Hallo

- Prof. Dr. Christian Bizer
- Professor for Information Systems
- Research Interests:
  - Global Data Spaces
  - Linked Data Technologies
  - Data- and Web Mining
- eMail: chris@informatik.uni-mannheim.de
Hello

- Dr. Heiko Paulheim
- Postdoctoral Researcher
- Research Interests:
  - Data Mining and Machine Learning on/with Linked Data
  - Ontology and Schema Matching
  - Data Quality
- eMail: heiko@informatik.uni-mannheim.de
1. Foundations of Linked Data
   - What is the vision and goal?

2. The Web of Linked Data
   - What data is out there?

3. How to publish and consume Linked Data?
   - Tasks and Tools
   - Sharing the Integration Effort

4. Alternative Web Data Publication Formats
   - RDFa, Microdata, Microformats

5. Challenges involved in using Web Data

6. Building Knowledge-intensive Applications
What does the classic Web offer us?
What do we actually want?

Use the Web like a single, global database
Access to structured Data on the Web

The Classic Document Web

The Web of Data

Web 2.0 APIs
Architecture of the classic Document 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
No single global dataspace

Shortcomings

1. APIs have proprietary interfaces
2. No hyperlinks between data items within different APIs
3. Mashups are based on a fixed set of data sources
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.
Linked Data

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.

Tim Berners-Lee, [http://www.w3.org/DesignIssues/LinkedData.html](http://www.w3.org/DesignIssues/LinkedData.html), 2006
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
The HTTP protocol brings together identification and retrieval again.
Following Links deeper into the Web

- **pd:cygri**
  - rdf:type: `foaf:Person`
  - foaf:name: Richard Cyganiak
  - foaf:based_near: `dbpedia:Berlin`

- **dbpedia:Hamburg**
- **dbpedia:Muenchen**
- **dp:Cities_in_Germany**
  - skos:subject: `dp:Cities_in_Germany`
  - dp:population: 3,405,259

- **dbpedia:Berlin**
  - skos:subject: `dp:Cities_in_Germany`

- **Richard Cyganiak**
  - foaf:based_near: `dbpedia:Berlin`

Christian Bizer, Heiko Paulheim: Linked Data on the Web. RR Summer School 2013 (07/30/2013)
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

- The Web of Linked Data can be used by generic applications
  - Linked Data Browsers
  - Linked Data Search Engines
<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.wiwiss.fu-berlin.de/suhl/bizer#d2rq">http://www.wiwiss.fu-berlin.de/suhl/bizer#d2rq</a></td>
<td>G3</td>
</tr>
<tr>
<td>depiction</td>
<td><img src="image.png" 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://en.wikipedia.org/wiki/Berlin">http://en.wikipedia.org/wiki/Berlin</a></td>
<td>G1</td>
</tr>
<tr>
<td>depiction</td>
<td><img src="http://en.wikipedia.org/wiki/Berlin" alt="Berlin" /></td>
<td>G2</td>
</tr>
<tr>
<td>is birthplace</td>
<td>Adolf von Baeyer</td>
<td>G2</td>
</tr>
</tbody>
</table>
Chris Bizer

picture:

given name: Chris [3,5,9,10,16]
family name: Bizer [3,5,9,10,16]
is creator of:
- DBpedia: A Nucleus for a Web of Open Data | Semantic Web Dog Food [6,18]
- The TriQLP Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies. [16]
- D2R Server - Publishing Relational Databases on the Semantic Web. [16]
- Named Graphs, Provenance and Trust [16]
- RAP: RDF API for PHP [16]
- Fresnel: A Browser-Independent Presentation Vocabulary for RDF [16]
- NG4L: Named Graphs API for Jena [16]
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
LOD Datasets on the Web: November 2011

- 31.6 billion RDF triples
- 503 million RDF links
## Distribution by Topical Domain (Nov 2011)

<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>Media</td>
<td>25</td>
<td>1,841,852,061</td>
<td>5.82%</td>
<td>50,440,705</td>
<td>10.01%</td>
</tr>
<tr>
<td>Geographic</td>
<td>31</td>
<td>6,145,532,484</td>
<td>19.43%</td>
<td>35,812,328</td>
<td>7.11%</td>
</tr>
<tr>
<td>Government</td>
<td>49</td>
<td>13,315,009,400</td>
<td>42.09%</td>
<td>19,343,519</td>
<td>3.84%</td>
</tr>
<tr>
<td>Library</td>
<td>87</td>
<td>2,950,720,693</td>
<td>9.33%</td>
<td>139,925,218</td>
<td>27.76%</td>
</tr>
<tr>
<td>Cross-domain</td>
<td>41</td>
<td>4,184,635,715</td>
<td>13.23%</td>
<td>63,183,065</td>
<td>12.54%</td>
</tr>
<tr>
<td>Life sciences</td>
<td>41</td>
<td>3,036,336,004</td>
<td>9.60%</td>
<td>191,844,090</td>
<td>38.06%</td>
</tr>
<tr>
<td>User content</td>
<td>20</td>
<td>134,127,413</td>
<td>0.42%</td>
<td>3,449,143</td>
<td>0.68%</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td>295</td>
<td>31,634,213,770</td>
<td></td>
<td>503,998,829</td>
<td></td>
</tr>
</tbody>
</table>

More statistics

http://lod-cloud.net/state/
Uptake in the Government Domain

- The EU is also starting to publish Linked Data
- Various other national efforts
Uptake in Life Sciences

- W3C Linking Open Drug Data Effort
- Bio2RDF Project

**Goal:** Smoothly integrate internal and external data in a pay-as-you-go-fashion.
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 Digital Library just released data about 4 million artifacts

- Goals:
  1. Integrate Library Catalogs on global scale.
  2. Interconnect resources between repositories (by topic, by location, by historical period, by ...).
Excursus: DBpedia
Excursus: DBpedia

- **DBpedia is a community effort**
  - to extract structured information from Wikipedia
  - make this data available on the Web under an open license

- **Contributors**
  - University of Mannheim (Germany)
  - Universität Leipzig (Germany)
  - OpenLink Software (UK)
Structured Data within Wikipedia

- **Title**: Bristol
- **Description**: Bristol is a city, unitary authority and ceremonial county in South West England, 105 miles (169 km) west of London, and 44 miles (71 km) east of Cardiff. With a population of about 410,900, it is England's second in the United Kingdom's most populous city, one of England's core cities and the most populous city in South West England. It received a Royal Charter in 1156 and was granted county status in 1373. For a half millennium it was the second or third largest English city. It is the capital of the ceremonial county of Gloucestershire (within which it is surrounded by the former administrative county of Avon) and also the county town of Gloucestershire (its seat until 1996 when it transferred to the city of Gloucester), and has a short coastline on the estuary of the Severn, which also forms the county boundary between Gloucestershire and Worcestershire. Bristol is one of the largest cities in the country, and is known for its historic docks which went into decline in the late 20th century. Bristol is a city of culture, industry and finance. It has a long history of trading and commerce, and the city centre docks have been regenerated as a centre of culture and commerce.

- **Language**: English
- **Links**:
  - **Geo-Coordinates**: 51°28′N 2°35′W
  - **Images**: Image of Bristol
  - **Infoboxes**: Infobox for Bristol including boundaries, government, and key statistics.
The DBpedia 3.8 Knowledge Base

- describes 3.77 million things, out of which 2.35 million are classified in a consistent ontology
  - 764,000 persons
  - 573,000 places
  - 192,000 organizations
  - 112,000 music albums

- Altogether 1.89 billion pieces of information (RDF triples)
  - 8,000,000 links to images
  - 24,000,000 links to external web pages
  - 27,200,000 external links into other RDF datasets

- DBpedia Internationalization Effort
  - provides data from 111 Wikipedia language editions for download
Highcliff
Highcliff is a 252.4-metre (828-foot) tall skyscraper located on a south slope of Happy Valley on the Hong Kong Island in Hong Kong. The 75 storey building’s construction began in 2000 and was completed in 2003 under a design by DLN Architects & Engineers. It was the Silver Winner of the 2003 Emporis Skyscraper Award, coming in second to 30 St Mary Axe in London.

The Harbourside
The Harbourside is a 255 m (836.6 ft) tall residential skyscraper located at 1 Austin Road West, in Union Square complex on Kowloon peninsula. The building is erected on the West Kowloon Reclamation west of Kwun Chung. Construction of the 74 storey building began in 2000 and was completed in 2003 under the design by P & T Architects & Engineers. The building is, in fact, three towers joined at the base, middle...
Other Examples of Linked Data Sets

- **Linked Geo Data**
  - Linked Data version of Open Street Maps
  - millions of places

- **Linked Movie Database**
  - data about movies, actors and directors
  - 40,000 films

- **Music Brainz**
  - musicians, albums
  - 22,000 albums, 40,000 musicians

- **DBLP**
  - computer science papers
  - 1.6 million articles
Questions so far?
3. How to Publish and Consume Linked Data?

Tasks involved in Publishing Linked Data:

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

Tom Heath and Christian Bizer:
Linked Data: Evolving the Web into a Global Data Space
http://linkeddatabook.com/
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

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

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

1. Pattern-based Approaches
   - Exploit naming conventions within URIs (for instance ISBNs, Gen IDs, …)

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

Link Generation Tools

- **Silk – Link Discovery Framework**
  - provides a user interface for specifying link conditions which may combine different similarity metrics

- **More tools**
  - [http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining](http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining)
A Silk Linkage Rule
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 / ontologies
  3. Publish schema mappings for proprietary terms
  4. Provide provenance metadata
  5. Provide licensing metadata

- Statistics about the compliance with these best practices
  - http://lod-cloud.net/state/
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

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
Usage of Common Vocabularies in the LOD Cloud

- Some terms from non-proprietary vocabularies: 191 (64.75%) of the 295 sources
- Only proprietary vocabularies: 104 (35.25%) of the 295 sources

**Common Vocabularies**

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc</td>
<td>92</td>
<td>31.19%</td>
</tr>
<tr>
<td>foaf</td>
<td>81</td>
<td>27.46%</td>
</tr>
<tr>
<td>skos</td>
<td>58</td>
<td>19.66%</td>
</tr>
<tr>
<td>geo</td>
<td>25</td>
<td>8.47%</td>
</tr>
<tr>
<td>akt</td>
<td>17</td>
<td>5.76%</td>
</tr>
<tr>
<td>bibo</td>
<td>14</td>
<td>4.75%</td>
</tr>
<tr>
<td>mo</td>
<td>13</td>
<td>4.41%</td>
</tr>
<tr>
<td>vcard</td>
<td>10</td>
<td>3.39%</td>
</tr>
<tr>
<td>sioc</td>
<td>10</td>
<td>3.39%</td>
</tr>
<tr>
<td>cc</td>
<td>8</td>
<td>2.71%</td>
</tr>
</tbody>
</table>
Terms for representing correspondences

- owl:equivalentClass, owl:equivalentProperty,
- rdfs:subClassOf, rdfs:subPropertyOf
- skos:broadMatch, skos:narrowMatch
Deployment of Vocabulary Links

Vocabulary links:
Vocabularies referencing "foaf" (119)
Vocabularies referenced by "mo" (17)


Christian Bizer, Heiko Paulheim: Linked Data on the Web. RR Summer School 2013 (07/30/2013)
3.3 Tasks involved in Consuming Linked Data

Application Layer

Data Access, Integration and Storage Layer

Web Access Module
Vocabulary Mapping Module
Identity Resolution Module
Quality Evaluation Module
Integrated Web Data

HTTP

Web of Linked Data

Publication Layer

LD Wrapper
HTTP
Database A

LD Wrapper
HTTP
Database B

RDFa
HTTP
Legacy App C

RDF/XML
HTTP
LDspider

- Flexible open-source Linked Data crawler
- Crawls RDF/XML and RDFa
- https://code.google.com/p/ldspider/
R2R Framework

- Tool for translating RDF data between different vocabularies
- http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/
- Alternative: Use SPARQL Construct queries to translate data
Silk Server

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

**Sieve Framework**
- Allows you to filter Web data using different data quality assessment policies
- Allows you to fuse data from different sources
- [http://sieve.wbsg.de/](http://sieve.wbsg.de/)

**WIQA Browser**
- Enables you to interactively employ different quality assessment policies
- Produces explanations about filtering decisions
- [http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/browser/](http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/browser/)
The WIQA Browser

Policy Selection Panel

Oh, yeah? Button

Christian Bizer, Heiko Paulheim: Linked Data on the Web. RR Summer School 2013 (07/30/2013)
The Triple:

Siemens Share positive analyst report Siemens agrees partnership with Novell unit SUSE. Siemens Business Services (SBS), the IT services arm of German technology conglomerate Siemens <SIEG.DE>, said on Tuesday it had agreed a partnership deal with Novell's (nasdaq: NOVL - news - people) newly acquired unit SUSE Linux. Linux software is open-source, meaning it can be freely copied and modified, unlike proprietary software such as Microsoft (nasdaq: MSFT - news - people) Windows. In the past months clients have been asking more and more for open-source platforms, SBS said in a statement which said SUSE would have premier partner status. SBS is one of Europe's top IT information technology service providers. Linux, once the exclusive province of a few dedicated enthusiasts, is now seen as the only serious rival to Windows and is supported by U.S. giant International Business Machines (nyse: IBM - news - people), among others. Its advocates, who include big businesses and government departments, argue it is cheaper, simpler and more secure than Windows.

fulfils the policy:

Use only information which has been asserted by German analysts.

because:

- it is stated in the document Information from Peter Smith, which is asserted by the German analyst Peter Smith.
Naive Reasoning on Web Data does not work!

Experiment: Naive RDF Schema reasoning on DBpedia data

- What are the rdf:types of dbpedia:Germany?
- Results: Place, Award, Populated Place, City, SportsTeam, Mountain, Agent, Organisation, Country, Stadium, RecordLabel, MilitaryUnit, Company, EducationalInstitution, PersonFunction, EthnicGroup, Architect, WineRegion, Language, MilitaryConflict, Settlement, RouteOfTransportation

What is going on here?

- DBpedia data is noisy as it was produced by many different people
- With naïve reasoning one wrong statement is enough for a wrong conclusion
- Germany example: 38,000 statements, 20 wrong types from 20 wrong statements (error rate of 0.05%)

Conclusion

- Always assess the quality of Web data before applying any reasoning
- Alternatively use robust reasoning methods
  (for instance: Paulheim/Bizer: Type inference on noisy RDF data. ISWC 2013)
The Dataspace Vision

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

- Properties of dataspaces
  - provide for data-coexistence
  - require no upfront investment into a global schema
  - give best effort answers to queries
  - rely on pay-as-you-go data integration

Franklin, M., Halevy, A., and Maier, D.: From Databases to Dataspaces

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

- for Identity Management
- for Schema/Vocabulary Management
Providing Integration Hints

- by publishing **Identity Links** on the Web

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

- You publish links pointing at other data sources.
- Somebody else publishes links pointing at your data source.
Effort Distribution between Publisher and Consumer

Publisher or third parties provides identity links

Consumer data mines identity links

Web of Linked Data

Application Layer

Data Access, Integration and Storage Layer

Web of Linked Data

Publication Layer

HTTP

LD Wrapper

Database A

HTTP

LD Wrapper

Database B

HTTP

RDFa

Legacy App C

HTTP

RDF/XML

Application Code

SPARQL

HTTP

Web Data Access Module

Vocabulary Mapping Module

Identity Resolution Module

Quality Evaluation Module

Integrated Web Data
Providing Integration Hints

- by publishing **Vocabulary Links** on the Web

Vocabulary Link

```
<http://xmlns.com/foaf/0.1/Person> owl:equivalentClass
<http://dbpedia.org/ontology/Person> .
```

- **Terms for expressing Correspondences**
  - owl:equivalentClass, owl:equivalentProperty
  - rdfs:subClassOf, rdfs:subPropertyOf
Effort Distribution between Publisher and Consumer

Consumer defines or data mines mappings

Effort Distribution

Publisher reuses vocabularies

Publisher or third party publishes mappings
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
4. Alternative Web Data Publishing Formats

More and more Websites semantically markup the content of their HTML pages.

- Microformats
- RDFa
- Microdata
- HTML
Microformats

- Microformat effort dates back to 2003
- Small set of fixed formats
  - hcard: people, companies, organizations, and places
  - XFN: relationships between people
  - hCalendar: calendaring and events
  - hListing: small-ads; classifieds
  - hReview: reviews of products, businesses, events
- Shortcoming of Microformats
  - can not represent any kind of data.
- indexed by Google and Yahoo since 2009
RDFa

- serialization format for embedding RDF data into HTML pages
- proposed in 2004, W3C Recommendation in 2008
- can be used together with any vocabulary
- can assign URIs as global primary keys to entities

```html
<html xmlns="http://www.w3.org/1999/xhtml"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:foaf="http://xmlns.com/foaf/0.1/>

...   <div about="http://example.com/Peter" typeof="foaf:Person">
       <span property="foaf:name">Peter Smith</span> knows
   </div>
...```

Christian Bizer, Heiko Paulheim: Linked Data on the Web. RR Summer School 2013 (07/30/2013)
Open Graph Protocol

- allows site owners to determine how entities are described in Facebook
- relies on RDFa for encoding data in HTML pages
- available since April 2010
Microdata

- alternative technique for embedding structured data
- proposed in 2009 by WHATWG as part of HTML5 work
- tries to be simpler than RDFa (5 new attributes instead of 8)
- W3C currently tries to reconcile the two alternative proposals

```html
1 <div itemscope itemtype="http://schema.org/Person"itemid="http://example.com/Peter">
2  <span itemprop="name">Peter Smith</span>
3  <a href="http://example.com/Paula" itemprop="knows">Paula Jones</a>
4 </div>
```
Schema.org

- ask site owners to embed data to enrich search results.
- 200+ Types: Event, Organization, Person, Place, Product, Review
- Encoding: Microdata or alternatively RDFa
Usage of Schema.org Data @ Google

Data snippets within search results

Data tables within search results

Answers to fact queries
Common Crawl is a non-profit foundation dedicated to building and maintaining an open crawl of the web, thereby enabling a new wave of innovation, education and research.
Web Data Commons

- **WebDataCommons.org Project**
  - extracts all Microformat, Microdata, RDFa data from the Common Crawl
  - provides the extracted data for free download

- **Two extractions runs**
  - 2009/2010 CC Corpus: 2.5 billion HTML pages → 5.1 billion RDF triples
  - 2012 CC Corpus: 3.0 billion HTML pages → 7.3 billion RDF triples

- **Jointed project of**
  - Universität Mannheim
  - Karlsruhe Institute of Technology
369 million of the 3 billion pages contain Microformat, Microdata or RDFa data (12.3%).

2.29 million websites (PLDs) out of 40.6 million provide Microformat, Microdata or RDFa data (5.65%).
### RDFa Topics (CC 2012)

#### Top Classes:

- CMS and Blog metadata
- Product data
- Ratings
- Company listings

<table>
<thead>
<tr>
<th>Class</th>
<th>PLDs Total #</th>
<th>PLDs in Alexa #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>183,046</td>
<td>17,002</td>
</tr>
<tr>
<td>2</td>
<td>58,971</td>
<td>5,820</td>
</tr>
<tr>
<td>3</td>
<td>56,573</td>
<td>9,533</td>
</tr>
<tr>
<td>4</td>
<td>49,252</td>
<td>2,802</td>
</tr>
<tr>
<td>5</td>
<td>44,644</td>
<td>2,794</td>
</tr>
<tr>
<td>6</td>
<td>33,141</td>
<td>2,188</td>
</tr>
<tr>
<td>7</td>
<td>19,331</td>
<td>1,327</td>
</tr>
<tr>
<td>8</td>
<td>19,107</td>
<td>3,389</td>
</tr>
<tr>
<td>9</td>
<td>13,477</td>
<td>1,135</td>
</tr>
<tr>
<td>10</td>
<td>9,054</td>
<td>2,123</td>
</tr>
<tr>
<td>11</td>
<td>6,994</td>
<td>691</td>
</tr>
<tr>
<td>12</td>
<td>6,758</td>
<td>1,067</td>
</tr>
<tr>
<td>13</td>
<td>6,236</td>
<td>1,410</td>
</tr>
<tr>
<td>14</td>
<td>4,139</td>
<td>845</td>
</tr>
<tr>
<td>15</td>
<td>3,936</td>
<td>308</td>
</tr>
<tr>
<td>16</td>
<td>3,339</td>
<td>456</td>
</tr>
<tr>
<td>17</td>
<td>3,303</td>
<td>606</td>
</tr>
<tr>
<td>18</td>
<td>3,167</td>
<td>401</td>
</tr>
<tr>
<td>19</td>
<td>3,155</td>
<td>392</td>
</tr>
<tr>
<td>20</td>
<td>2,502</td>
<td>367</td>
</tr>
</tbody>
</table>

og = Facebook’s Open Graph Protocol
Microdata Topics (CC 2012)

- **Top Classes:**
  - CMS and Blog metadata
  - Navigational metadata
  - Products and offers
  - Business listings
  - Ratings

<table>
<thead>
<tr>
<th>Class</th>
<th>PLDs Total</th>
<th>PLDs in Alexa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>1 schema:BlogPosting</td>
<td>25,235</td>
<td>17.98</td>
</tr>
<tr>
<td>2 datavoc:Breadcrumb</td>
<td>21,729</td>
<td>15.49</td>
</tr>
<tr>
<td>3 schema:PostalAddress</td>
<td>19,592</td>
<td>13.96</td>
</tr>
<tr>
<td>4 schema:Product</td>
<td>16,612</td>
<td>11.84</td>
</tr>
<tr>
<td>5 schema:LocalBusiness</td>
<td>16,383</td>
<td>11.68</td>
</tr>
<tr>
<td>6 schema:Article</td>
<td>15,718</td>
<td>11.20</td>
</tr>
<tr>
<td>7 datavoc:Review-aggregate</td>
<td>8,517</td>
<td>6.07</td>
</tr>
<tr>
<td>8 schema:Offer</td>
<td>8,456</td>
<td>6.03</td>
</tr>
<tr>
<td>9 datavoc:Rating</td>
<td>7,711</td>
<td>5.50</td>
</tr>
<tr>
<td>10 schema:AggregateRating</td>
<td>7,029</td>
<td>5.01</td>
</tr>
<tr>
<td>11 schema:Organization</td>
<td>7,011</td>
<td>5.00</td>
</tr>
<tr>
<td>12 datavoc:Product</td>
<td>6,770</td>
<td>4.82</td>
</tr>
<tr>
<td>13 schema:WebPage</td>
<td>6,678</td>
<td>4.76</td>
</tr>
<tr>
<td>14 datavoc:Organization</td>
<td>5,853</td>
<td>4.17</td>
</tr>
<tr>
<td>15 datavoc:Address</td>
<td>5,559</td>
<td>3.96</td>
</tr>
<tr>
<td>16 schema:Person</td>
<td>5,237</td>
<td>3.73</td>
</tr>
<tr>
<td>17 schema:GeoCoordinates</td>
<td>4,677</td>
<td>3.33</td>
</tr>
<tr>
<td>18 schema:Place</td>
<td>4,131</td>
<td>2.94</td>
</tr>
<tr>
<td>19 schema:Event</td>
<td>4,102</td>
<td>2.92</td>
</tr>
<tr>
<td>20 datavoc:Person</td>
<td>2,877</td>
<td>2.05</td>
</tr>
<tr>
<td>21 datavoc:Review</td>
<td>2,816</td>
<td>2.01</td>
</tr>
</tbody>
</table>

datavoc = Google‘s Rich Snippet Vocabulary
schema = Schema.org
## Microformats

### Top Classes:

#### Topics
- Persons
- Organisations
- Events

#### Listings and Reviews
- Recipes

<table>
<thead>
<tr>
<th>Class</th>
<th>PLDs Total #</th>
<th>PLDs Total %</th>
<th>PLDs in Alexa #</th>
<th>PLDs in Alexa %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1   hCard:VCard</td>
<td>1,511,467</td>
<td>84.03</td>
<td>87,758</td>
<td>83.79</td>
</tr>
<tr>
<td>2   hCard:Organization</td>
<td>195,493</td>
<td>10.87</td>
<td>10,430</td>
<td>9.96</td>
</tr>
<tr>
<td>3   hCard:Location</td>
<td>48,415</td>
<td>2.69</td>
<td>2,784</td>
<td>2.66</td>
</tr>
<tr>
<td>4   hCalendar:vcalendar</td>
<td>37,620</td>
<td>2.09</td>
<td>4,614</td>
<td>4.41</td>
</tr>
<tr>
<td>5   hCalendar:Vevent</td>
<td>36,349</td>
<td>2.02</td>
<td>4,400</td>
<td>4.20</td>
</tr>
<tr>
<td>6   hReview:Review</td>
<td>20,781</td>
<td>1.16</td>
<td>3,659</td>
<td>3.49</td>
</tr>
<tr>
<td>7   hListing:Lister</td>
<td>4,030</td>
<td>0.22</td>
<td>244</td>
<td>0.23</td>
</tr>
<tr>
<td>8   hListing:Listing</td>
<td>4,030</td>
<td>0.22</td>
<td>244</td>
<td>0.23</td>
</tr>
<tr>
<td>8   hRecipe:Recipe</td>
<td>3,281</td>
<td>0.18</td>
<td>1,068</td>
<td>1.02</td>
</tr>
<tr>
<td>10  hListing:Item</td>
<td>2,957</td>
<td>0.16</td>
<td>164</td>
<td>0.16</td>
</tr>
<tr>
<td>11  hRecipe:Ingredient</td>
<td>2,658</td>
<td>0.15</td>
<td>891</td>
<td>0.85</td>
</tr>
<tr>
<td>12  hRecipe:Duration</td>
<td>1,323</td>
<td>0.07</td>
<td>473</td>
<td>0.45</td>
</tr>
<tr>
<td>13  hRecipe:Nutrition</td>
<td>818</td>
<td>0.05</td>
<td>300</td>
<td>0.29</td>
</tr>
<tr>
<td>14  species:species</td>
<td>91</td>
<td>0.01</td>
<td>38</td>
<td>0.04</td>
</tr>
<tr>
<td>15  species:Genus</td>
<td>61</td>
<td>0.00</td>
<td>24</td>
<td>0.02</td>
</tr>
<tr>
<td>16  species:Family</td>
<td>60</td>
<td>0.00</td>
<td>24</td>
<td>0.02</td>
</tr>
<tr>
<td>17  species:Kingdom</td>
<td>59</td>
<td>0.00</td>
<td>24</td>
<td>0.02</td>
</tr>
<tr>
<td>18  species:Order</td>
<td>59</td>
<td>0.00</td>
<td>25</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Linked Data vs. HTML-embedded Data

Compared to Microformats, Microdata, RDFa

- the LOD Cloud covers a wider range of topics
- the LOD Cloud contains more complex data structures
- Emphasis on setting RDF Links between sources
Overall Topology of the Web of Data
5. Challenges involved in using Web Data

Applications hate heterogeneity and uncertain data quality!

The wild wild west

My little world
Current Research Challenges

1. More research on **data space profiling** is needed.
   - What is in the data space and how does the content change over time?

2. More research on **data quality assessment and SPAM detection** is needed.

3. More research on learning **mappings and identity resolution heuristics** within the Web context.
   - Identity links make it easier to learn vocabulary links.
   - Vocabulary links make it easier to learn identity links.

4. More research on **pay-as-you-go data integration** is needed.
   - How do human, community and machine contributions play together over time?
Conclusion

- The Web of Data is growing rapidly
  - Active deployment communities exist in various domains
  - Value-able resource of background knowledge for many applications

- Web search is evolving into query answering
  - Search engines 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 Web Data
  - Examples: Web-scale data integration, data quality assessment
Thanks!

References

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

- Tom Heath, Christian Bizer: Linked Data – Evolving the Web into a global data space.
  http://linkeddatabook.com/editions/1.0/

- 4th Workshop on Consuming Linked Data at ISWC 2013
  http://db.uwaterloo.ca/cold2013/

- 6th Linked Data on the Web Workshop at WWW 2013
  http://events.linkeddata.org/ldow2013/