ARISTOTLE UNIVERSITY OF THESSALONIKI

DEPARTMENT OF INFORMATICS

# DBpedia citations & references challenge

| | | |
|---|---|---|
| Project | : | Mapping DBpedia's citations to existing bibliographical data |
| Editor | : | David Nazarian, Mr. |
| Supervisor | : | Nick Bassiliades, Dr., Associate Professor |

In this report, a brief overview is given about a project that aims to map the DBpedia's citations to existing bibliographical data. The project is part of an MSc's degree thesis that will be submitted in January 2017. A number of properties of the *enwiki-20160305-citation-data.ttl* file have been used in order to facilitate the linking of the triples' subjects (found in the file) to URIs from other bibliographical sources. As a result, a total of 1,084,445 links were discovered, with 761,235 corresponding to distinct subjects, covering 6% of the entire file. Emphasis has been given to the properties that represent identifiers, that can be found in other data sources and are relatively common. In particular, the properties *isbn*, *issn*, *doi*, *journal*, *series*, *periodical, magazine*, *oclc, pmid, arxiv and lccn* have been used combined with the *title* and *year*. The linking of the data has been based on a number of LOD dumps that are available for download and bibliographical websites that provide their metadata through APIs. The project comprises of an application written in Java that processes and links the data and a triplestore[1] which stores the original and the processed data. For example, in order to create links based on the isbn property, the following steps are taken:

1) Hyphens and extra characters are removed from the ISBN strings
2) The resulting ISBNs are checked for validity based on the ISBN algorithm
3) All the ISBN10s are converted to their equivalent ISBN13 form for uniformity
4) The year property (when exists) is filtered from any extra characters
5) The title property is filtered based on a list of symbols and user-defined strings
6) All the processed data are saved in a separate repository
7) A link is created when ISBNs are joined successfully from two sources, the absolute difference between their year properties is ≤ 1 year and the titles have a similarity ≥ 80% based on overlap string similarity metric configured for 2-shingles. In cases when the year property is absent, the titles must have a similarity ≥ 90%.

Similar steps are taken for the other properties too. For the issn, journal, series, periodical and magazine properties, the title similarity is checked based on the Dice coefficient with a stricter threshold (≥ 97%) and the year properties must match.

The following data sources have been used in the project:

| Data source | Type | Unique triples in local data dump |
|---|---|---|
| DBpedia citations | Data dump | 76.2M |
| DBLP - Digital Bibliography & Library Project | Data dump | 88.1M |
| BNB - British National Bibliography | Data dump | 111M |
| DNB - Deutsche Nationalbibliografie | Data dump | 231.3M |
| BNE - Biblioteca Nacional de España | Data dump | 68.7M |
| Springer | Data dump | 3.3M |
| WorldCat | API | 2.5M |
| PubMed | API | 1M |

[1] Ontotext GraphDB Free 7

| | | |
|---|---|---|
| arXiv | API | 0.04M |
| Open Library | API | 1.1M |
| HathiTrust | API | 0.33M |

The complete data dumps of WorldCat, PubMed, arXiv, Open Library and HathiTrust are not available; instead specific data have been downloaded in local repositories by utilizing their APIs.

The *enwiki-20160305-citation-data.ttl* file contains 76,223,926 unique triples with 12,391,363 distinct subjects. The queries that have been used to extract information from the repository containing the data of the file are described below:

| Query | Subject count | Distinct subject count | Identifier |
|---|---|---|---|
| subjects that have the isbn property, the title property, optionally the year property and don't refer to a specific chapter | 712,834 | 594,244 | isbn |
| subjects that have the journal or series or periodical or magazine property combined with the year and title properties and don't refer to a specific chapter | 1,072,592 | 542,923 | journalTitle |
| subjects that have the issn, year and title properties and don't refer to a specific chapter | 98,264 | 24,434 | issn |
| subjects that have the doi and title properties, optionally the year property and don't refer to a specific chapter | 629,571 | 576,530 | doi |
| subjects that have the oclc and title properties, optionally the year property and don't refer to a specific chapter | 124,082 | 43,483 | oclc |
| subjects that have the pmid and title properties, optionally the year property and don't refer to a specific chapter | 379,274 | 346,613 | pmid |
| subjects that have the arxiv and title properties, optionally the year property and don't refer to a specific chapter | 14,986 | 13,373 | arxivID |
| subjects that have the lccn and title properties, optionally the year property and don't refer to a specific chapter | 17,699 | 4,873 | lccn |

The eight query categories together refer to a total of 1,380,387 distinct subjects since there are overlaps between them.

The results found in the project correspond to 761,235 / 1,380,387 = 55.15% of the distinct subjects extracted and to 761,235 / 12,391,363 = 6.14% of the entire file. The following table contains the number of links found for each category and data dump.

| Identifier | Destination | Link count | File |
|---|---|---|---|
| isbn | BNB | 197,405 | dbpedia_to_bnb_isbn_links.nt.zip |
| | BNE | 10,584 | dbpedia_to_bne_isbn_links.nt.zip |
| | DBLP | 2,232 | dbpedia_to_dblp_isbn_links.nt.zip |
| | DNB | 23,442 | dbpedia_to_dnb_isbn_links.nt.zip |
| | Springer | 2,191 | dbpedia_to_springer_isbn_links.nt.zip |
| | Open Library | 393,276 | dbpedia_to_openlibrary_isbn_links.nt.zip |
| | HathiTrust | 101,473 | dbpedia_to_hathitrust_isbn_links.nt.zip |
| journalTitle | BNB | 3,674 | dbpedia_to_bnb_journal_links.nt.zip |
| | DBLP | 4,989 | dbpedia_to_dblp_journal_links.nt.zip |
| | DNB | 1,076 | dbpedia_to_dnb_journal_links.nt.zip |
| issn | BNB | 64 | dbpedia_to_bnb_issn_links.nt.zip |
| | DNB | 19 | dbpedia_to_dnb_issn_links.nt.zip |
| | BNE | 3 | dbpedia_to_bne_issn_links.nt.zip |
| doi | DBLP | 14,824 | dbpedia_to_dblp_doi_links.nt.zip |
| | Springer | 676 | dbpedia_to_springer_doi_links.nt.zip |
| oclc | DNB | 143 | dbpedia_to_dnb_oclc_links.nt.zip |
| | WorldCat | 15,386 | dbpedia_to_worldcat_oclc_links.nt.zip |
| | Open Library | 8,504 | dbpedia_to_openlibrary_oclc_links.nt.zip |
| | HathiTrust | 10,531 | dbpedia_to_hathitrust_oclc_links.nt.zip |
| pmid | PubMed | 293,689 | dbpedia_to_pubmed_pmid_links.nt.zip |
| arxivID | arXiv | 10,891 | dbpedia_to_arxiv_arxivid_links.nt.zip |
| lccn | Open Library | 2,967 | dbpedia_to_openlibrary_lccn_links.nt.zip |
| | HathiTrust | 1,578 | dbpedia_to_hathitrust_lccn_links.nt.zip |

The triples from all the files can be found in the dbpedia_combined_links.nt.zip file.
Also the triples can be queried from the following GraphDB Free SPARQL endpoint:
http://lod.csd.auth.gr:7200/sparql

Even though the deadline for the DBpedia citations & references challenge has passed we would be grateful for your feedback about the project.