1st International Workshop on Consuming Linked Data (COLD2010)
November 8, 2010, Shanghai, China

The R2R Framework: Publishing and Discovering Mappings on the Web

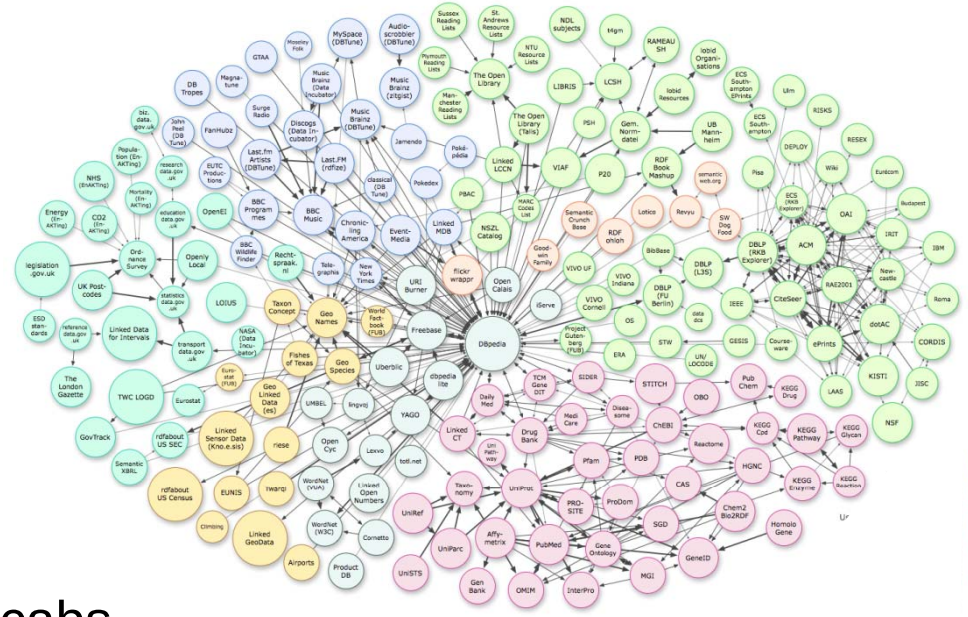
Christian Bizer, Andreas Schultz

Freie Universität Berlin

Outline

1. **Web-scale Data Integration**
   - Diversity and Agreement in the LOD Cloud
   - How would a webby solution look like?
2. **The R2R Framework**
   - The R2R Mapping Language
   - The R2R Mapping Engine
3. **Conclusions**

Diversity and Partial Agreement in the LOD Cloud



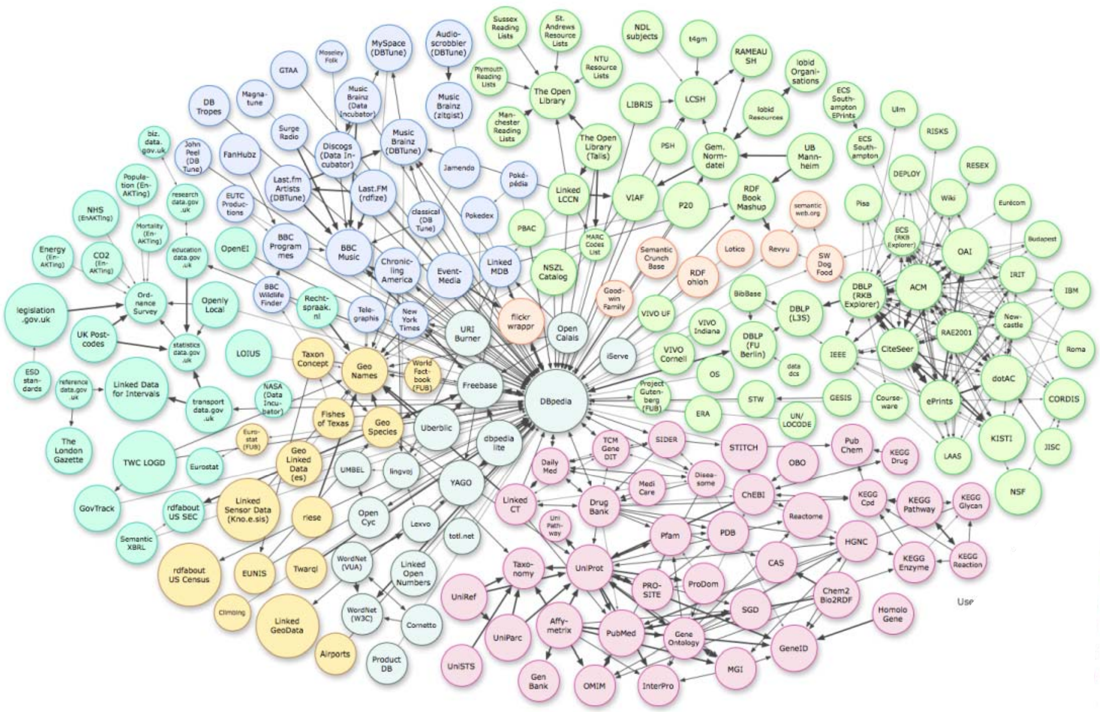
- **Some data sources reuse terms from widely-used vocabularies**
  - 77 data sources use only terms from widely-used vocabularies
  - 110 use some terms from widely-used vocabs
- **Others only use proprietary vocabularies**
  - 93 data sources use only proprietary terms

Source: http://www.lod-cloud.net/state

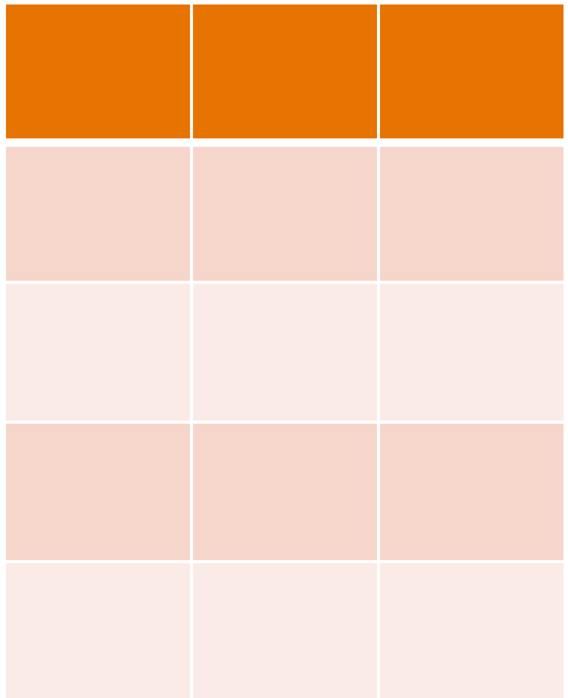
- **Data sources use terms from widely-used vocabularies differently**
  - foaf:name “C. Bizer” vs. foaf:name “Bizer, Christian”
  - units of measurement: distances, temperatures, …

Applications love Homogenity

Heterogeneity and data quality are the mayor challenges for Linked Data applications



The wild wild west

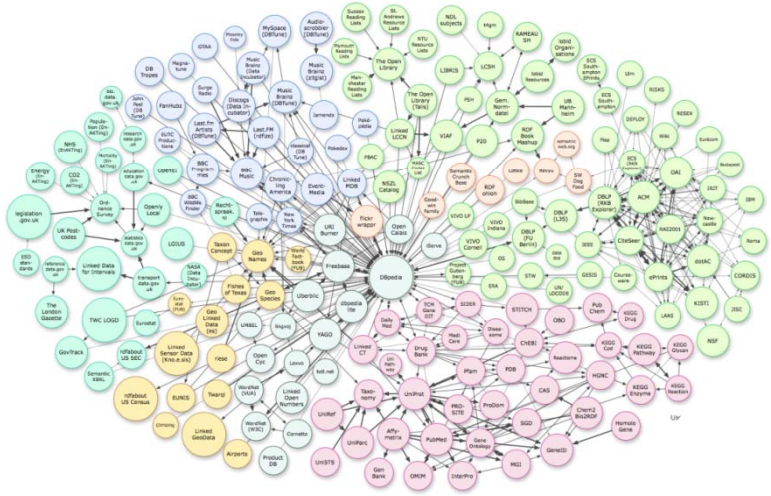


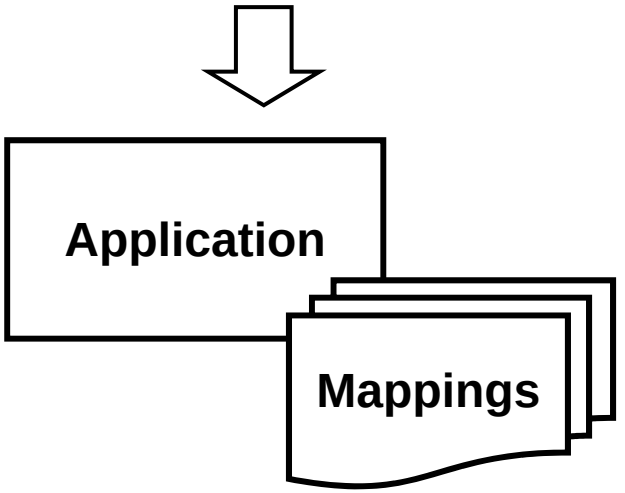
My little world

Classic Data Integration

The more data sources an application wants to integrate, the more mappings it needs.

- **How to generate mappings?**
  - write them by hand (expensive)
  - data mine mappings (uncertain quality)
- **Pay-as-you-go Integration**
  - integration effort can be split over time
  - Madhavan, et. al.: Web-scale Data Integration: You can only afford to pay-as-you-go. http://research.yahoo.com/files/paygo.pdf



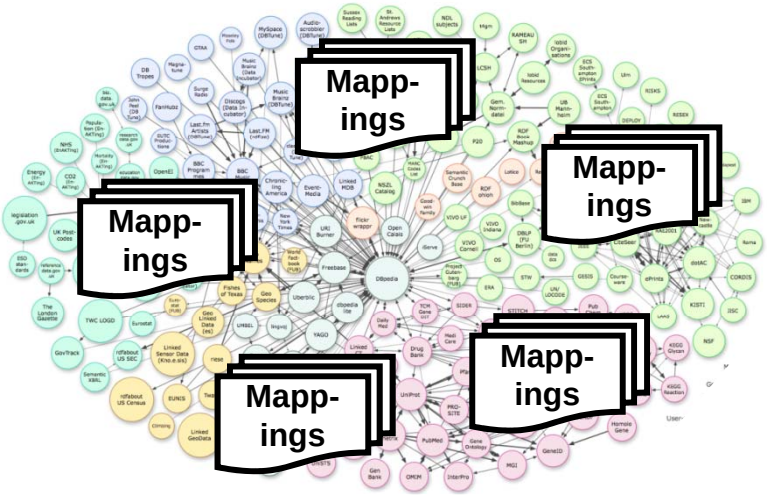


A more Webby Approach: Self-descriptive Data

Enable applications to discover everything on the Web that they need to integrate data.

1. **Publish mappings on the Web**
2. **Interlink Data and Mappings**

- **Terms from W3C recommendations to represent mappings**
  - owl:equivalentClass, owl:equivalentProperty
  - rdfs:subClassOf, rdfs:subPropertyOf
  - SKOS mapping properties



Mapping Discovery

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

Term definition including mappings

```
<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> ;
           owl:equivalentClass <http://dbpedia.org/ontology/Person> .
```

Somebody-Pays-As-You-Go

On the Web, the overall data integration effort can be split between data publishers, data consumers and third parties.

Data Publisher

- publishes data as RDF
- sets links and publishes mappings

Third Parties

- set links pointing at your data
- publish mappings to the Web

Data Consumer

- has to do the rest
  (for instance using data mining)
- has to assess the quality of mappings

Drawbacks of the OWL, RDFS, and SKOS Terms

1. **Not very expressive**
   - don’t provide for structural transformations
   - don’t provide value transformation functions (i.e. for units of measurement)
   - no literal modifiers to add data types or language tags
2. **Can only represent schema-level mappings**
   - no way to deal with the specifics of how a term is used by different data sources (foaf:name “C. Bizer” vs. foaf:name “Bizer, Christian”)
3. **Mappings have no URIs**
   - not easy to provide mapping specific meta-information
   - i.e. provenance or ratings for later mapping quality assessment

The R2R Framework

R2R Mapping Language

1. provides more expressivity
   - structural transformations
   - property value transformation functions
   - literal modifiers to add data types and language tags
2. distinguishes between schema-level and dataset-level mappings
3. allows meta-information to be published about mappings

R2R Mapping Engine

1. translates data to given target schema
2. assembles mappings into chains in order to overcome missing mappings
3. takes mapping quality into account

The R2R Mapping Language

- builds on SPARQL
- Example: Film runtime – Freebase to DBpedia

```
01: <http://mappings.dbpedia.org/r2r/FilmRuntimeFreebaseToDBpedia>
02:  rdf:type r2r:Mapping ;
03:  r2r:prefixDefinitions "dbpedia-owl: <http://dbpedia.org/ontology/>
04:                         . fb: <http://rdf.freebase.com/ns/>" ;
05:  r2r:sourcePattern  "?SUBJ fb:film.film.runtime ?ro .
06:                 ?ro fb:film.film_cut.runtime ?runtimeInMinutes" ;
07:  r2r:targetPattern "?SUBJ dbpedia-owl:runtime
08:                   ?runtimeInSeconds^^xsd:double" ;
09:  r2r:transformation "?runtimeInSeconds = ?runtimeInMinutes * 60" ;
10:  dc:creator www4:is-group/resource/persons/Person4;
11:  dc:date "2010-06-23"^^xsd:date.
```

Evaluation of the Expressivity

Class : Data sources	URI repl-ace	Struc trans 1:1	Struc trans 1:n	Val trans	UoM trans	DT mod	LG mod	L2U mod
Place : GeoNames / DBpedia	x					x		
Artist : MusicBrainz / DBpedia	x	x						
Place : NYT / DBpedia	x							
Country : Factbook / DBpedia	x	x	x		x	x		
Book : BookMashup / DBpedia	x	x		x		x	x	x
Author : Gutenberg / DBpedia	x	x	x	x			x	
County : US Census / DBpedia	x			x		x		
Organiza. : Dailymed /DBpedia	x							
Film : Linkedmdb / DBpedia	x	x			x	x		
Drug : Drugbank / DBpedia	x	x	x	x		x	x	x
Film : Freebase / DBpedia	x	x		x	x	x		
Musician : Freebase / DBpedia	x	x		x	x	x		

How to discover R2R Mappings?

Interlink Mappings with Vocabulary Terms

Some data on the Web

```
<http://richard.cyganiak.de/foaf.rdf#cygri>
      foaf:name "Richard Cyganiak" ;
      rdf:type <http://dbpedia.org/ontology/Person> .
```

Resolve `http://dbpedia.org/ontology/Person`

Term definition including link to mapping

```
<http://dbpedia.org/ontology/Person>
           rdf:type owl:Class ;
           rdfs:label "Person";
   r2r:hasMapping <http://dbpedia.org/mappings/r2r/PersonDBpediToFoaf>
```

Resolve `http://dbpedia.org/mappings/r2r/PersonDBpediToFoaf`

R2R Mapping

```
<http://dbpedia.org/mappings/r2r/PersonDBpediToFoaf>
          rdf:type r2r:Mapping ;
          r2r:sourcePattern "?SUBJ rdf:type dbpedia-owl:Person" ;
          r2r:targetPattern "?SUBJ rdf:type foaf:Person" .
```

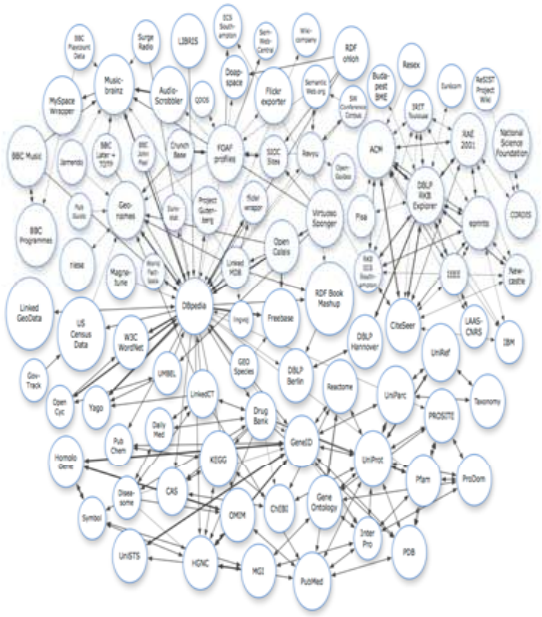
Interlink Mappings with VoiD Dataset Descriptions

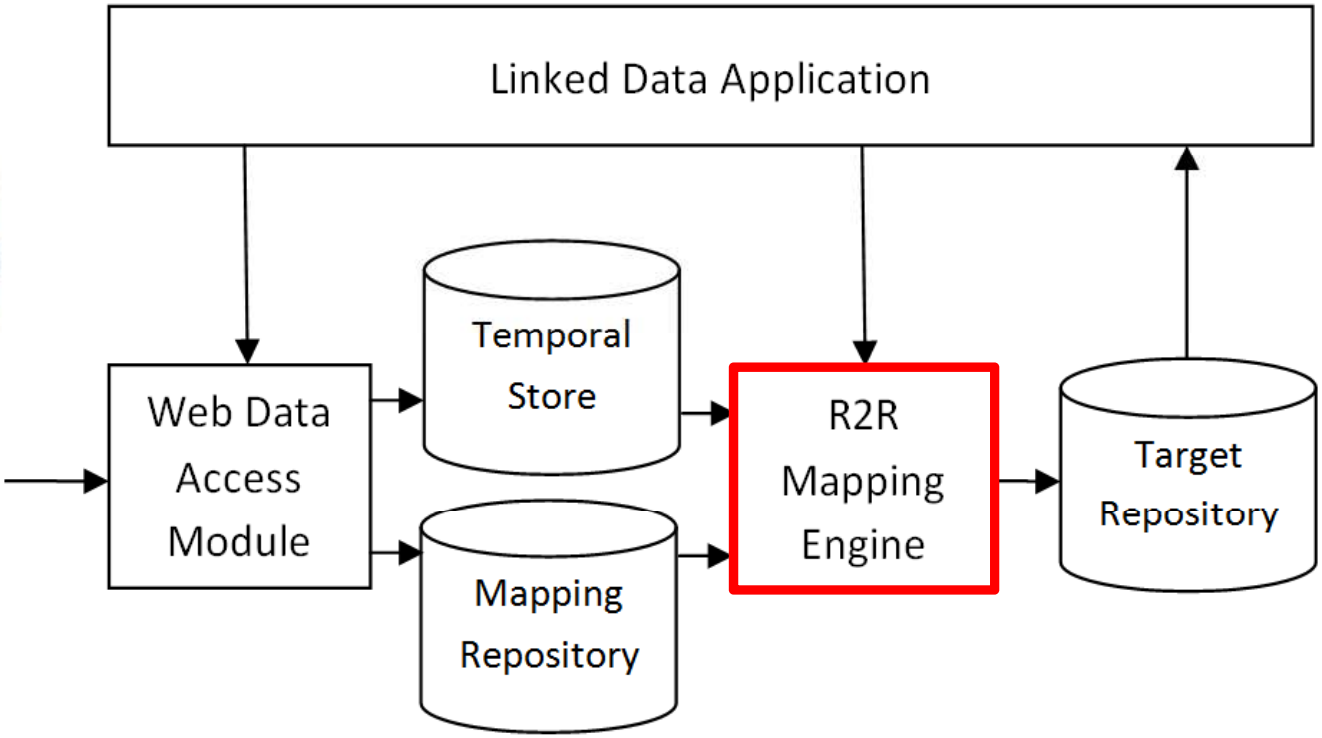
- provides for expressing dataset-specific mappings
- the R2R Mapping Engine prefers dataset-level mappings to vocabulary-level mappings.

```
01: <http://mappings.dbpedia.org/r2r/FilmRuntimeFreebaseToDBpedia>
02:  rdf:type r2r:Mapping ;
03:  r2r:sourcePattern "?SUBJ fb:film.film.runtime ?ro .
04:                     ?ro fb:film.film_cut.runtime ?runtimeInMinutes" ;
05:  r2r:targetPattern "?SUBJ dbpedia-owl:runtime
06:                     ?runtimeInSeconds^^xsd:double" ;
07:  r2r:transformation "?runtimeInSeconds = ?runtimeInMinutes * 60" ;
08:  r2r:sourceDataset <http://mappings.dbpedia.org/r2r/freebaseVOID> ;
09:  r2r:targetDataset <http://dbpedia.org/DBpediaVOID> ;
10:  dc:creator www4:is-group/resource/persons/Person4;
11:  dc:date "2010-06-23"^^xsd:date.
```

The R2R Mapping Engine

- **translates Web data to given target schema**
  - list of target classes and properties; different namespaces possible
- **combines mappings into mapping chains**
- **applies quality assessment heuristic while chaining mappings**





The Quality Assessment Heuristic

is build on the following assumptions:

1. **prefer vocabulary-level mappings provided by vocabulary maintainers to mappings provided by third parties**
2. **prefer dataset-level mappings provided by dataset maintainers to mappings provided by third parties**
3. **prefer dataset-level mappings to vocabulary-level mappings**
4. **prefer short mapping chains**

Conclusions

The R2R Framework

1. introduces an expressive SPARQL-based mapping language
2. introduces the distinction between dataset-level and vocabulary-level mappings
3. uses a quality assessment heuristic for choosing mapping from the Web
4. is available under Apache license

Publish mappings on the Web !

- currently only 5% of the LOD sources publish mappings ☹
- use OWL, RDFS, or SKOS terms
- if you want to be specific, use R2R

Thanks!

References

- R2R Website:
  http://www4.wiwiss.fu-berlin.de/bizer/r2r/
- R2R Paper: The R2R Framework – Publishing and Discovering Mappings on the Web
  http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/BizerSchultz-COLD-R2R-Paper.pdf
- Madhavan, et. al.: Web-scale Data Integration: You can only afford to pay-as-you-go. CIDR-07. http://research.yahoo.com/files/paygo.pdf