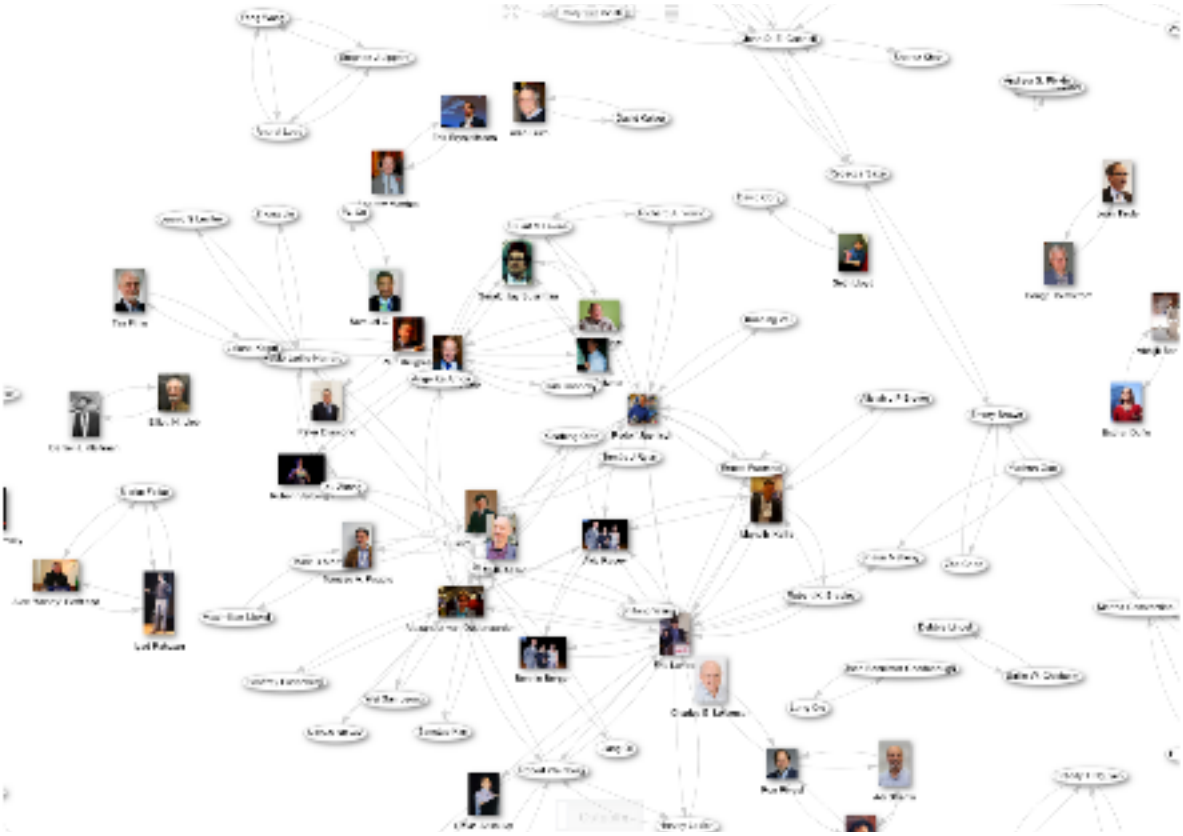

Open Climate library

21.01.2019

Human-Centered Computing
Institut für Informatik
Königin-Luise-Str. 24-26
14195 Berlin

Supervision

Prof. Dr. Claudia Müller-Birn
clmb@inf.fu-berlin.de

Collaboration

Simon David Hirsbrunner
simon.hirsbrunner@uni-siegen.de

Dr. Moritz Schubotz
schubotz@uni-wuppertal.de

Area

Data science, semantic web technologies, bibliometrics, databases

Degree

BSc./MSc. Bachelor of Science

Requirements

- Literacy in modern programming languages such as python or javascript
- Database technology, fluency in SQL and/or SPARQL
- Virtualization and webserver technology

Content

(1) Context of the project

- The work for the thesis is connected to an interdisciplinary research project by Wikimedia fellows assessing the potential of Wikidata and Wikimedia labs to be used as a source and tool environment for data scientific analysis.

- Concretely, the project aims at translating the complete bibliographic information of a scientific library on climate science (8000 entries) to the Wikidata format.
- Building on the outcome of the thesis, a team of interdisciplinary researchers together with stakeholders will experiment with existing data analysis tools such as Scholia (https://tools.wmflabs.org/scholia/) and recommend further developments of tools and applications. These stakeholders include the Wikidata, Wikicite community (https://meta.wikimedia.org/wiki/WikiCite) and the Potsdam Institute for Climate Impact Research (PIK, https://www.pik-potsdam.de/) who kindly provided the dataset with a CC0 licence (https://creativecommons.org/share-your-work/public-domain/cc0/).

(2) What is the problem

- While there are comprehensive tools available to conduct data analysis via the Wikidata Query Service and tool environments such as Scholia, the data quality in Wikidata is often not good enough to conduct such analysis beyond mere exploration. To assess the potential of Wikidata for data science and bibliographic analysis in particular, we would need a complete and meaningful dataset.
- In contrast, we have a very well documented and interesting dataset available - all bibliographic metadata of a major climate institute available, but this format is not fit for data analysis (Excel table). How can this dataset be made more FAIR (findable, accessible, interoperable and re-usable) (Wilkinson et al. 2016) via Wikidata?

(3) What are the objectives

- The objective of the bachelor project is to develop a solution for the challenges described in (2)
- To map the fields of the dataset to the Wikidata data structure;
- Develop exploratory queries via Wikidata Query Service / Scholia

(4) What is a possible procedure

- Install private Wikibase instance by using the Wikibase Docker image https://github.com/wmde/wikibase-docker;
- Set up data import pipeline with data quality tests;
- Import the provided PIK data (see description in section (1)) to this instance;
- Set up Sparql Endpoint on private instance;
- Install Scholia on private instance;
- Implement exploratory citation analysis with the existing Scholia tools;
- Carefully document all steps for reproducibility.

References

Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. "Scholia, Scientometrics and Wikidata." In *The Semantic Web: ESWC 2017 Satellite Events*, edited by Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig, 237–59. Lecture Notes in Computer Science. Springer International Publishing, 2017.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March 15, 2016): 160018. https://doi.org/10.1038/sdata.2016.18.