Extending the Web with a global Data Space

Prof. Dr. Christian Bizer
Freie Universität Berlin
Germany
Outline

1. Foundations of Linked Data
   - What is the vision and goal?

2. The Web of Linked Data
   - What data is out there?
   - What is being done with the data?

3. How to publish Linked Data?
   - Tasks and Tools

4. How to consume Linked Data?
   - Tasks and Tools

5. Splitting the Data Integration Effort
   - How to reach global-scale data integration in an evolutionary fashion?
**Linked Data Principles**

Set of best practices for publishing structured data on the Web in accordance with the general architecture of the Web.

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 **RDF** information.
4. Include RDF statements that link to other URIs so that they can discover related things.

Architecture of the classic Web

Single global information space

Small set of simple standards

1. HTML as document format
2. HTTP URLs as
   • globally unique IDs
   • retrieval mechanism
3. Hyperlinks to connect everything
Web 2.0 APIs and Mashups

No single global dataspace

Shortcomings
1. APIs have proprietary interfaces
2. Mashups are based on a fixed set of data sources
3. No hyperlinks between data items within different APIs
Web APIs slice the Web into Walled Gardens
Extend the Web with a **single global dataspace**

1. by using RDF to publish structured data on the Web
2. by setting links between data items within different data sources.
The Basis: RDF Data Model

Flexible graph-based data model.
Data items are identified with HTTP URIs.

HTTP URIs take the role of global primary keys.

pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri

dbpedia:Berlin = http://dbpedia.org/resource/Berlin
Resolving URIs over the Web

The HTTP protocol brings together identification and retrieval again.
Following Links deeper into the Web

pd:cygri \(\rightarrow\) foaf:Person
- foaf:name \(\rightarrow\) Richard Cyganiak
- foaf:based_near \(\rightarrow\) dbpedia:Berlin

dbpedia:Berlin \(\rightarrow\) dp:population \(= 3.405.259\)
- skos:subject

dbpedia:Hamburg
- skos:subject

dbpedia:Muenchen
- skos:subject
<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>event</td>
<td>...</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>G1 G2 G3 G4</td>
</tr>
<tr>
<td>seeAlso</td>
<td><a href="http://richard.cyganiak.de/cygri.rdf">http://richard.cyganiak.de/cygri.rdf</a></td>
<td>G2</td>
</tr>
<tr>
<td>seeAlso</td>
<td><a href="http://richard.cyganiak.de/foaf.rdf">http://richard.cyganiak.de/foaf.rdf</a></td>
<td>G3</td>
</tr>
<tr>
<td>nearest airport</td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td>phone</td>
<td>tel:+49-175-5630408</td>
<td>G1</td>
</tr>
<tr>
<td>sameAs</td>
<td>Richard Cyganiak</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td>Berlin</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G1</td>
</tr>
<tr>
<td>currentProject</td>
<td><a href="http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs">http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs</a></td>
<td>G3</td>
</tr>
<tr>
<td>currentProject</td>
<td><a href="http://www.wiwi.fu-berlin.de/suhl/bizer#d2rq">http://www.wiwi.fu-berlin.de/suhl/bizer#d2rq</a></td>
<td>G3</td>
</tr>
<tr>
<td>depiction</td>
<td><img src="image.jpg" alt="Image" /></td>
<td>G4</td>
</tr>
<tr>
<td>gender</td>
<td>male</td>
<td>G1</td>
</tr>
<tr>
<td>Property</td>
<td>Value</td>
<td>Sources</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>population</td>
<td>3398888</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://dbpedia.org/City">http://dbpedia.org/City</a></td>
<td>G2</td>
</tr>
<tr>
<td>comment</td>
<td>Berlin is the capital city and one of the sixteen Federal States of Germany. It is the country's largest city in area and population, and the second most populous city in the European Union.</td>
<td>G2</td>
</tr>
<tr>
<td>comment</td>
<td>Berlin ist die deutsche Bundeshauptstadt und als Stadtstaat ein eigenständiges Land der Bundesrepublik Deutschland. Berlin ist die bevölkerungsreichste und flächengrößte Stadt Deutschlands und nach Einwohnern die zweitgrößte Stadt der EU.</td>
<td>G2</td>
</tr>
<tr>
<td>label</td>
<td>Berlin</td>
<td>G2</td>
</tr>
<tr>
<td>sameAs</td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Berlin">http://dbpedia.org/resource/category/Berlin</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Capitals_in_Europe">http://dbpedia.org/resource/category/Capitals_in_Europe</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Cities_in_Germany">http://dbpedia.org/resource/category/Cities_in_Germany</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/German_state_capitals">http://dbpedia.org/resource/category/German_state_capitals</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games">http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/States_of_Germany">http://dbpedia.org/resource/category/States_of_Germany</a></td>
<td>G2</td>
</tr>
<tr>
<td>sourceURL</td>
<td><a href="http://dbpedia.org/resource/category/Berlin">http://dbpedia.org/resource/category/Berlin</a></td>
<td>G1</td>
</tr>
<tr>
<td>depiction</td>
<td><img src="http://dbpedia.org/resource/category/Berlin" alt="Image" /></td>
<td>G2</td>
</tr>
<tr>
<td>is birthplace</td>
<td>Adolf von Baeyer</td>
<td>G2</td>
</tr>
</tbody>
</table>
Properties of the Web of Linked Data

- **Global, distributed dataspace build on a simple set of standards**
  - RDF, URIs, HTTP

- **Entities are connected by links**
  - creating a global data graph that spans data sources and
  - enables the discovery of new data sources

- **Provides for data-coexistence**
  - Everyone can publish data to the Web of Linked Data
  - Everyone can express their personal view on things
  - Everybody can use the vocabularies/schema that they like
The Structural Continuum

The Web of Linked Data is interwoven with the classic Web.

- Unstructured data: HTML
- Semi-structured data: RDFa embed into HTML
- Structured data: RDF/XML

- Services using named entity recognition to annotate texts with Linked Data URIs
  - Open Calais (Thomsons Reuters) for news
  - Zemanta (startup) for blog posts
  - DBpedia Spotlight
2. Linked Data Deployment on the Web

Is this real?
Grassroots community effort to

- publish existing open license datasets as Linked Data on the Web
- interlink things between different data sources
- Over 500 million RDF triples
- Around 120,000 RDF links between data sources
LOD Datasets on the Web: September 2008
Over 13.1 billion RDF triples
Over 142 million RDF links between data sources
- Over 24.7 billion RDF triples
- Over 436 million RDF links between data sources
### LOD data set statistics as of November 2010

<table>
<thead>
<tr>
<th>Domain</th>
<th>Data Sets</th>
<th>Triples</th>
<th>Percent</th>
<th>RDF Links</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-domain</td>
<td>20</td>
<td>1,999,085,950</td>
<td>7.42</td>
<td>29,105,638</td>
<td>7.36</td>
</tr>
<tr>
<td>Geographic</td>
<td>16</td>
<td>5,904,980,833</td>
<td>21.93</td>
<td>16,589,086</td>
<td>4.19</td>
</tr>
<tr>
<td>Government</td>
<td>25</td>
<td>11,613,525,437</td>
<td>43.12</td>
<td>17,658,869</td>
<td>4.46</td>
</tr>
<tr>
<td>Media</td>
<td>26</td>
<td>2,453,898,811</td>
<td>9.11</td>
<td>50,374,304</td>
<td>12.74</td>
</tr>
<tr>
<td>Libraries</td>
<td>67</td>
<td>2,237,435,732</td>
<td>8.31</td>
<td>77,951,898</td>
<td>19.71</td>
</tr>
<tr>
<td>Life sciences</td>
<td>42</td>
<td>2,664,119,184</td>
<td>9.89</td>
<td>200,417,873</td>
<td>50.67</td>
</tr>
<tr>
<td>User Content</td>
<td>7</td>
<td>57,463,756</td>
<td>0.21</td>
<td>3,402,228</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>203</strong></td>
<td><strong>26,930,509,703</strong></td>
<td><strong>395,499,896</strong></td>
<td><strong>Total</strong></td>
<td></td>
</tr>
</tbody>
</table>

**LOD Cloud Data Catalog on CKAN**

http://www.ckan.net/group/lodcloud

**More statistics**

http://www4.wiwiss.fu-berlin.de/lodcloud/state/
## The Growth in Numbers

<table>
<thead>
<tr>
<th>Year</th>
<th>Datasets</th>
<th>Triples</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>12</td>
<td>500,000,000</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>45</td>
<td>2,000,000,000</td>
<td>300%</td>
</tr>
<tr>
<td>2009</td>
<td>95</td>
<td>6,726,000,000</td>
<td>236%</td>
</tr>
<tr>
<td>2010</td>
<td>203</td>
<td>26,930,509,703</td>
<td>300%</td>
</tr>
</tbody>
</table>
Uptake in Life Sciences

- W3C Linking Open Drug Data Effort
- Bio2RDF Project
- Allen Brain Atlas

**Goal:** Smoothly integrate internal and external data in a pay-as-you-go-fashion.
The EU is pushing Linked Data (LOD2, LATC, EuroStat)

W3C eGovernment Interest Group
Uptake in the Libraries Community

- Institutions publishing Linked Data
  - Library of Congress (subject headings)
  - German National Library (PND dataset and subject headings)
  - Swedish National Library (Libris - catalog)
  - Hungarian National Library (OPAC and Digital Library)
  - Europeana project just released data about 4 million artifacts

- Growth of Library Linked Data (2009-2010): 1000%

- W3C Library Linked Data Incubator Group

- Open Archives ORE Standard

- Goals:
  1. Integrate Library Catalogs on global scale.
  2. Interconnect resources between repositories (by topic, by location, by historical period, by ...).
Uptake in the Media Industry

- Publish data as RDF/XML or RDFa
- Goal: Drive traffic to websites via search engines
Linked Data Applications

What can I do with this?

Linked Data Browsers

Linked Data Mashups

Search Engines

What can I do with this?
Linked Data Browsers

Provide for navigating between data sources and for exploring the data space.

- Tabulator Browser (MIT, USA)
- Marbles (FU Berlin, DE)
- OpenLink RDF Browser (OpenLink, UK)
- Zitgist RDF Browser (Zitgist, USA)
- Disco Hyperdata Browser (FU Berlin, DE)
- Fenfire (DERI, Irland)
Web of Data Search Engines

Crawl the dataspace and provide best-effort query answers over crawled data.

- Falcons (IWS, China)
- Sig.ma (DERI, Ireland)
- Swoogle (UMBC, USA)
- VisiNav (DERI, Ireland)
- Watson (Open University, UK)
Chicago - Begriff
- label: Chicago
- type: Begriff
http://www4.wiwiss.fu-berlin.de/bookmashup/subject/Chicago

Chicago - City, Community
- label: Chicago
- sameAs: http://www.rdfabout.com/rdf/usgov/geo/us/il/counties/cook_county/chicago
- image:
- type: Community
http://dbpedia.org/resource/Chicago

chicago
- Title: chicago
http://www.deadjournal.com/interests.bml?int=chicago

Chicago Cubs players - Begriff
- label: Chicago Cubs players
- bevorzugter Name: Chicago Cubs players
- hat Oberbegriff: Chicago Cubs field personnel
- hat Oberbegriff: Chicago Cubs
- type: Begriff
http://dbpedia.org/resource/Category:Chicago_Cubs_players

People from Chicago - Begriff
- label: People from Chicago
- http://dbpedia.org/resource/Category:People_from_Chicago
Ivan Herman
http://www.ivan-herman.net/
Document Resource Document

Ivan’s private site
http://ivan-herman.name/
RSS1.0 News Channel Document Resource

open source
http://www.advogato.org/person/connolly/
RSS1.0 News Channel Document organization
Advogato blog for connolly

Paul Downey
http://blog.whatfettle.com/
Document Resource Document
Domain-specific applications using Linked Data from the Web
See UK :: Schools (normalised by Population) at Ward level

Visualise

Schools

normalised
by Population
by Area
actual values

zoom
Ward
County
Region

West End North :: 1.2 schools per 1000 people

Compton and Otterbourne (2.8)
Bishopstoke East (0.73)
Eastleigh North (1.2)
Eastleigh Central (1.3)
Chandler's Ford East (1.9)
Eastleigh South (1.4)
Chandler's Ford West (1.1)

Hedge End Grange Park (0.57)
Hedge End St. John's (1.9)
Bursledon and Old Netley (1.0)
Botley (0.78)
Harefield (1.1)
Hedge End Wildern (0.37)
Peartree (0.79)
West End North (1.2)
West End South (1.0)
Bevois (0.82)

The picture is centred on West End North, and the circles get further away. Colour indicates the "worst" (red) and "best" (green) areas from those shown. Things can be viewed as a raw value, or normalised by population or area.

Find out more about this visualisation.

http://apps.seme4.com/see-uk/
DBpedia Mobile

- Geospatial entry point into the Web of Data
- Starts with DBpedia, Revyu and Flickr data
DERI Semantic Web Pipes
What are the big players doing?
Yahoo!

- crawls Linked Data in its RDFa serialization as well as Microformats.
- provides access to crawled data through the Yahoo BOSS API
- is using the data to make search results more useful and visually appealing.
Facebook’s Open Graph Protocol

- imports RDFa data from external web sites.
- For instance:
  - IMDb, Microsoft, NHL, Posterous
  - Rotten Tomatoes, TIME, Yelp
Google

- uses crawled data to enhance search results snippets and to feed its Social Graph API
- is developing Google Squared and Google Fusion Tables
- has recently bought MetaWeb
  - which maintains Freebase, a DBpedia/YAGO competitor
- is starting to use structured data in its applications:

![Data Snippets](image1.png)

![Query Answer](image2.png)
3. How to publish Linked Data?

Tasks:
1. Make data available as RDF via HTTP
2. Set RDF links pointing at other data sources
3. Make your data self-descriptive

Textbook

Tom Heath, Christian Bizer
Linked Data: Evolving the Web into a Global Data Space
http://linkeddatabook.com/
**Linked Data Publishing Patterns**

![Diagram of Linked Data Publishing Patterns](image)

1. **Data Preparation**
   - Structured Data
     - RDFizers for CVS, XML, Excel, ...
     - Entity Extractor (e.g., Calais)
   - Text

2. **Data Storage**
   - Relational Database
     - RDB-to-RDF Wrapper (e.g., D2R)
   - Data Source with API
     - CMS with RDFa Output (e.g., Drupal)
   - RDF Store
     - Custom Linked Data Wrapper
     - Linked Data Interface (e.g., Pubby)
   - RDF files

3. **Data Publication**
   - Linked Data on the Web
     - Web Server (e.g., Apache)
3.1 Make Data available as RDF via HTTP

Ready to use tools (examples)

1. D2R Server
   - provides for mapping relational databases into RDF and for serving them as Linked Data

2. Pubby
   - Linked Data Frontend for SPARQL Endpoints

3. More tools
   - http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/PublishingTools
3.2 Set RDF links pointing at other data sources

- **Examples of RDF links**

  ```xml
  <http://dbpedia.org/resource/Berlin> owl:sameAs
  <http://sws.geonames.org/2950159> .
  ```

  ```xml
  <http://richard.cyganiak.de/foaf.rdf#cygri> foaf:topic_interest
  ```

  ```xml
  <http://example-bookshop.com/book006251587X> owl:sameAs
  <http://www4.wiwiss.fu-berlin.de/bookmashup/books/006251587X> .
  ```
How to generate RDF links?

- **Pattern-based Approaches**
  - Exploit naming conventions within URIs (for instance ISBNs, ISINs, …)

- **Similarity-based Approaches**
  - Compare items within different data sources using various similarity metrics

**Ready to use tools** (Examples)

1. **Silk – Link Discovery Framework**
   - provides a declarative language for specifying link conditions which may combine different similarity metrics
   - Silk Single Machine, Silk MapReduce

2. **More tools**
   - [http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining](http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining)
3.3 Make your Data Self-Descriptive

- Increase the usefulness of your data and ease data integration

- Aspects of self-descriptiveness
  1. Enable clients to retrieve the schema
  2. Reuse terms from common vocabularies
  3. Publish schema mappings for proprietary terms
  4. Provide provenance metadata
  5. Provide licensing metadata
  6. Provide data-set-level metadata using voiD
  7. Refer to additional access methods using voiD
Enable Clients to retrieve the Schema

Clients can resolve the URIs that identify vocabulary terms in order to get their RDFS or OWL definitions.

Some data on the Web

```
<http://richard.cyganiak.de/foaf.rdf#cygri>
  foaf:name "Richard Cyganiak" ;
  rdf:type <http://xmlns.com/foaf/0.1/Person> .
```

Resolve unknown term

```
http://xmlns.com/foaf/0.1/Person
```

RDFS or OWL definition

```
<http://xmlns.com/foaf/0.1/Person>
  rdf:type owl:Class ;
  rdfs:label "Person" ;
  rdfs:subClassOf <http://xmlns.com/foaf/0.1/Agent> ;
  rdfs:subClassOf <http://xmlns.com/wordnet/1.6/Agent> .
```
Common Vocabularies

- **Friend-of-a-Friend** for describing people and their social network
- **SIOC** for describing forums and blogs
- **SKOS** for representing topic taxonomies
- **Organization Ontology** for describing the structure of organizations
- **GoodRelations** provides terms for describing products and business entities
- **Music Ontology** for describing artists, albums, and performances
- **Review Vocabulary** provides terms for representing reviews

Common sources of identifiers (URIs) for real world objects

- **LinkedGeoData** and **Geonames** locations
- **GeneID** and **UniProt** life science identifiers
- **DBpedia** wide range of things
4. How to consume Linked Data?
LDspider

- Flexible open-source Linked Data crawler
- Crawls RDF/XML and RDFa
R2R Framework

- Tool for translating RDF data between different vocabularies
- Provides for publication and discovery of mappings on the Web
Silk Server

- Add missing links while consuming Linked Data
- Designed to work together with LDspider
WIQA Framework

- Allows you to filter Web data using different data quality assessment policies
- Will be extended towards Data Fusion
Named Graphs

- Extension of the RDF Data Model for representing meta-information about RDF Graphs
- Implemented by most SPARQL stores
- Used by many Linked Data applications for provenance tracking

Provenance vocabularies

- Are compared by W3C Provenance XG
- Open Provenance Model is gaining traction
1. **Get some data using a crawler**
   - for instance: LDspider (GPL license)

2. **Store the data**
   - using for instance: Virtuoso (GPL), Sesame (BSD), Jena TDB (BSD)
   - or any relational database or column store you like
   - decision help: Berlin SPARQL Benchmark (February 2011)

3. **Query and analyze the data**
   - using the SPARQL query language
   - current version SPARQL 1.1 adds support for aggregates, subqueries, negation
Shortcut: Billion Triples Challenge Dataset

Download the Billion Triples Challenge Dataset

- 3.2 billion triples (27GB gzipped)
- crawled from the public Web of Linked Data in March/April 2010
- http://challenge.semanticweb.org/

If you do something interesting with the data

- submit your results to the challenge until October 1st
- present your results at the 10th International Semantic Web Conference (ISWC2011), October 2011, Koblenz, Germany
5. Splitting the Data Integration Effort

- Fix Overall Data Integration Effort
- Publisher’s Effort
- Third Party Effort
- Consumer’s Effort
The Dataspace Vision

Alternative to classic data integration systems in order to cope with growing number of data sources.

Properties of dataspaces

- require no upfront investment into a global schema
- rely on pay-as-you-go data integration
- give best effort answers to queries


Linked Data relies on the Pay-as-You-Go Idea

- for Identity Management
- for Vocabulary Management
Identity on the Web of Linked Data

Real world objects are identified with multiple URIs.

- Everybody can say everything about everything.
- Cheap to set up.

Linked Data website of our research group

http://www4.wiwiss.fu-berlin.de/is-group/resource/persons/Person4

Wrapper around the DBLP bibliography

http://dblp.l3s.de/d2r/resource/authors/Christian_Bizer
Publish Identity Links on the Web

**Identity Link**

```xml
<http://www4.wiwiss.fu-berlin.de/is-group/resource/persons/Person4>
owl:sameAs
<http://dblp.l3s.de/d2r/resource/authors/Christian_Bizer> .
```

**Pay-as-you-go Aspect**

1. First: Just put a wrapper in front of your DB
2. Later: You or somebody else invests effort into identity resolution
3. Publishes the results as Identity Links on the Web
Effort Distribution between Publisher and Consumer

Consumer data mines identity links

Effort Distribution

Publishers or third parties provides identity links
Everyone can use whatever vocabularies she likes to publish Linked Data on the Web.
Common Vocabularies

- Friend-of-a-Friend for describing people and their social network
- SIOC for describing forums and blogs
- SKOS for representing topic taxonomies
- Organization Ontology for describing the structure of organizations
- GoodRelations provides terms for describing products and business entities
- Music Ontology for describing artists, albums, and performances
- Review Vocabulary provides terms for representing reviews

Common sources of identifiers (URIs) for real world objects

- LinkedGeoData and Geonames locations
- GeneID and UniProt life science identifiers
- DBpedia wide range of things
Publish Vocabulary Links on the Web

**Simple Mappings: RDFS, OWL**
- rdfs:subClassOf, rdfs:subPropertyOf
- owl:equivalentClass, owl:equivalentProperty

**Complex Mappings: R2R**
- provides value transformation functions
- structural transformations

**Pay-as-you-go Aspect**
1. Use a mix of common vocabularies and proprietary terms
2. You or somebody else publishes schema mappings afterwards

Vocabulary Link

```
<http://xmlns.com/foaf/0.1/Person>
owl:equivalentClass
<http://dbpedia.org/ontology/Person> .
```
Effort Distribution between Publisher and Consumer

- Consumer defines or data mines mappings
- Publisher or third party publishes mappings
- Publisher reuses vocabularies

Diagram:
- Application Layer
  - Application Code
  - SPARQL
  - HTTP
- Data Access, Integration and Storage Layer
  - Web Data Access Module
  - Vocabulary Mapping Module
  - Identity Resolution Module
  - Quality Evaluation Module
- Web of Linked Data
- Publication Layer
  - LD Wrapper
  - Database A
  - Database B
  - Legacy App C
  - RDFa
  - RDF/XML

The overall data integration effort is split between the data publisher, the data consumer and third parties.

- **Data Publisher**
  - publishes data as RDF
  - sets identity links
  - reuses terms or publishes mappings

- **Third Parties**
  - set identity links pointing at your data
  - publish mappings to the Web

- **Data Consumer**
  - has to do the rest
  - using record linkage and schema matching techniques
Conclusion

- Linked Data realizes the data space vision on global scale and adds the social aspect to it.

- Web search is evolving into query answering
  - Search engines will increasingly rely on structured data from the Web

- Next step: Linked Data within Enterprises
  - alternative to data warehouses and EAI middleware
  - advantages: schema-less data model, pay-as-you go data integration

- You are looking for a topic for your PhD thesis?
  - There are many exciting research challenges around consuming Linked Data
  - Examples: Web-scale data integration, data quality assessment, UIs, …
References


- Christian Bizer, Tom Heath, Tim Berners-Lee: *Linked Data – The Story So Far*  

- *Linking Open Data* Project Wiki  
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

- 4th Linked Data on the Web Workshop at WWW 2011  
  http://events.linkeddata.org/ldow2011/

- 1st Workshop on Consuming Linked Data at ISWC 2010  
  http://people.aifb.kit.edu/aha/2010/cold/