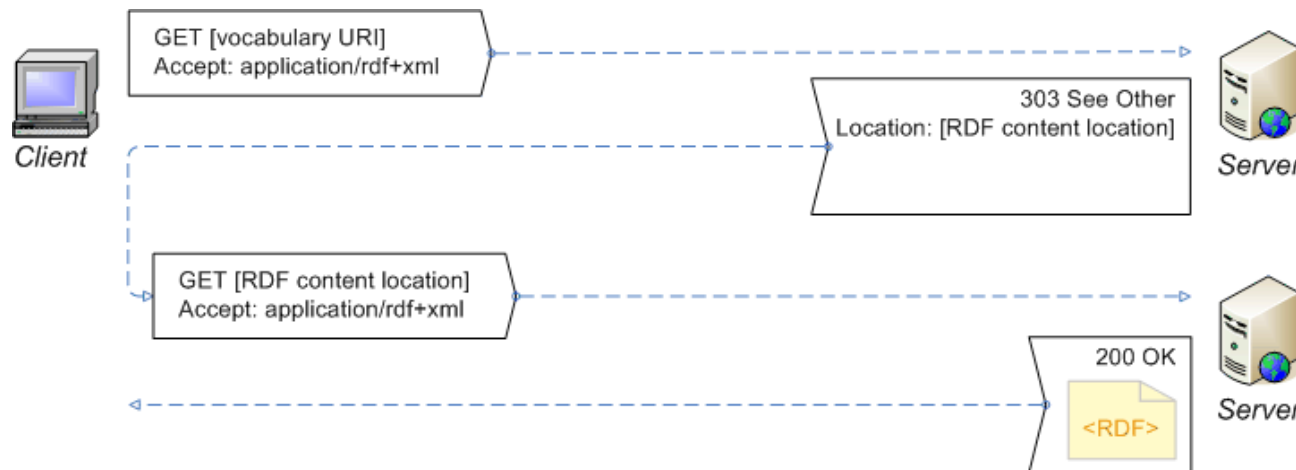
→ Each data source has 3 URIs related to the non-information resource.

- <http://www4.wiwiss.fu-berlin.de/factbook/resource/Russia>
(URI identifying the non-information resource Russia)
- <http://www4.wiwiss.fu-berlin.de/factbook/data/Russia>
(information resource with an RDF/XML representation describing Russia)
- <http://www4.wiwiss.fu-berlin.de/factbook/page/Russia>
(information resource with an HTML representation describing Russia)

Linked Open Data: Principles

Architecture (3):

- Content Negotiation:



- URI Aliases:

- Different URI's describing the same non-information resource

<http://dbpedia.org/resource/Berlin> = <http://sws.geonames.org/2950159/>

→ Link to URI alias by owl:sameAs

Linked Open Data: Principles

- Things must be identified with dereferencable HTTP URIs.
- Serve two representations of a resource: HTML and RDF.
- URIs that identify non-information resources must be set up in one of these ways:
 - *HTTP 303 redirect* to an information resource describing the non-information resource.
 - The URI for the non-information resource must be formed by taking the URI of the related information resource and appending a *fragment identifier* (e.g. #foo).
- RDF descriptions should also contain RDF links to resources provided by other data sources, so that clients can navigate the Web of Data as a whole by following RDF links.

Interlinking Data

“The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data.”

Interlinking Data

Goal:

Turning islands of data into a web of data, where all resources are linked with each other.

These links enable Linked Data browsers and crawlers to navigate between data sources and to discover additional resources.

Interlinking Data

Extending records with other useful links to other data sources.

owl:sameAs

rdfs:seeAlso

foaf:holdsOnlineAccount

sioc:User

...

Interlinking Data

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description>
    <dc:publisher>sam coppens</dc:publisher>
    <dc:subject>Berlin</dc:subject>
    <dc:description>
      Berlin is the capital city and one of sixteen states of Germany...
    </dc:description>
    <dc:coverage>Germany</dc:coverage>
    <dc:language>en</dc:language>
  </rdf:Description>
</rdf:RDF>
```

Interlinking Data

```

<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description>
    <dc:publisher>sam coppens</dc:publisher>
    <dc:subject>Berlin</dc:subject>
    <dc:description>
      Berlin is the capital city and one of sixteen states of Germany...
    </dc:description>
    <dc:coverage>Germany</dc:coverage>
    <dc:language>en</dc:language>
    <owl:sameAs>http://dbpedia.org/resource/Berlin</owl:sameAs>
    <owl:sameAs>http://sws.geonames.org/2950159/</owl:sameAs>
    <rdfs:seeAlso>http://dbpedia.org/page/Germany</rdfs:seeAlso>
    <rdfs:seeAlso> fbase:Duitsland </rdfs:seeAlso>
    <foaf:page>http://en.wikipedia.org/wiki/Berlin</foaf:page>
  </rdf:Description>
</rdf:RDF>
  
```

Interlinking Data

Interlinking:

Manually for small datasets

Automatically for big datasets

SPARQL-endpoint:

“People who were born in Berlin before 1900”

```
SELECT distinct ?name ?birth ?death ?person WHERE {  
  ?person dbpedia2:birthPlace <http://dbpedia.org/resource/Berlin> .  
  ?person dbpedia2:birth ?birth .  
  ?person foaf:name ?name .  
  ?person dbpedia2:death ?death  
  FILTER (?birth < '1900-01-01'^^xsd:date) .  
}  
ORDER BY ?name
```

Interlinking Data

DBpedia:

- ~Wikipedia
- Huge dataset
- SPARQL endpoint: <http://DBpedia.org/sparql>

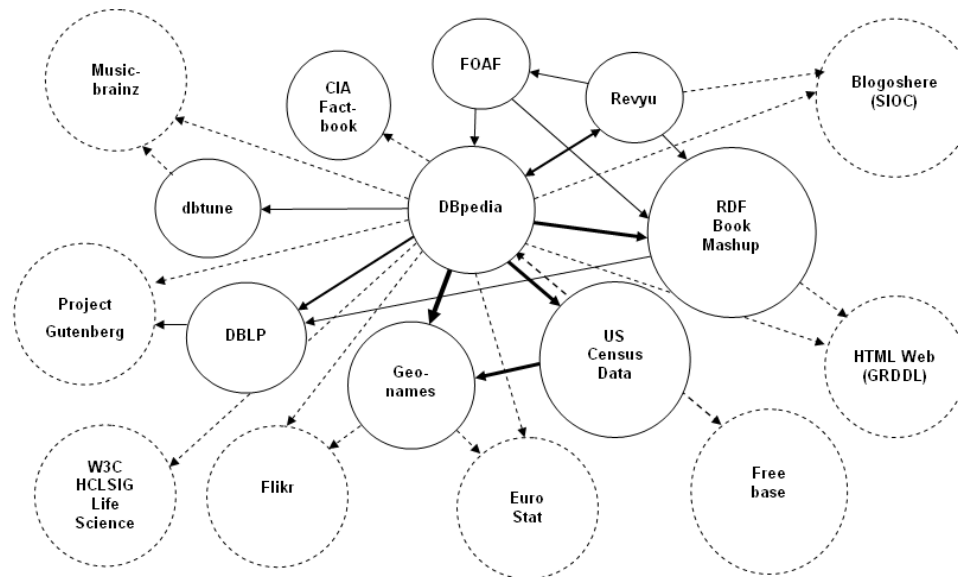
Interlinking Data

DBpedia: Interlinking Hub

Call all RDF Dumps and load them into
Virtuoso 6.0 Cluster edition.

>2billion triples

<http://esw.w3.org/topic/DataSetRDFDumps>



Interlinking Data

GeoNames:

- Location Information (- geocoordinates, map)
- Wide range of web services for querying the database

<http://www.geonames.org/export/ws-overview.html>

Interlinking Data

OpenCalais

- Concepts out plain text by natural language processing:

Anniversary, City, Company, Continent, Country, Currency, EmailAddress, EntertainmentAwardEvent, Facility, FaxNumber, Holiday, IndustryTerm, MarketIndex, MedicalCondition, MedicalTreatment, Movie, MusicAlbum, MusicGroup, NaturalDisaster, NaturalFeature, OperatingSystem, Organization, Person, PhoneNumber, Product, ProgrammingLanguage, ProvinceOrState, PublishedMedium, RadioProgram, RadioStation, Region, SportsEvent, SportsGame, SportsLeague, Technology, TVShow, TVStation, URL

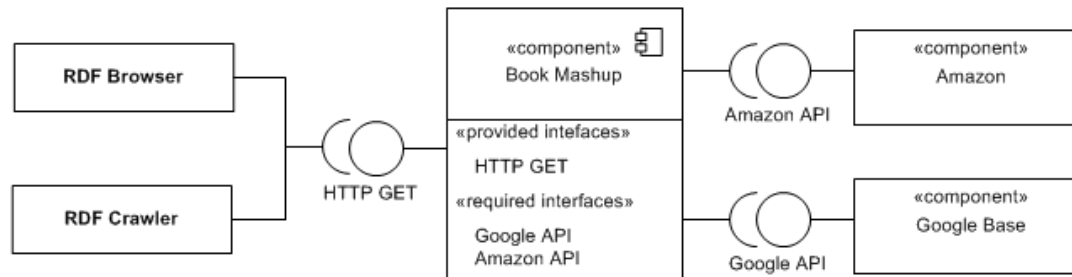
- Available by webservice

<http://www.opencalais.com/>

Interlinking Data

RDF Book Mashup

- Information on books, authors, reviews, bookstores...
- Amazon, Google base



- SPARQL endpoint

<http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/>

Interlinking Data

MusicBrainz

- user-maintained community music metadata database
- Artist title, release title, tracks, ...
- Webservice

<http://musicbrainz.org/doc/WebService>

Interlinking Data

How to start with automatic enrichment?

- Selecting the properties which will be enriched.
For example: dc.subject, dc.coverage, dc.description,
dc:creator, ...
- Selecting the datasets to enrich with for each property.
For example:
dc:subject DBpedia, RDFBookMashup
dc:coverage GeoNames
dc:description OpenCalais, DBpedia, GeoNames
- Start querying for each property the datasets for linked data.

Interlinking Data

Automatic Enrichment

Problems:

- False links due to bad spelling.

- Multiple answers to the query.

Solutions:

- Use thesauri.

- Use finer grained queries (context).

- Use other algorithms to check for other spellings.

LOD server tools

Demo LOD Server:

- OAICat
- OAI2LOD

Open Source Tools:

- Joseki
- Pubby
- D2R

LOD server tools

Demo LOD Server:

Build on two open source applications:

- OAICat
- OAI2LOD

OAICat: Web Application (Tomcat)

Builds OAI-PMH service on top of xml files or database.

OAI2LOD: Server

Builds LOD server on top of OAI-PMH service

LOD server tools

OAI-Cat

Configuration:

- Which data and where can it be found: File System, Database
- Which Format: Dublin Core (provide at least a mapping to DC)

→ OAI-PMH service <http://localhost:8080/oaicat>

<http://www.oclc.org/research/software/oai/cat.htm>

LOD Server Tools

OAI2LOD

Configuration:

- Which OAI-PMH service:
<http://localhost:8080/oaicat:OAIHandler>
- Which mapping for converting xml records in rdf records
- Which uri for the LOD server: <http://localhost:9090>

<http://www.mediaspaces.info/tools/oai2lod/>

LOD server tools

Joseki

Implements an HTTP engine that supports the SPARQL protocol and the SPARQL RDF Query language.

- RDF data from files or databases
- HTTP (Get and Post) implementation of SPARQL protocol.

<http://www.joseki.org/>

LOD server tools

Pubby

A Linked Data Frontend for SPARQL Endpoints.
Originally developed for DBpedia.

Configuration:

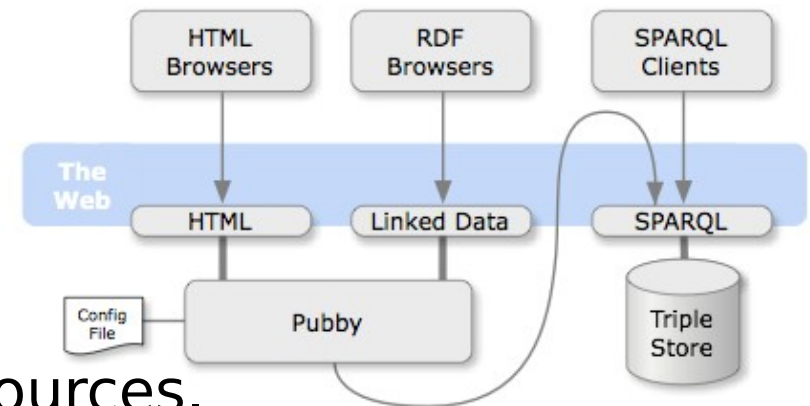
Define SPARQL endpoint.

Define URL for server.

Define mappings for the resources.

Define mappings for the classes and properties.

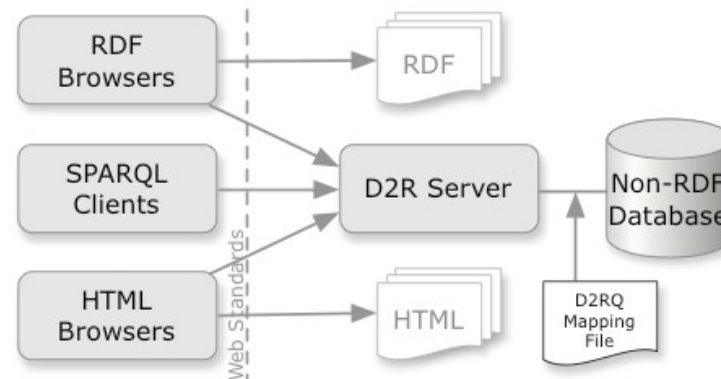
<http://www4.wiwiss.fu-berlin.de/pubby/>



LOD server tools

D2R

A tool for publishing relational databases on the Semantic Web as Linked Open Data with SPARQL endpoint.



<http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/>

Q&A