DBpedia

Manuel Fiorelli

fiorelli@info.uniroma2.it

Note

Questo pacco di slide contiene esempi e figure dal seguente articolo pubblicato dal Semantic Web journal:

> 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.

Una copia gratuita può essere recuperata dal sito della rivista:

> http://www.semantic-web-journal.net/system/files/swj558.pdf

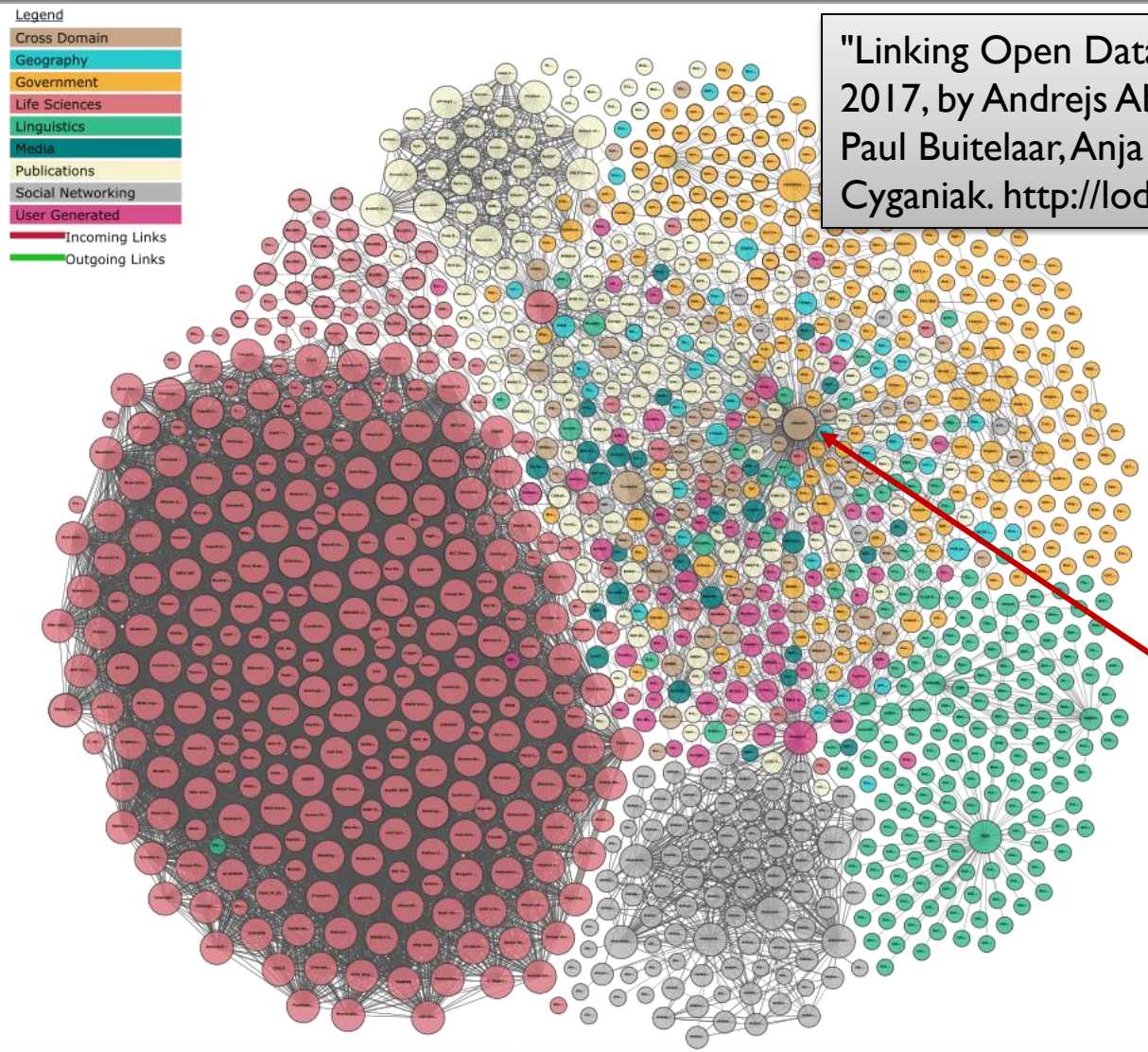
Cos'è DBpedia?

DBpedia è una base di conoscenza costruita a partire da informazione strutturata trovata nelle diverse edizioni di Wikipedia (in Italiano, in Inglese, etc.)

DBpedia è caratterizzata da un'ampiezza di contenuti, in diversi domini e linguaggi, ed include numerosi collegamenti ad altri dataset..

In aggiunta, molti altri dataset contengono link a DBpedia, che è diventata un hub della LOD cloud.

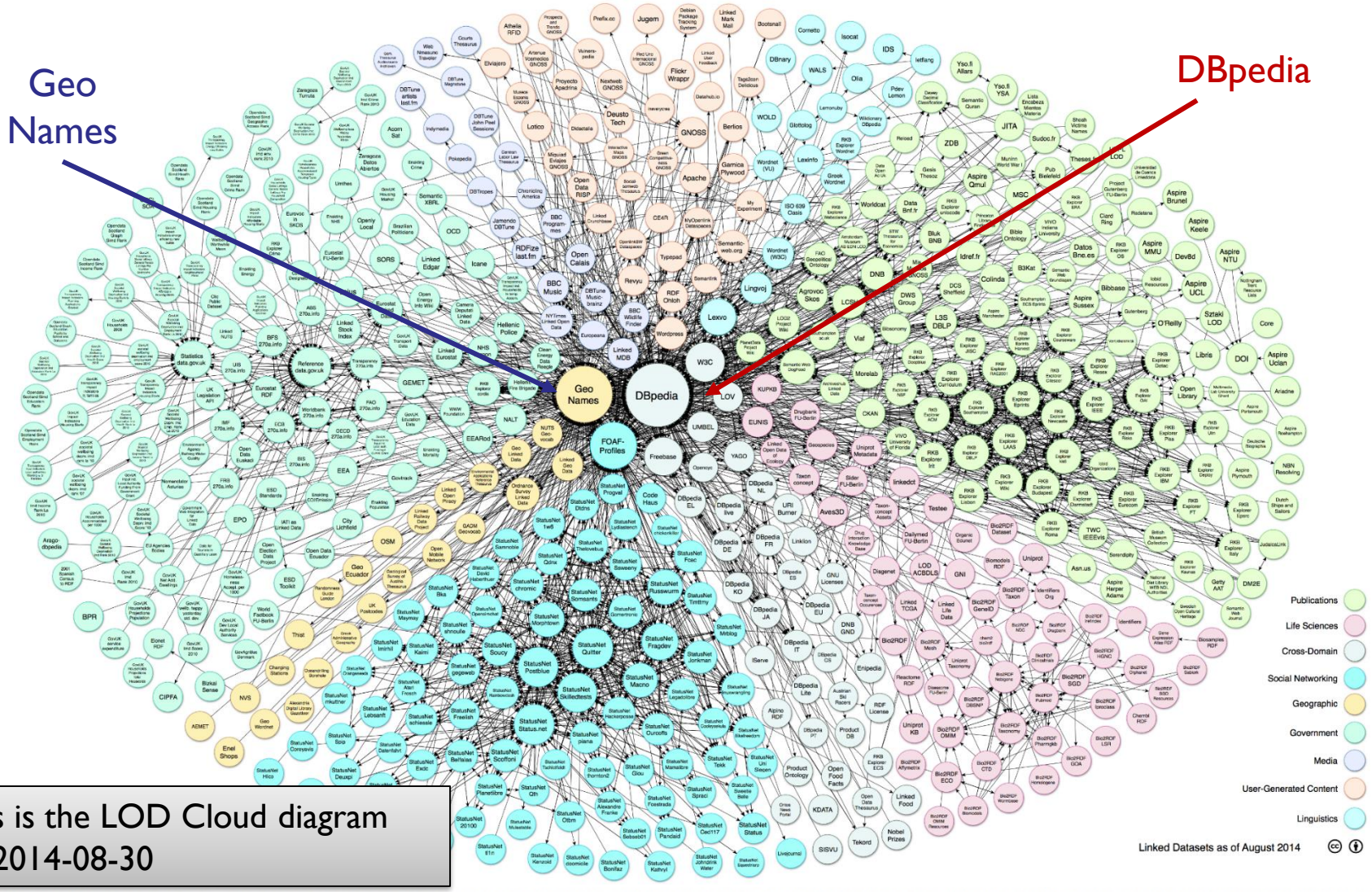
Un hub del LOD



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

DBpedia

Un hub del LOD (cont)



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

Manuel Fiorelli fiorelli@info.uniroma2.it
http://art.uniroma2.it/fiorelli

DBpedia: meccanismi di accesso

- URIs dereferenziabili (con content negotiation)

  http://dbpedia.org/resource/Rome può reindirizzare a:

  - 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/)
- Download (http://wiki.dbpedia.org/develop/datasets), inclusa la DBpedia ontology!

Struttura di DBpedia

- DBpedia è un dataset RDF con un'ontologia associata (la DBpedia ontology)
- Principali namespace:
  - 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)
- Il suo VoID dataset è http://dbpedia.org/void/Dataset

Struttura di Dbpedia (cont)

- La DBpedia ontology usata dalla versione corrente di DBpedia (2016-10) può essere trovata tra i download: http://downloads.dbpedia.org/2016-10/dbpedia_2016-10.owl
- La DBpedia ontology è editata attraverso il *mappings server* usando una interfaccia di tipo Wiki.
  Lo snapshot corrente dalla Dbpedia ontology può essere trovato a questo indirizzo: http://mappings.dbpedia.org/server/ontology/

DBpedia in cifre

Un estratto di

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 cifre

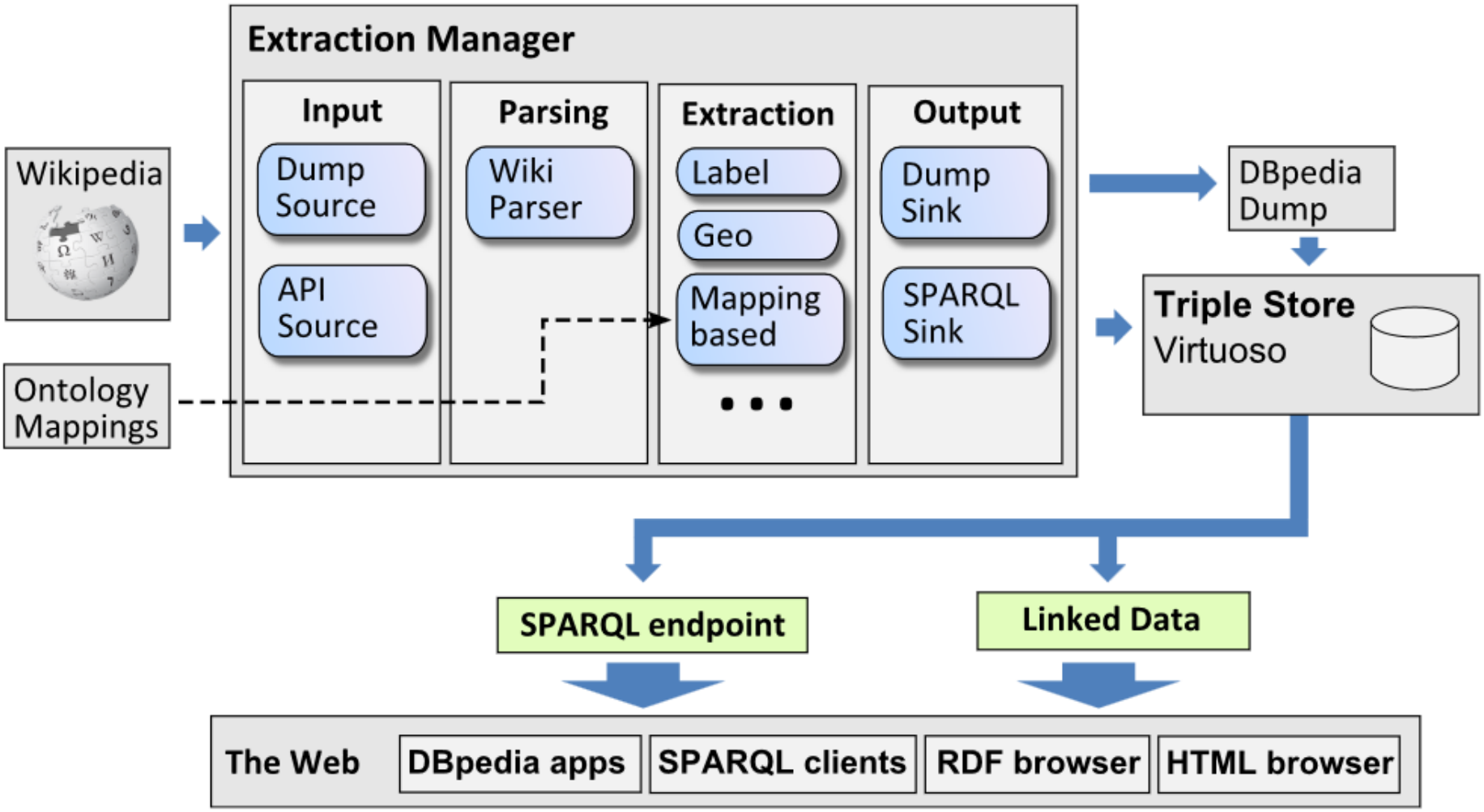
Alcune statistiche sulla DBpedia ontology (2016-10).

Numero di classi	760
Radici dell'albero delle classi	50
Numero massimo di sottoclassi (dirette)	50 (dbo:Person)
Foglie dell'albero delle classi	603 (79% del totale)
Numero medio di sottoclassi (dirette), ristretto alle classi con almeno un figlio	4.5
Numero di object property	1105
Numero di datatype property	1760

Informazione strutturata in Wikipedia

- infobox template
- categorie
- immagini
- coordinate geografiche
- collegamenti a pagine web esterne
- pagine di disambiguazione
- reindirizzamento tra pagine
- collegamenti tra diversi edizioni di Wikipedia (nelle diverse lingue)

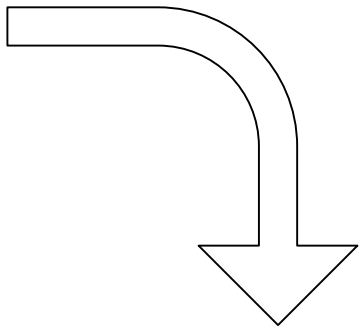
Framework di estrazione



Fonte (Lehmann et al, 2012)

Raw Infobox Extraction

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



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

L'estrattore prova ad interpretare i literal nell'ordine come:

- date,
- coordinate,
- numeri,
- collegamenti,
- stringhe

Assegnazione non deterministica del tipo di dato dei valori di una proprietà

Esempio preso da (Lehmann et al, 2012)

Limiti della Raw Infobox Extraction

La Raw Infobox Extraction soffre di alcune limiti:

- Assegnazione non deterministica del tipo di dato alle proprietà
- Mancata assegnazione di un tipo alla risorsa soggetto
- L'uso di template, proprietà o convenzioni differenti per la rappresentazione dei valori delle proprietà produce risultati differenti
- Ogni edizione di DBpedia (nelle diverse lingue) usa il proprio insieme di raw property

Mapping-Based Infobox Extraction

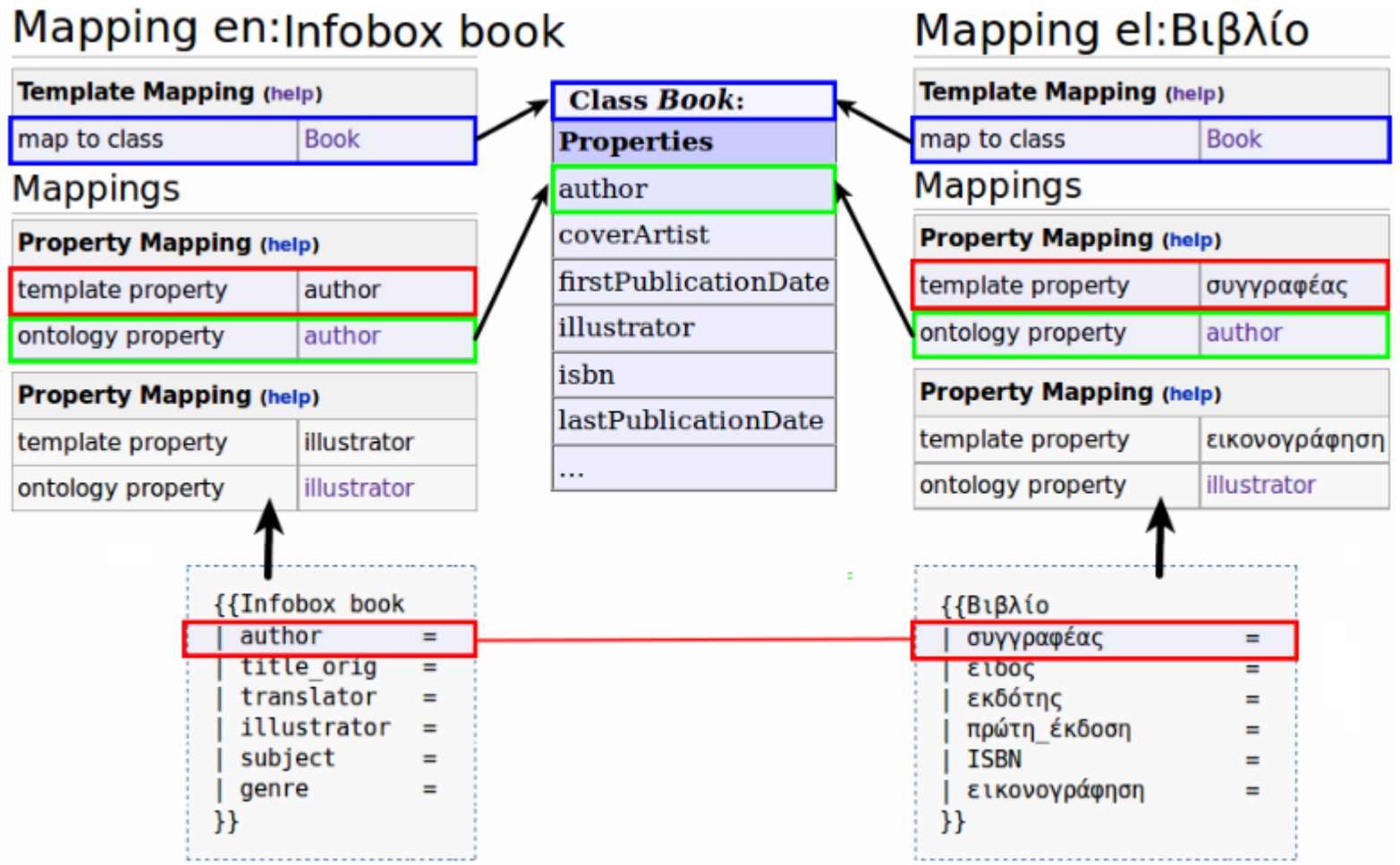
L'estrazione è guidata da *mapping* delle *infobox* in triple conformi alla *DBpedia ontology*.

I mapping sono espressi usando il linguaggi di template di Mediawiki, e sono editati attraverso il Mappings Server.

I mapping possono:

- Standardizzare le unità (es. convertire tutti i volumi in m^3)
- Scomporre valori complessi (es. un intervallo può essere scomposto nei sui istanti di inizio e fine)
- Omogenizzare i nomi delle proprietà
- Aggiungere tipi

Mapping-Based Infobox Extraction (cont)



Disegno preso da (Lehmann et al, 2012)

DBpedia Live

- È il framework per la triplificazione continua di Wikipedia come avvengono dei cambiamenti
- Usa il protocollo OAI-PMH per ricevere il flusso degli aggiornamenti da Wikipedia e dal mappings server, così da poter rieseguire gli estrattori in modo intelligente (incrementale), processando esclusivamente:
  - Le pagine che utilizzano un infobox template il cui mapping è stato aggiornato
  - Le pagine che sono state modificate
- Il risultato è un changeset che consiste delle triple aggiunte e rimosse

Collegamenti ad altri dataset

Un progetto su Github permette di caricare linkset da DBpedia verso altri dataset:

https://github.com/dbpedia/links

Va sottolineato che il caricamento di file statici viene scoraggiato; in realtà, si suggerisce di inviare un link per il download, una query su uno SPARQL ednpoint, uno script, etc. in modo che i link possano essere rigenerati (secondo una frequenza indicata)

Dopo un controllo di qualità, questi linkset saranno integrati in DBpedia.