An Introduction to Linked Open Data

Felix.Ostrowski@googlemail.com (@literarymachine)
Pascal Christoph@hbz-nrw.de (@dr0ide)
Adrian Pohl@hbz-nrw.de (@acka47)

#swib12 Pre-Conference Workshop
Monday, November 26th 2012
Cologne
Schedule

- Organize in teams
- Introduction: Data – Graphs – Triples
- Groupwork
- URIs and Namespaces
- Groupwork
- Open Data
- Groupwork
- Identification vs. Description
- Content-Negotiation
- Groupwork
- Triple Stores & SPARQL
- Groupwork
- RDF Schema
- Groupwork
- Summary, Questions & Discussion
Linked Open Data

- It's about **data** ...
- ... more precisely: about **open** data ...
- ... even more precisely: about **linked** open data!
(To be honest, we might actually be the only ones knowing such data. And there aren't too many things that one can describe in this way.)
Data, how others know it

(Of course, „others“ does not mean „everybody“. But at least you can describe many things this way. Maybe even everything.)
Data, how the web likes it

(No wonder, it actually looks like a web. Or, if you will, a directed labelled graph.)
Obviously a computer will have trouble interpreting such a diagram. The **graph data model** is an abstract one, but we can concrete it for the computer.
Graphs, (almost) how computers like them

(This notation is called Turtle and it is one of several writing styles for a data model called RDF. RDF stands for „Resource Description Framework“; this is the de-facto standard for publishing Linked Data. A big advantage of the Turtle notation: humans can actually read it!)
Basic element: the **triple**

\[
\text{Weaving the Web} \quad \text{is written by} \quad \text{Tim Berners-Lee}
\]

\[
\langle \text{Weaving the Web} \rangle \langle \text{is written by} \rangle \langle \text{Tim Berners-Lee} \rangle .
\]

(A triple is the smallest possible graph. It's components are called **subject**, **predicate** and **object**.)
Your turn!
Create an etherpad for your group at http://etherpad.lobid.org/

In this pad, describe the members of your group using RDF expressed in Turtle. You can just make up the predicates for now.

(If you'd rather do so, go ahead and describe a fictional character.)
What does ... 

... <Tim Berners-Lee>,

... <London> and

... <England>

stand for, and what does

... <has first name>,

... <is located in> and

... <has population>

mean?
We need **unambiguous reference**!

Controlled vocabularies are a good start, but again we'll be the only ones understanding those. On the web, people use **URI**s!

(URI stands for **Uniform Resource Identifier**)
URI

= scheme "::" hier-part [ "?" query ] [ "#" fragment ]

(??????)
ftp://ftp.is.co.za/rfc/rfc3986.txt
file:///home/fo/doc/elag12/slides.odp
urn:isbn:978-1608454303

(URIs don't necessarily need domain names; IP-adresses also work: http://192.168.0.124)
Graphs, how computers really like them

(A pleasant side-effect when using HTTP-URIs – which is what Linked Data is based upon, is that they can be dereferenced. When following such a link, one should get a description of the resource. More on that later.)
Graphs, (sort of) readable for humans and machines

@prefix dc: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix viaf: <http://viaf.org/viaf/> .

viaf:121649091 foaf:givenName "Tim" .
viaf:121649091 foaf:familyName "Berners-Lee" .
viaf:121649091 foaf:birthday "06/08/1955" .
But isn't some data we had missing!?

(There may not be a URI for everything you want to refer to, neither for entities nor for vocabularies. When looking for URIs for entities, DBpedia is a good place to start. It's a bit tougher for vocabularies, but LOV (Linked Open Vocabularies) is worth a try.)
A couple of tricks
(When something you want to describe does not have a URI yet, you can use IDs that are relative to the describing document. Since two documents can't be at the same place at the same time, these IDs only have to be unique within that document. „<>“ stands for the document itself.)
(When making several statements about the same resource, you don't have to repeat its identifier. The predicate-object pairs are separated by a semicolon. When listing several values for the same predicate of a resource, that predicate also does not have to be repeated. Simply separate the values by comma. You can check here whether you are creating valid turtle.)
Your turn!
Within you group, agree on an identifier for each member and describe them using the FOAF vocabulary. Don't forget to state that you know each other! Also, use DC Terms to assert that you are the authors of the describing document. You can also add further metadata about the document.
Break
Open Data
„Open Data“ - Definition
Open Knowledge

“A piece of knowledge is open if you are free to use, reuse, and redistribute it.”

http://www.opendefinition.org/okd/
Open Data is a question of...

- Access
- Licenses
- Formats
Open Data is a question of...

- Access
- Licenses
- Formats
Access

➔ ...to the whole data
➔ No more than a reasonable reproduction cost
➔ Preferably downloading via the Internet without charge
➔ Technically: XMPP > OAI-PMH > http > ftp > scp > paper-based
Open Data is a question of...

- Access
- Licenses
- Formats
Open Data Licenses

- Attribution (ODC-BY)
- Attribution-Share-Alike (OdbL)
- Public-Domain (CC0, PDDL)
- CC-BY, CC-BY-SA for some uses
- **No** non-commercial licenses

http://www.opendefinition.org/licenses/#Data
Open Data is a question of...

- Access
- Licenses
- Formats
Formats

➡ Machine-readibility counts!
➡ Examples:
  rdf > csv > ods > xls > PDF >
  docx > Hardcopy
Data vs. Databases
Database

“a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means.”

From: European Database Directive
Data

- Content of a database → can be anything
- Recorded facts → aren’t copyrightable, only as collection
Different legal status?

- Legal status of database and its contents may differ
- Example: Public Domain content and copyrighted collection
Opening up data in 7 steps
1. Getting willing people together
2. Clarify potential legal problems

➔ Check your national legislation
➔ Bought data?
➔ From which vendors?
➔ What usage rights & restrictions do contracts give?
3. Export the data
4. Publish data on the web

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE-605.base_0.rdfmab.tar.bz2</td>
<td>13-Mar-2012 19:33</td>
<td>235M</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_0.rdfmab.tar.bz2.md5</td>
<td>14-Mar-2012 13:17</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_1.rdfmab.tar.bz2</td>
<td>13-Mar-2012 19:33</td>
<td>232M</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_1.rdfmab.tar.bz2.md5</td>
<td>14-Mar-2012 13:17</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_2.rdfmab.tar.bz2</td>
<td>13-Mar-2012 19:32</td>
<td>232M</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_2.rdfmab.tar.bz2.md5</td>
<td>14-Mar-2012 13:17</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_3.rdfmab.tar.bz2.md5</td>
<td>14-Mar-2012 13:17</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_4.rdfmab.tar.bz2.md5</td>
<td>14-Mar-2012 13:17</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>DE-605.base_5.rdfmab.tar.bz2</td>
<td>13-Mar-2012 19:33</td>
<td>237M</td>
<td></td>
</tr>
</tbody>
</table>
5. Apply an open license

All data which is available for download here is published under a Creative Commons CC0 licence. The data thus is in the Public Domain which means it belongs to everybody and can be used without restrictions for any purpose.

@prefix cc: <http://creativecommons.org/ns#> .

<dataset_URI> cc:license
<http://creativecommons.org/publicdomain/zero/1.0/>.
6. Register your dataset

Open Data from the hbz Union Catalog

There are daily updated library metadata dumps from the hbz Union Catalog, those records which have holdings from libraries who want their data to be open. There are more than 10 million records.

Resources (edit)

- Base dump of the Open Data from the hbz union catalog 25 🗒️ tar.bz2
- The updates of the Open Data from the hbz union catalog 21 🗒️ tar.bz2
- example of rdfmab 7 🗒️ text/turtle

Additional Information (settings)

- **Field**: Source, **Value**: https://wiki1.hbz-nrw.de/display/SEM/Recently%20published%20Open%20Data%20exports
- **Field**: Author, **Value**: North Rhine-Westphalian Library Service Center
- **Field**: Maintainer, **Value**: North Rhine-Westphalian Library Service Center
- **Field**: Version, **Value**: 2012-03-08
- **Field**: State, **Value**: active

License: Creative Commons CCZero
7. Let others know

March 2010: Releasing catalogue data: Cologne-based libraries to pioneer Open Data practices

Deutsche Version der Mitteilung

Joint statement of the North Rhine-Westphalian Library Service Center, the University and Public Library of Cologne, the University Library of the University of Applied Science of Cologne, the Library of the Academy of Media Arts Cologne and Library Centre of Rhineland-Palatinate.

March 2010: Cologne-based libraries and the Library Centre of Rhineland-Palatinate (LBZ) in cooperation with the North Rhine-Westphalian Library Service Center (hbz) are the first German libraries to adopt the idea of Open Access for bibliographic data. They are releasing their catalogue data for free public use. The University Library of Cologne Library of the Academy of Media Arts Cologne, the University Library of Public Library of Cologne has announced to follow shortly. The release of these libraries into the Open Access movement for a long time has only few libraries have done so with their own data. Roll Thiele, deputy director of the Cologne Public Library, has announced to follow shortly.

Libraries have been involved in the Open Access movement for a long time. They have decided to provide access to knowledge without barriers. Providing this time disregarded until now. Up to this point, it was not possible to download library holdings on the internet. The library of the European Organization &

Public data is placed in the public domain

The publication of data enables anybody to download, modify and use it. It is important to stick up for the traditional duty of libraries and librarians with the lowest restrictions possible. "said Sinead Schomburg, deputy director to the public without restrictions," she continued.

Cooperation and data exchange between libraries have been firmly established. For example, the data is exchanged between libraries by the German Wikipedia. The German Wikipedia for example has been edited by users and is distributed under the Creative Commons Attribution-Share Alike 3.0 license. Schomburg noted.

Data for the Semantic Web

The North Rhine-Westphalian Library Service Center has recently begun exploring the emerging Semantic Web. The liberalization of bibliographic data provides libraries with the opportunity to include data in the semantic web. The goal is to make use of the data and to make it accessible to other libraries. The goal is to make use of the data and to make it accessible to other libraries.

Further information and links to the published datasets are available at http://www.hbz-nrw.de/projekte/linked_open_data.

(Our new website will be launched on Monday, March 15th.)

For further questions contact:

Adrian Pohl
@acka47
The time has come! Cologne-based libraries & the LBZ in cooperation with the hbz open up their data:
http://tr.im/RCnV #opendata #CC0

8:16 PM - 12 Mar 10 via TweetDeck - Embed this Tweet
Your turn!
Agree on an open license within your group and link your document to that license.

(The predicate `<http://creativecommons.org/ns#license>` is well suited for this link, but searching the Web will reveal alternatives.)
Linked Data in Action
**Identification** and **description** of a resource ought to be distinguished! But in the Linked Data paradigm, there are currently two common variants to link both together.
Hash URIs

The server ignores the fragment identifier and delivers the description

http://www.example.org/people#alice

http://www.example.org/people
HTTP 303 Redirects

The server redirects to the URL of the description:

- http://www.example.org/people/alice
- http://www.example.org/data/alice

Location: http://www.example.org/data/alice
The description of a resource can be made available in various formats. Which format will be delivered can be decided by Content Negotiation.
Content Negotiation

http://www.example.org/data/alice

Accept: application/rdf+xml  Accept: text/html

Content-Location: alice.rdf  Content-Location: alice.html

http://www.example.org/data/alice

http://www.example.org/data/alice.rdf

http://www.example.org/data/alice.html
The easiest way to publish Linked Data is to serve static RDF-files that use Hash-URIs from a webserver that is set up to use Multiviews for Content-Negotiation.
With this simple setup, the **URIs of the people** are:
http://localhost/personen#ostrowski
http://localhost/personen#christoph

The **URL of the description** is:
http://localhost/personen

Content-Negotiation for Turtle and RDF/XML is activated.
Demo: Consuming Linked Data with cURL
Your turn!
In your description, link yourself to people from other groups that you know. This doesn't have to be reciprocal.

Also, link (approximately) to the place you live or work. Use DBpedia for this.
Break
Scattered machine-readable descriptions are useful, but we can do better than that! RDF is a distributed data model that makes it easy to combine several descriptions. Furthermore, special databases exist that allow to query RDF data.
Triple Stores

http://www.example.org/data/alice

http://de.dbpedia.org/page/Berlin

http://www.example.org/data/carol

http://de.dbpedia.org/page/Köln
SPARQL facilitates queries on the data in a triple store. The foundations for this are simply graph patterns. These look almost like triples, the difference being that the contain variables.
@prefix ex: <http://example.org/people#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:alice foaf:name „Alice“ .

PREFIX ex: <http://example.org/people#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT * WHERE {
  ex:alice foaf:name ?name .
}

name „Alice“
@prefix ex: <http://example.org/people#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:alice foaf:name "Alice" ;
    foaf:knows ex:bob .
ex:bob foaf:name "Bob" ;
    foaf:knows ex:carol .
ex:carol foaf:name "Carol" ;
    foaf:knows ex:alice .

PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?name1 ?name2 WHERE {
  ?person1 foaf:knows ?person2 .
  ?person1 foaf:name ?name1 .
}

name1         name2
"Alice"       "Bob"
"Bob"         "Carol"
"Carol"       "Alice"
Your turn!
Use SPARQL to analyse your connections. For example you might want to determine who you know directly or indirectly or who comes from the same city as you.
Break
Let's put some **Semantic** in the **Web**

The **classes** and **properties** being used can be using **description languages for vocabularies**. The relatively simple RDF Schema (**RDFS**) is wide spread, but more complex issues can be expressed in the Web Ontology Language (**OWL**).
ex:alice rdfs:knows ex:bob

foaf:Person rdfs:domain foaf:knows

foaf:knows rdf:type foaf:knows

ex:alice rdf:type foaf:Person

ex:bob rdf:type foaf:Person
# RDF Schema

foaf:knows rdfs:type rdfs:Property ;
    rdfs:range foaf:Person ;
    rdfs:domain foaf:Person .

foaf:Person rdf:type rdfs:Class .

# Explicit triples

ex:bob foaf:knows ex:alice .
# RDF Schema as a ”bridge“ across vocabularies

```
ex:colleague rdfs:subPropertyOf foaf:knows ;
    rdfs:domain        ex:Employee ;
    rdfs:range         ex:Employee .
```

```
ex:Employee rdf:type           rdfs:Class ;
    rdfs:subClassOf    foaf:Person .
```

---

# Implicit triple, that follow from the schema

```
ex:bob   foaf:knows ex:alice .
ex:bob   rdf:type foaf:Person .
ex:alice rdf:type foaf:Person .
ex:bob   rdf:type foaf:Employee .
ex:alice rdf:type foaf:Employee .
```
Your turn!
Create an RDF Schema so that from these assertions the following triples can be inferred.
team:player rdf:type rdfs:Property ;
   rdfs:subPropertyOf foaf:member ;
   rdfs:domain foaf:Person ;
   rdfs:range foaf:Group .

team:home rdf:type rdfs:Property ;
   rdfs:domain team:Game .

team:away rdf:type rdfs:Property ;
   rdfs:domain team:Game .

team:Game rdf:type rdfs:Class .
The expressiveness and the possibilities of inference of RDFS and OWL are not always needed. For controlled vocabularies, the **Simple Knowledge Organization System (SKOS)** is a simpler alternative that is also based on RDF. The **Dewey Decimal Classification** and the **Library of Congress Subject Headings** have already found their way into the Linked-Data-world.
Linked Data Principles

1) Use URIs as names for things.
2) Use HTTP URIs so that people can look up those names.
3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4) Include links to other URIs. So that they can discover more things.
Thanks.

Questions?

Here and now or also later to semweb@hbz-nrw.de
Credits

„Shunting“ by John Spooner CC-BY
http://www.flickr.com/photos/johnspooner/2199685678/

Exit arrow by ccarlstead CC-BY http://www.flickr.com/photos/cristic/553170147/
License

These slides are published under a Creative Commons license:
http://creativecommons.org/licenses/by/3.0/de/