An overview of Scholia

Finn Årup Nielsen

DTU Compute
Technical University of Denmark

May 18, 2017
How do we show data from Wikidata?
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 Wikimmedia Tool Labs
Presenting Wikidata: SQID

Markus Krötzsch, Michael Günther et al. SQID, https://tools.wmflabs.org/sqid/

Wikidata class browser.

Displays typical properties

Runs from Wikimedia Tool Labs
How can we show scientific (bibliographic) data from Wikidata?
How can we show scientific (bibliographic) data from Wikidata?

For instance, a scholarly researcher profile, like we find in Google Scholar, ResearchGate, Scopus et al.
Scholia is a website with scholarly information extracted from Wikidata running from https://tools.wmflabs.org/scholia/ (Nielsen et al., 2017).

Almost entirely built by using Wikidata Query Service (WDQS), — the extended SPARQL endpoint available at https://query.wikidata.org/ maintained by the Wikimedia Foundation. Able to not only return tables with SPARQL results but also format the results with charts: maps, bar chart, graphs, etc.

Multiple “panels” on “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.
“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.

and some hidden aspects (work in progress)
Scholia: Author aspect publications per year

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

Number of publications per year.

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

Using default “BarChart” https://query.wikidata.org/#%23defaultView...
Scholia: Work aspect citation graph

Citation panel on work aspect for partial citation graph.

For A principal component analysis of 39 scientific impact measures.
Scholia: Work aspect citation graph

Citation panel on work aspect for partial citation graph.

For *A principal component analysis of 39 scientific impact measures.*

Actually a bit difficult to make good citation graphs.
Panel on publisher aspect with an overview of number of papers published and their citations across journals published by the publisher.

Here for BioMedCentral (which may be an imprint)
Incomplete statistics on page production per year for DTU Cognitive Systems.
Co-author-normalized citations per year for Technical University of Denmark.
Scholia: Organization aspect

Co-author graph for **DTU Cognitive Systems**.

Finn Årup Nielsen 15 May 18, 2017
Citation distribution

Citation distribution for PLOS ONE.
Citation distribution

Citation distribution for PLOS ONE. Here we would like a logarithm.
Citation distribution

Citation distribution for PLOS ONE, — with logarithms using WDQS’ interactive Graph Builder.
What questions from real life can Scholia answer?
Top 10 researchers with most Nature/Science articles on Unicph
Top 10 researchers with most Nature/Science articles on Unicph

Not (yet?) in Scholia, but WDQSable: http://tinyurl.com/kn3r4wz
Top 10 researchers with most Nature/Science articles on Unicph

Not (yet?) in Scholia, but WDQSable: http://tinyurl.com/kn3r4wz

<table>
<thead>
<tr>
<th>KU</th>
<th>Wikidata</th>
<th>Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>21</td>
<td>Eske Willerslev</td>
</tr>
<tr>
<td>83</td>
<td>18</td>
<td>Jun Wang</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>Ludovic Orlando</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>Søren Brunak</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>Niels Grarup</td>
</tr>
<tr>
<td>—</td>
<td>2</td>
<td>Eline D. Lorenzen</td>
</tr>
<tr>
<td>—</td>
<td>2</td>
<td>Thomas Werge</td>
</tr>
<tr>
<td>—</td>
<td>2</td>
<td>Albin Sandelin</td>
</tr>
<tr>
<td>—</td>
<td>2</td>
<td>Lars Juhl Jensen</td>
</tr>
<tr>
<td>—</td>
<td>2</td>
<td>Anders Krogh</td>
</tr>
</tbody>
</table>

Missing: Torben Hansen (27), Oluf Borbye Pedersen (24), Guojie Zhang (19), Rasmus Nielsen (16), Tom Gilbert (15)

Data is lacking due to the problem of resolving names like Wang, Zhang, Hansen, Pedersen, etc.
Give me an introductory paper

What is the best introductory/overview paper on word embeddings?
Give me an introductory paper

What is the best introductory/overview paper on word embeddings?

We are not there yet.
Give me an introductory paper

What is the best introductory/overview paper on word embeddings?

We are not there yet.

But we can get “Most cited works from works on the topic” from the topic aspect of word embedding pages.
**Give me an introductory paper**

What is the best introductory/overview paper on *word embeddings*?

We are not there yet.

But we can get “Most cited works from works on the topic” from the topic aspect of word embedding pages.

This gives: (Mikolov et al., 2013b; Mikolov et al., 2013a; Dhillon et al., 2012) in a table.

### Citations

**Most cited works from works on the topic**

<table>
<thead>
<tr>
<th>count</th>
<th>cited_work</th>
<th>cited_workLabel</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>wd:Q24731579</td>
<td>Distributed Representations of Words and Phrases and their Compositionality</td>
</tr>
<tr>
<td>3</td>
<td>wd:Q24699014</td>
<td>Efficient Estimation of Word Representations in Vector Space</td>
</tr>
<tr>
<td>1</td>
<td>wd:Q28646033</td>
<td>Two Step CCA: A new spectral method for estimating vector models of words</td>
</tr>
</tbody>
</table>
Scholia access statistics

Based on WMF toollabs' uwsgi.log log file with anonymized IP address.
Lookup ID on arXiv homepage, extract metadata and format it for Magnus Manske’s quick-statement web-service.
Scholia

Wikidata-based BIBTeX generation

A rough-in-the-edges implementation in Scholia can generate BIBTeX .bib files from .aux files

My .tex file:

\bibliographystyle{Nielsen2012Slides}
\bibliography{Nielsen2017Overview_slides}

Commands:

latex Nielsen2017Overview_slides.tex
python -m scholia.tex write-bib-from-aux Nielsen2017Overview_slides.aux
bibtex Nielsen2017Overview_slides
latex Nielsen2017Overview_slides.tex
latex Nielsen2017Overview_slides.tex
More command-line interfacing

```bash
> python -m scholia --help

Usage:
scholia arxiv-to-quickstatements [options] <arxiv>
scholia orcid-to-q <orcid>

Options:
- o --output=file   Output filename, default output to stdout

Examples:
$ python -m scholia orcid-to-q 0000-0001-6128-3356 Q20980928

References:
https://tools.wmflabs.org/wikidata-todo/quick_statements.php
```
Development

Developed from Github at https://github.com/fnielsen/scholia under GPL with work/input from Daniel Mietchen, Egon Willighagen, Jakob Voß, Magnus Manske, Andy Mabbett
Scholia :-( issues

Citation data in Wikidata far from complete meaning that Scholia’s representation may be quite biased. Scholia might disappoint researchers.

Paper affiliations are not made, thus scientometrics with precise affiliation resolving is not possible at the moment, and Scholia does not yet handle this issue well. Example: Dario Taraborelli’s paper assigned to UCL because of previous affiliation.

Query times: Large-scale analysis may be difficult with WDQS because of time-out. Perhaps Scholia should implement cache?
Scholia :) issues

An open alternative to commercial researcher profiler.

SPARQL with Blazegraphs graph queries on Wikidata quite powerfull.

Scholia exposes the possibilities with the different output formats in WDQS.

General idea: Other example “cvrminer” for (Danish) business data: https://tools.wmflabs.org/cvrminer/cvr/27761291
What’s next for Scholia?

Building scrapers. Initial work on community venues: JMLR, CEUR, . . .

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

Better search, better aspect switching, better . . .

“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, . . .?
Looking for the killer

What about uploading all of Danish research available at the Danish National Research Database?

What analysis can we (or Scholia) perform that Google Scholar, ResearchGate, Scopus, et al. cannot do?
Looking for the killer

What about uploading all of Danish research available at the Danish National Research Database?

What analysis can we (or Scholia) perform that Google Scholar, ResearchGate, Scopus, et al. cannot do? (note the gender panel in some of Scholia’s aspects)
Thanks
References


