FREME Research Corpora: ORCID and CORDIS

Milan Dojchinovski[1,3], Martin Brümmer[1], Robert Roessling[1], Felix Sasaki[2]

[1]Agile Knowledge Engineering and Semantic Web (AKSW)
Institute for Applied Informatics, Germany
firstname.lastname@informatik.uni-leipzig.de

[2] Language Technology Lab
German Research Center for Artificial Intelligence (DFKI) - W3C Fellow, Germany
firstname.lastname@dfki.de

[3] Web Intelligence Research Group
Faculty of Information Technology, Czech Technical University in Prague
firstname.lastname@fit.cvut.cz

Abstract

TODO

Keywords: research corpora, linked data, enrichment, named entities

1. Introduction

In the last few years, the Linked Open Data (LOD) cloud attract many institutions to open and publish their data as Linked Open Data. According to (Schmachtenberg et al., 2014), from only 299 Linked Data datasets published in September 2011, the Linked Open Cloud has increased to 1024 published Linked Data datasets. Although datasets from various domains have been published, there are still many other datasets of high value for specific use cases.
In this paper, we focus datasets from the research domain. In particular, we look at the CORDIS and ORCID datasets.

TODO: to be further extended

2. Requirements for Research Corpora

TODO: requirements and motivation

3. The ORCID Corpus

ORCID[1] (Open Researcher and Contributor ID) is an open, non–profit, community effort to provide a registry of unique identifiers for researchers and related activities and results. A researcher can register, provide his professional information such as fundings, and link them to works such research papers and journal papers. Upon successful registration, a research obtains a unique ORCID identifier.

3.1. The Data Source

On an annual basis ORCID publishes all this these valuable information as dump. The dump is published under a CC0 1.0 Public Domain Dedication waiver for free download[2]. In our work, we take up the latest ORCID dump from 2014. It contains over 1.3 million of JSON files amounting to 41GB of data.

3.2. Conversion

We processed each JSON file which describes a single person researcher. For each person we extracted the unique ORCID identifier, their general personal information and works the person was involved in. As general information we extracted their name and other alternative names the person is known as, other identifiers such as Scopus, ResearcherID or LinkedIn, their short biography, work location, and email.
Further, we extract the works associated with each person. For each work we capture its title, type, date of publication, external identifiers and the contributors to the work. Following list of external identifiers are consider: DOI, ISSN, ISBN, LCCN, OCLC, PMID and ASIN. Also, if the work contributor is identified with an ORCID ID, then we link the work with the the contributor. Otherwise, the contributor is encoded as literal.
Note that not always all the information described above is available, therefore we only convert the available information.
The RDF model used to capture the available information builds on top of existing and well established ontologies. We use the FOAF[3] ontology to describe researchers, the Schema.org[4] vocabulary to describe places, the Bibliographic ontology[5] to describe bibliographic information such as articles, books, chapters, thesis, reports, manuals and other type of documents, and terms from the Dublin Core[6] vocabulary to describe some metadata information such as title, date of creation or associate person with a work.
The converted dataset contains 126 millions of RDF triples with a total size of 13GB in the N-Triples RDF serialization format. Table 1 provides first order statistics for the dataset.

[1] http://orcid.org/

[2] http://support.orcid.org/knowledgebase/articles/223698-how-do-i-get-a-public-data-file-

[3] http://xmlns.com/foaf/spec/

[4] http://schema.org/

[5] http://bibliontology.com/

[6] http://dublincore.org/

Type of record	Counts
Persons	1.369.762
Places	11.875
Articles	5.498.470
Books	143.971
Chapters	103.583
Reports	9.176
Manuals	582
Thesis	7.241
Documents	497.999
Total	7.642.659

Table 1: Size metrics for the ORCID dataset.

Availability and License. The dump of the dataset is available for download[7] and it can be also queried at via a Linked Fragments Endpoint at `http://rv2622.1blu.de:5000/orcid`. We publish the converted dataset under a CC0 1.0 Public Domain Dedication free license.

4. The CORDIS Corpus

CORDIS[8] (Community Research and Development Information Service) is the European Commission's core public repository providing dissemination information for all EU-funded research projects.

5. Conclusion

TODO: conclusion

Acknowledgement. The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644771.

6. References

Max Schmachtenberg, Heiko Paulheim, and Christian Bizer. 2014. Adoption of linked data best practices in different topical domains. In *The Semantic Web - ISWC 2014*, Lecture Notes in Computer Science. Springer Berlin Heidelberg.

[7] `http://freme.aksw.org/datasets/orcid/`

[8] `http://cordis.europa.eu/home_en.html`