DBPedia

G. Falquet

April 6, 2014

Source

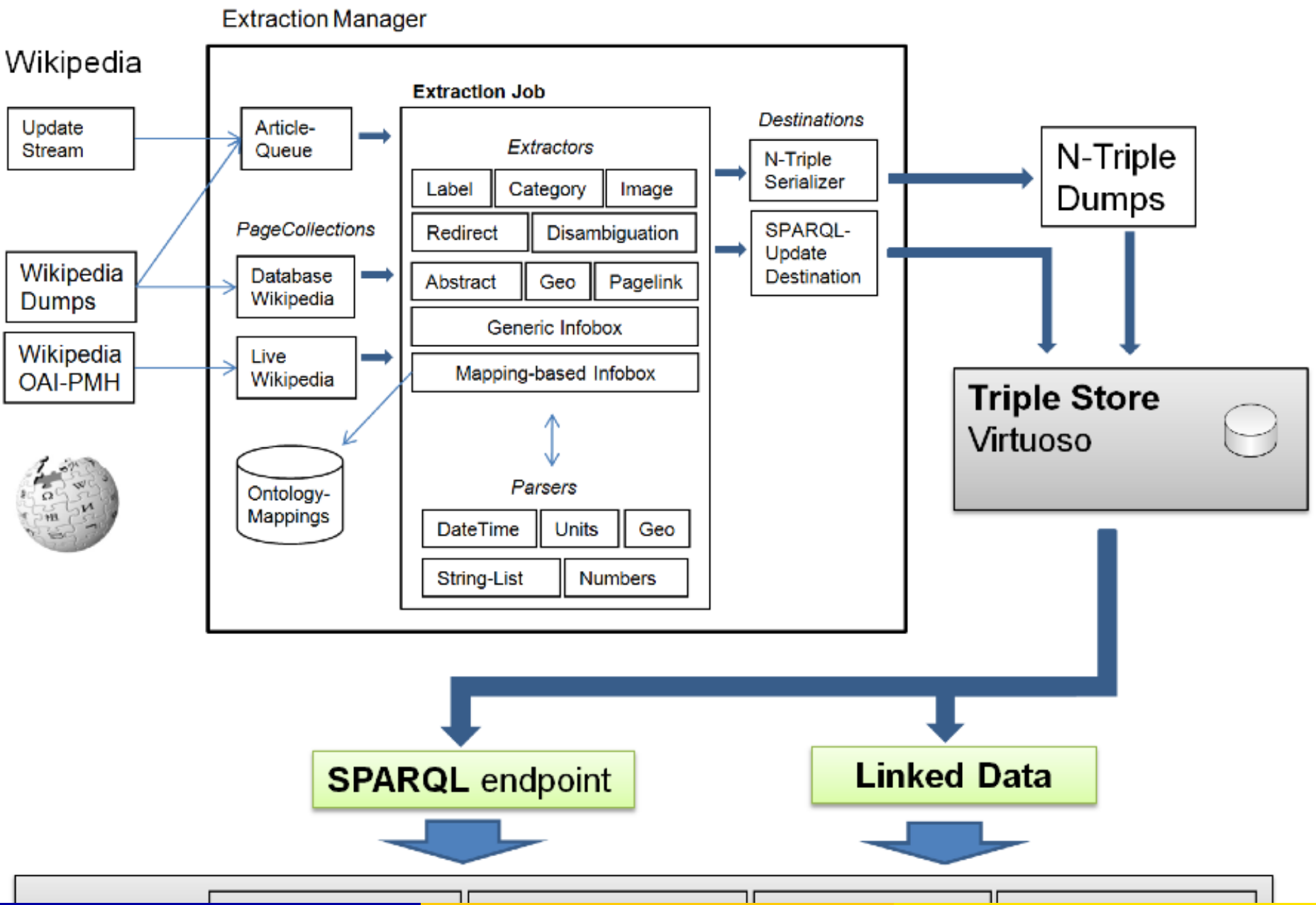
Christian Bizer , Jens Lehmann , Georgi Kobilarov , Soren Auer , Christian Becker , Richard Cyganiak, Sebastian Hellmann (2009) DBpedia - A Crystallization Point for the Web of Data Christian. Journal of Web Semantics, Volume 7, Issue 3, September 2009, Pages 154–165.

DBPedia

Extracting structured information from Wikipedia and making this information accessible on the Web.
The resulting DBpedia knowledge base (2009) describes more than 2.6 million entities,

- 198,000 persons,
- 328,000 places,
- 101,000 musical works,
- 34,000 films, and
- 20,000 companies.
- 3.1 million links to external web pages;
- 4.9 million RDF links into other Web data sources.

Knowledge Extraction



Extractors

Labels. article titles → rdfs:label for the corresponding DBpedia resource.

Abstracts. first paragraph → rdfs:comment ; text before a table of contents, at most 500 words → dbpedia:abstract

Interlanguage links. links that connect articles about the same topic in different language editions of Wikipedia → labels and abstracts in different languages to DBpedia resources.

Images. Links pointing at Wikimedia Commons images depicting a resource → foaf:depiction property.

Redirects. to identify synonymous terms, extract the redirects and use them to resolve references between DBpedia resources.

(cont.)

Disambiguation. Wikipedia disambiguation pages explain the different meanings of homonyms. → predicate dbpedia:disambiguates .

External links. references to external Web resources → dbpedia:reference .

Pagelinks. links between Wikipedia articles → dbpedia:wikilink property.

Homepages. obtain links to the homepages of entities by looking for the terms *homepage* or *website* within article links (represented using foaf:homepage).

Categories. Categories become skos:concepts ; category relations are represented using skos:broader .

Geo-coordinates. The geo-extractor expresses coordinates using the Basic Geo (WGS84 lat/long) Vocabulary and the GeoRSS Simple encoding of the W3C Geospatial Vocabulary .

Live Extraction

- the Wikipedia OAI-PMH live feed that instantly reports all Wikipedia changes.
- the live extraction work ow uses this update stream to extract new RDF whenever a Wikipedia article is changed.
- SPARQL-Update Destination deletes existing and inserts new triples into a separate triple store.
- about 1.4 article pages are updated each second on Wikipedia.
- the framework can handle up to 8.8 pages per second on a 2.4 GHz dual-core machine
  - includes consumption from the stream, extraction, diffing and loading the triples into a Virtuoso triple store.
- the time lag for DBpedia to reflect Wikipedia changes lies between one or two minutes.

Infobox extraction

```
{{Infobox Actor
| birthname      = Thomas Jeffrey Hanks
| birthdate      = {{birth date and age|1956|7|9}}
| birthplace = [[Concord, California|Concord]],
               [[California]]
| yearsactive    = 1979 - present
| occupation     = Actor, producer, director,
                   [[voice over artist]], writer,
                   speaker
```

```
{{Infobox Tennis player
|country = United States
|playername = Andre Agassi
|residence = [[Las Vegas metropolitan area|Las Vegas]],
             [[Nevada]], United States
|datebirth = {{birth date and age|mf=yes|1970|4|29}}
|placebirth = [[Las Vegas, Nevada]], United States
|height = {{convert|1.80|m|ftin|abbr=on}}
|weight = {{convert|177|lb|kg|abbr=on}}
```

- Different communities use different templates to describe the same type of things (e.g. infobox_city_japan , infobox_swiss_town and infobox_town_de).
- Different templates use different names for the same attribute (e.g. birthplace and placeofbirth).
- As many Wikipedia editors do not strictly follow the recommendations given on the page that describes a template, attribute values are expressed using a wide range of different formats and units of measurement.

Generic technique

Advantage

- Complete coverage of all infoboxes and infobox attributes.

...

Disadvantages

- Synonymous attribute names are not resolved, which makes writing queries against generic infobox data rather cumbersome.
- As Wikipedia attributes do not have explicitly defined datatypes, a further problem is the relatively high error rate of the heuristics that are used to determine the datatypes of at- tribute values.

Mapping technique

- Based on an analysis of the 350 most commonly used infobox templates within the English edition of Wikipedia.
- Map 2350 attributes from within these templates to 720 properties.
  - define fine-grained rules on how to parse infobox values
  - define 55 target datatypes, which help the parsers to process attribute values.

Disadvantage

- it currently covers only 350 Wikipedia templates;
- therefore it only provides data about 843,000 entities compared to 1,462,000 entities that are covered by the generic approach

Content (2009)

The DBpedia knowledge base currently consists of around

- 274 million RDF triples, which have been extracted from the English, German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish, Norwegian, Catalan, Ukranian, Turkish, Czech, Hungarian, Romanian, Volapuk, Esperanto, Danish, Slovak, Indonesian, Arabic, Korean, Hebrew, Lithuanian, Vietnamese, Slovenian, Serbian, Bulgarian, Estonian, and Welsh versions of Wikipedia.
- The knowledge base describes more than 2.6 million entities.
- It features labels and short abstracts in 30 different languages;
- 609,000 links to images;
- 3,150,000 links to external web pages;
- 415,000 Wikipedia categories,
- 286,000 YAGO categories

Classifying Entities

Four classification schemes

Wikipedia categories

YAGO classes

UMBEL. The Upper Mapping and Binding Exchange Layer

DBpedia Ontology. The DBpedia ontology consists of 170 classes that form a shallow subsumption hierarchy. It includes 720 properties with domain and range definitions. The ontology was manually created from the most commonly used infobox templates within the English edition of Wikipedia