Scholia

Finn Årup Nielsen

Cognitive System, DTU Compute, Technical University of Denmark

28 October 2017
Can we make a Wikidata presenter tool targeted at bibliographic information?
Presenting Wikidata: Reasonator

Magnus Manske’s Reasonator, https://tools.wmflabs.org/reasonator/

Extracts information from Wikidata and makes templated (“natural language”) text, maps, timelines, fetches relevant images, formats other information nicely and adds internal and external links.

Runs from Wikimedia Toolforge.

What about citation information, aggregation of publication for an organization, etc.?
Scholia

Web site with scholarly information extracted from Wiki-data running from https://tools.wmflabs.org/scholia/.

Developed from GitHub under GPL https://github.com/fnielsen/scholia with work/input from Egon Willighagen, Daniel Mietchen, Jakob Voß, Magnus Manske, Andy Mabbett.

Almost entirely built by using Wikidata Query Service: We generate tables, bubble charts, time lines, graphs, etc.
“Aspects”

Scholia presents the data in different “aspects”: author, work, organization (e.g., university, research group), venue (journal or conference), series (e.g., conference proceedings series), publisher, sponsor, award, topic.

Researcher can be viewed as an author or a topic. University could be an organization or a publisher.
Author aspect publications per year

Inspired by Shubhanshu Mishra’s and Vetle I. Torvik’s LEGOLAS visualization.

Number of publications per year from https://tools.wmflabs.org/scholia/author/Q20980928.

Color-coding based on author-role (first author, last author, middle author, solo author)

Using default “BarChart” of Wikidata Query Service with complex SPARQL query: https://query.wikidata.org/#%23defaultView...
Work aspect citation graph

Citation panel on work aspect for partial citation graph with WDQS's Graph output.

For a principal component analysis of 39 scientific impact measures paper.

Network constructed with the BlazeGraph’s RDF GAS API for graph queries.
Overview of number of papers published and their citations across journals published by the publisher with Scatter output from WDQS.

Here for BioMedCentral (which may be an imprint):

https://tools.wmflabs.org/scholia/publisher/Q463494
Incomplete statistics on page production per year for DTU Cognitive Systems. Yet another color-coded WDQS Bar chart.
Award aspect

Turing award locations

The points are colored according to the “layer” column in the SPARQL result.

https://tools.wmflabs.org/scholia/award/Q185667
Multiple items

Aspects for multiple items: two or more works, two or more authors, two or more publishers.

Comparison of authors, universities, etc.

For instance, two works works/Q20900776,Q25938983
Software aspect

There are several topic-related aspects: disease, protein, gene, chemical, biological pathway, software.

Software aspect: Here the panel with Top-ics of works using the software queried from P2283 with SPM.

Based on work by Katherine Thornton: Wikidata for digital preservation: A Wikidata Portal.
Scholia tool: arxiv-to-quickstatement

Conversion of arXiv identifier to Magnus Manske’s quickstatements which can setup up a new Wikidata item.
Implementation of Scholia

Python (Python27 and Python35).

Uses the Flask, mostly selected because of its simplicity and nice tutorial.

JavaScript to query the Wikidata API (https://www.wikidata.org/w/api.php) for labels, inclusion of Wikipedia extract and setup of tables: jQuery, DataTables (elements have links to Scholia rather than Wikidata).

Test with tox, flake8 and py.test.
Contributing

For contribution, GitHub users can fork the repo at GitHub and make pull requests under GPL.

Particularly Egon Willighagen has made a number of pull requests around proteins, biological pathways and chemicals, see, e.g., citric acid where he has made SPARQL queries on qualifiers and references.

We got lots of issues.
The canonical version of Scholia runs from Toolforge: https://tools.wmflabs.org/

Toolforge (rebranded from Wikimedia Tool Labs), — the free cloud service provided by the Wikimedia Foundation.

Currently running gridengine. Considering Kubernetes to get better stability in response.
Wikidata-based \LaTeX\ \TeX{}\ generation

A rough-in-the-edges implementation in Scholia can generate \LaTeX\ \TeX{}\ \BIBTeX\ \.bib\ files from \.aux\ files

My \texttt{.tex}\ file:

\begin{verbatim}
\bibliographystyle{Nielsen2012Slides}
\bibliography{Nielsen2017ScholiaWikidataCon_slides}
\end{verbatim}

Commands:

\begin{verbatim}
latex Nielsen2017ScholiaWikidataCon_slides.tex
python -m scholia.tex write-bib-from-aux \\
   Nielsen2017ScholiaWikidataCon_slides.aux
bibtex Nielsen2017ScholiaWikidataCon_slides
latex Nielsen2017ScholiaWikidataCon_slides
latex Nielsen2017ScholiaWikidataCon_slides.tex
latex Nielsen2017ScholiaWikidataCon_slides.tex
\end{verbatim}
Scholia

Scholia issues :(

Wikidata far from complete. Particular non-Pubmed.

Citation data lacking, but some released with I4OC.

Some queries run into WDQS time-out.

Some queries generate too large results. This is a problem from graphs such as citation networks (we put in SPARQL “LIMIT”’s) and co-occurrence topics graphs (Egon Willighagen pointed to aluminum).
Scholia issues :) 

Wikidata act as a hub for different resources linking Google Scholar, Twitter, Scopus, VIAF, ResearchGate, ...

Good author disambiguation possible, — even for authors that do not have an account on the site.

Data description more detailed with many different properties: main theme, genre, multiple affiliation with time points, sex of author, license, sponsor, etc.

Linking to much more than science: Wikidata is becoming the “Internet duct tape that can solve anything” (light-hearted comment by Andrew Lih, somewhere on Facebook)
What’s next for Scholia and Wikicite?

Building scrapers. Could focus on non-PubMed and non-DOI publications.

Better integration between panels and aspects in Scholia (Javascript and D3 work)

“Editable Scholia”: Edit Wikidata items from Scholia. (Magnus Manske implements editing with his Listeria tool).

“Social Scholia”: User login, followers, followees, messages between users, messages when new relevant data appears in Wikidata.

Specialized aspects: Neuroinformatics, Bioinformatics, . . . ?

Feeds (Egon Willighagen is working on this issue)
Thanks