DBpedia

Manuel Fiorelli

fiorelli@info.uniroma2.it

Notes

The following slides contain some examples and pictures taken from:

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167-195.

An author-archived copy can be found here:

https://www.researchgate.net/profile/Christian_Bizer/publication/259828897_DBpedia_-_A_Large-scale_Multilingual_Knowledge_Base_Extracted_from_Wikipedia/links/0deec52e78a6e95b73000000/DBpedia-A-Large-scale-Multilingual-Knowledge-Base-Extracted-from-Wikipedia.pdf

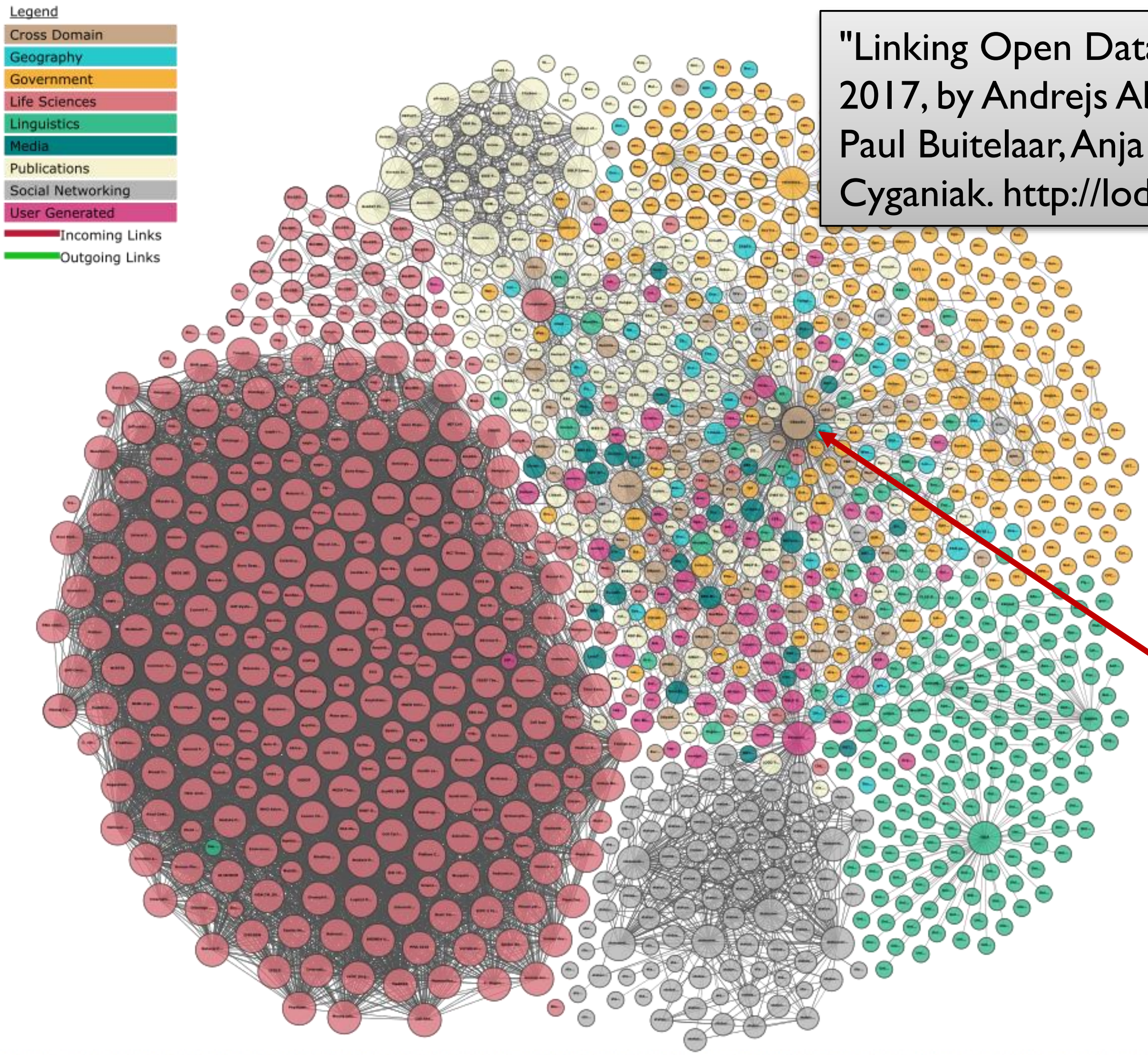
What is DBpedia?

DBpedia is a knowledge base constructed from structured information found in (different language editions) of Wikipedia.

DBpedia is broad-coverage, cross-domain, multi-lingual and includes a number of links to other datasets.

Because of the many incoming links from other datasets, DBpedia has become the central hub of the LOD cloud.

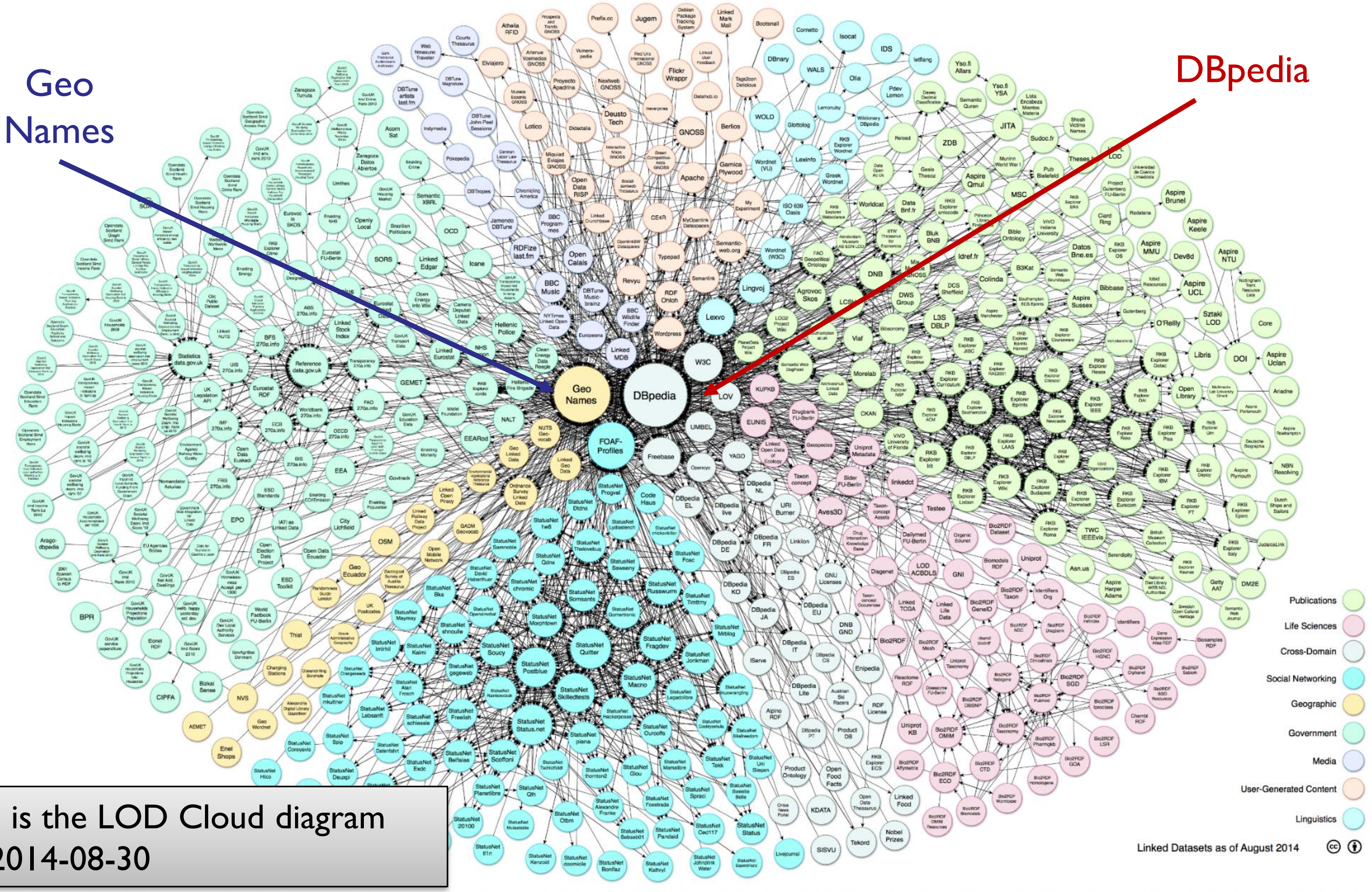
A hub for the LOD



"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

DBpedia

A hub for the LOD (cont'd)



Geo Names

DBpedia

This is the LOD Cloud diagram on 2014-08-30

DBpedia access mechanisms

- Dereferenceable URIs (with content negotiation)

  http://dbpedia.org/resource/Rome may redirect to:

  - http://dbpedia.org/page/Rome (HTML page)
  - http://dbpedia.org/data/Rome.xml (RDF/XML)
  - http://dbpedia.org/data/Rome.ttl (Turtle)
  - http://dbpedia.org/data/Rome.nt (N-Triples)
- SPARQL 1.1 Endpoint (http://dbpedia.org/sparql)
- Faceted Search (http://dbpedia.org/fct/)
- Downloads (http://wiki.dbpedia.org/develop/datasets), including the DBpedia ontology!

Structure of DBpedia

- DBpedia is an RDF dataset with an associated OWL ontology (the DBpedia ontology)
- Main namespaces:
  - http://dbpedia.org/resource/ (e.g. http://dbpedia.org/resource/Rome)
  - http://dbpedia.org/ontology/ (e.g. http://dbpedia.org/ontology/elevation)
  - http://dbpedia.org/property/ (e.g. http://dbpedia.org/property/latd)
- Its VoID dataset is http://dbpedia.org/void/Dataset

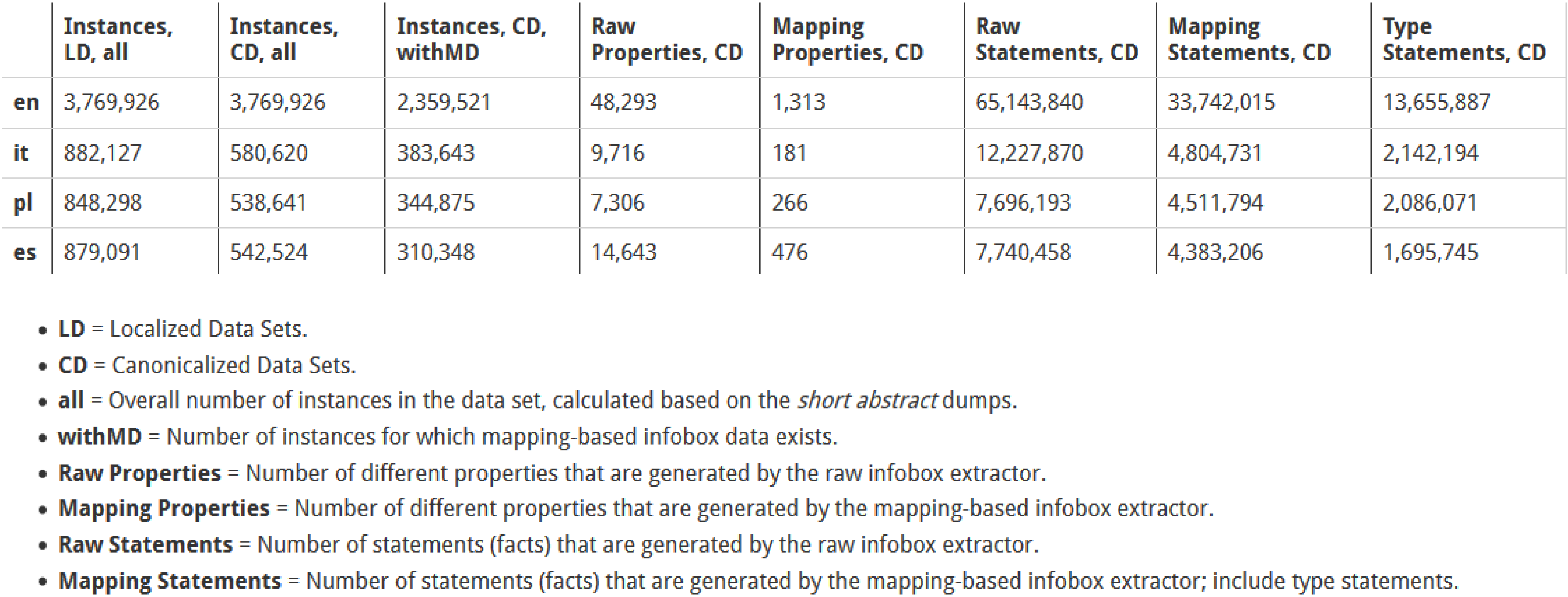
Structure of Dbpedia (cont'd)

- The DBpedia ontology used by the current version of DBpedia (2016-10) can be found among the downloads: http://downloads.dbpedia.org/2016-10/dbpedia_2016-10.owl
- The DBpedia ontology is edited via the *mappings server* using a Wiki-style interface. The current snapshot of the ontology can be accessed here: http://mappings.dbpedia.org/server/ontology/

DBpedia in figures

An excerpt of:

http://wiki.dbpedia.org/services-resources/datasets/data-set-38/data-set-statistics



	Instances, LD, all	Instances, CD, all	Instances, CD, withMD	Raw Properties, CD	Mapping Properties, CD	Raw Statements, CD	Mapping Statements, CD	Type Statements, CD
en	3,769,926	3,769,926	2,359,521	48,293	1,313	65,143,840	33,742,015	13,655,887
it	882,127	580,620	383,643	9,716	181	12,227,870	4,804,731	2,142,194
pl	848,298	538,641	344,875	7,306	266	7,696,193	4,511,794	2,086,071
es	879,091	542,524	310,348	14,643	476	7,740,458	4,383,206	1,695,745

- **LD** = Localized Data Sets.
- **CD** = Canonicalized Data Sets.
- **all** = Overall number of instances in the data set, calculated based on the *short abstract* dumps.
- **withMD** = Number of instances for which mapping-based infobox data exists.
- **Raw Properties** = Number of different properties that are generated by the raw infobox extractor.
- **Mapping Properties** = Number of different properties that are generated by the mapping-based infobox extractor.
- **Raw Statements** = Number of statements (facts) that are generated by the raw infobox extractor.
- **Mapping Statements** = Number of statements (facts) that are generated by the mapping-based infobox extractor; include type statements.

DBpedia Ontology in figures

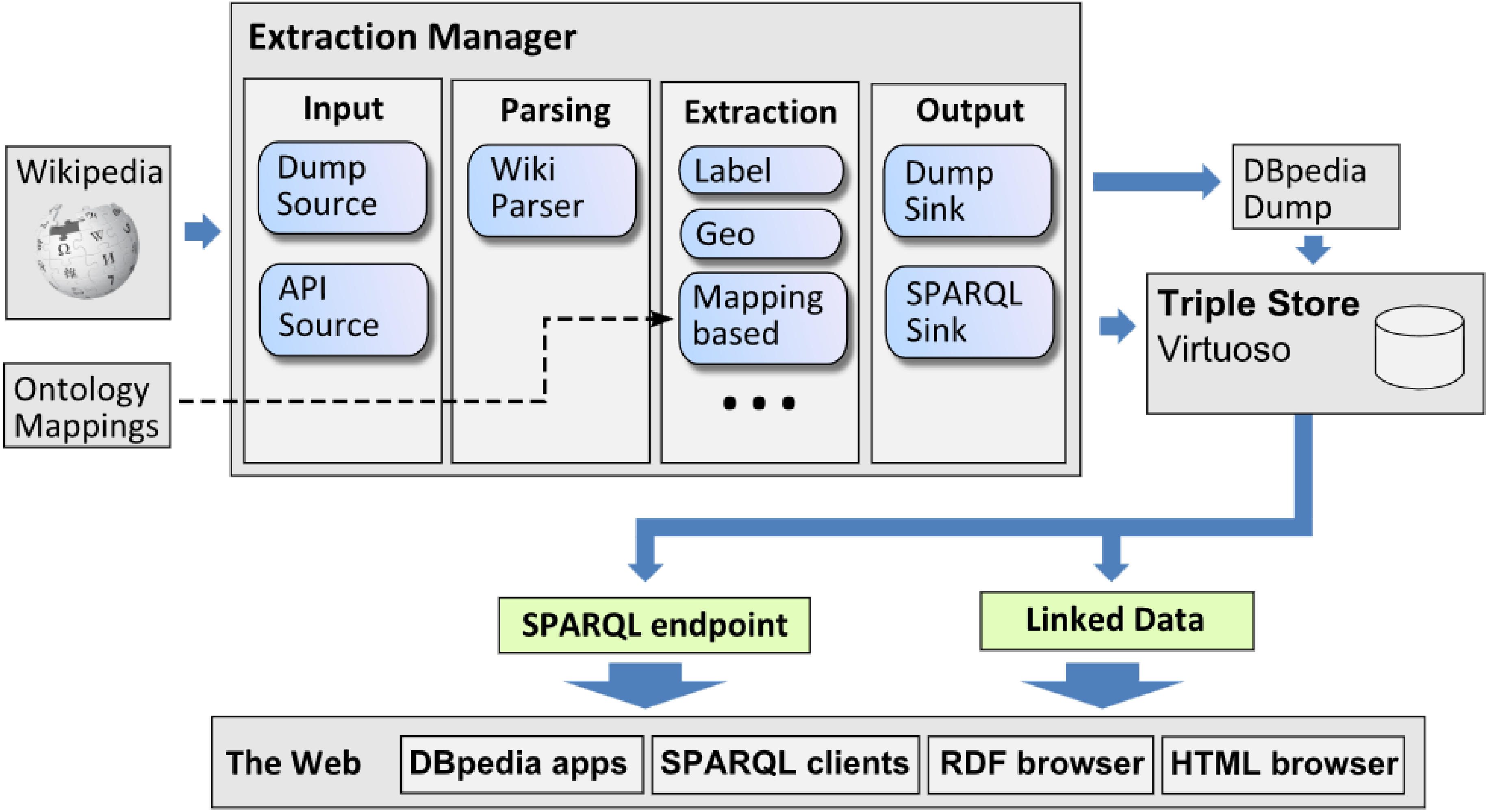
The following statistics were computed on DBpedia ontology (2016-10).

Number of classes	760
Root classes	50
Max number of (direct) subclasses	50 (dbo:Person)
Leaves classes	603 (79% of total)
Avg number of (direct) subclasses (restricted to classes with at least one child)	4.5
Number of object properties	1105
Number of datatype properties	1760

Structured information in Wikipedia

- infobox templates
- categorisation information
- images
- geo-coordinates
- links to external web pages
- disambiguation pages
- redirects between pages
- links across different language editions of Wikipedia

Extraction Framework



Source (Lehmann et al, 2012)

Raw Infobox Extraction

```
{{Infobox automobile
| name = Ford GT40
| manufacturer = [[Ford Advanced Vehicles]]
| production = 1964-1969
| engine = 4181cc
(...)
}}
```

The extractor attempts to interpret literals orderly as:

- dates,
- coordinates,
- numbers,
- links,
- strings (as default)

```
dbr:Ford_GT40
    dbp:name "Ford GT40"@en;
    dbp:manufacturer dbr:Ford_Advanced_Vehicles;
    dbp:engine 4181;
    dbp:production 1964;
    (...).
```

Non deterministic datatype assignment per property

Example taken from (Lehmann et al, 2012)

Limitations of Raw Infobox Extraction

The Raw Infobox Extraction suffers from several limitations:

- Non deterministic assignment of datatypes to properties
- No type information is generated
- Use of different templates, properties or conventions in representing property values produce different results
- Each language edition of DBpedia uses its own set of raw properties

Mapping-Based Infobox Extraction

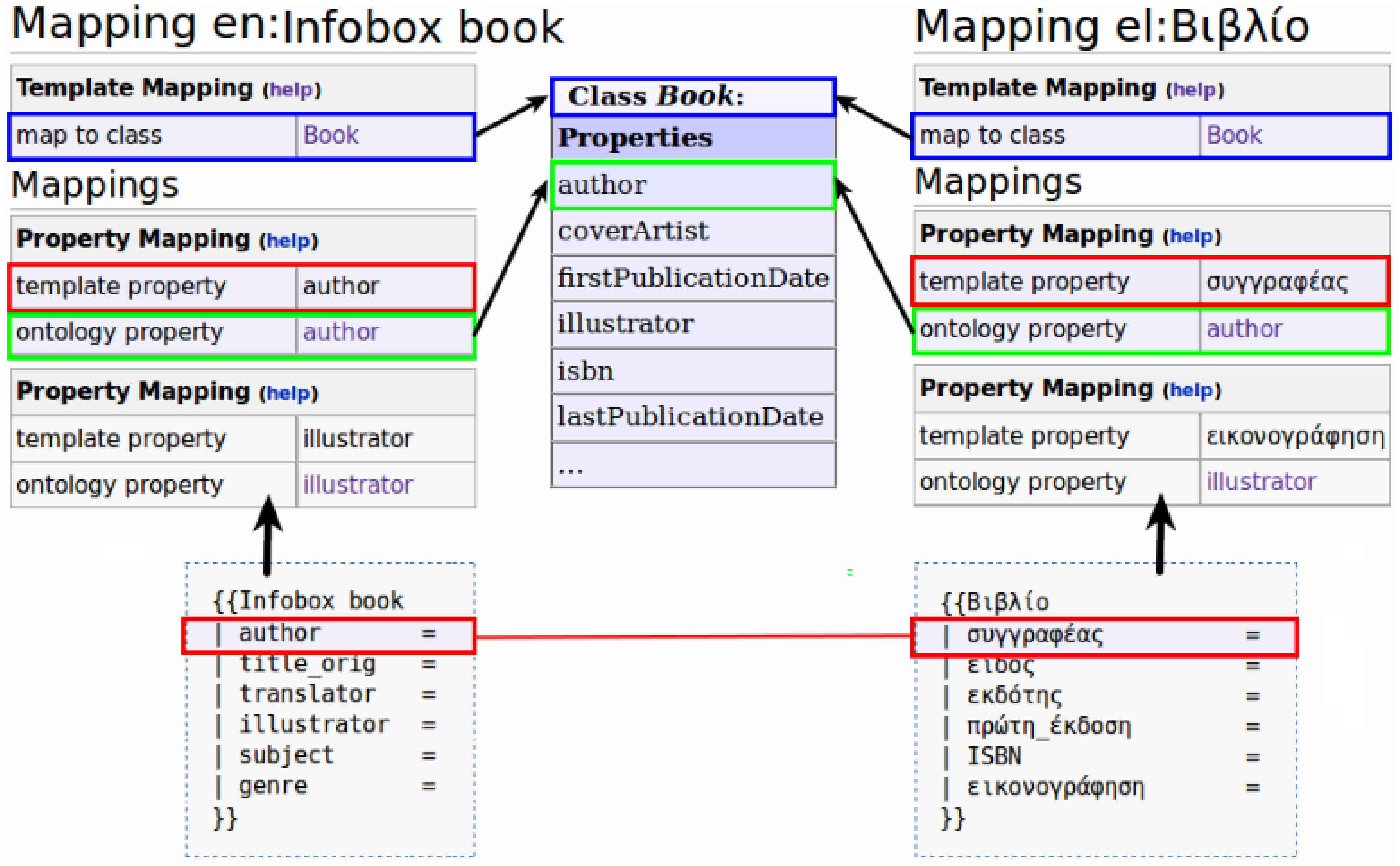
Extraction is guided by mapping of *infoboxes* to triples conforming to the *DBpedia ontology*.

Mappings are expressed using the Mediawiki Template Language, and edited through the Mappings Server.

Mappings can:

- Standardize units (e.g. covert every volume to m^3)
- Break down complex values (e.g. an interval into a start and end dates)
- Homogenize different property names
- Add types

Mapping-Based Infobox Extraction (cont’d)



Mapping en:Infobox book

Template Mapping (help)	
map to class	Book

Mappings

Property Mapping (help)	
template property	author
ontology property	author

Property Mapping (help)	
template property	illustrator
ontology property	illustrator

```
{{Infobox book
| author        =
| title_orig    =
| translator    =
| illustrator   =
| subject       =
| genre         =
}}
```

Class *Book*:
Properties
author
coverArtist
firstPublicationDate
illustrator
isbn
lastPublicationDate
…

Mapping el:Βιβλίο

Template Mapping (help)	
map to class	Book

Mappings

Property Mapping (help)	
template property	συγγραφέας
ontology property	author

Property Mapping (help)	
template property	εικονογράφηση
ontology property	illustrator

```
{{Βιβλίο
| συγγραφέας        =
| είδος             =
| εκδότης           =
| πρώτη_έκδοση      =
| ISBN              =
| εικονογράφηση     =
}}
```

Picture taken from (Lehmann et al, 2012)

DBpedia Live

- It is a framework for the continuous triplification of Wikipedia as changes occur
- It uses the OAI-PMH to get a stream of updates from Wikipedia and the mappings server, so that extractors can be execute again intelligently
  - Only process pages affected by a mapping update
  - Only process modified pages
- The result is a changeset consisting of triple additions and deletions