Scholia: A Wikidata-based site for analytics and visualization of science

Finn Árup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

3 oktober 2018
Page production

Scientific article page production per year per author. The number of pages for a multiple-author paper is distributed among the authors. The statistics is only for papers where the "number of pages" property has been set.
Scholia is a webservice from https://tools.wmflabs.org/scholia/ and a Python package from https://github.com/fnielsen/scholia.

The webservice generates overview of science with Wikidata Query Service and is built with the Flask web framework, HTML, Bootstrap, Javascript and templated SPARQL.

For researcher profiles, scientometrics, bibliographic reference management, information discovery (find relevant papers, scientific meetings, researchers, funding opportunities, ...).
Where does the data come from?
“Wikidata: Verifiable, Linked Open Knowledge That Anyone Can edit” (Dario Taraborelli)

CC0-licensed data available on website, API, SPARQL endpoint or dump files.

Each page is an “item” with labels, aliases, properties and property values, as well as Wikipedia links.

Wikidata site UI mockup from 2012 for Berlin (Q64).
Wikidata Query Service (WDQS) is the SPARQL endpoint for the RDF-transformed data in Wikidata.

There is a “Query Helper” for non-programmatic formation of SPARQL queries, predefined prefixes, identifier lookup.

Several results output formats: table, bubble chart, line chart, graphs, etc.
WikiCite

“WikiCite: Building the sum of all human citations” (Dario Taraborelli)

Use Wikidata to hold metadata about works (scientific articles, book, etc.)

Properties: authors, publication date, where it is published, reviewed by, editor, main subject, language, retracted by, erratum, volume, issue number, page range, number of pages, type or genre (retraction notice, retracted paper), series, publisher, and a lot of identifiers: DOI, ACM, Semantic Scholar, PMCID, PMID, arXiv, etc.
WikiCite Statistics

<table>
<thead>
<tr>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6110672735</td>
<td>Total number of triples</td>
</tr>
<tr>
<td>121065663</td>
<td>Citations</td>
</tr>
<tr>
<td>77862349</td>
<td>Author name strings on items about works</td>
</tr>
<tr>
<td>17160242</td>
<td>Items with a PubMed ID</td>
</tr>
<tr>
<td>13835584</td>
<td>Items with a DOI</td>
</tr>
<tr>
<td>6889517</td>
<td>Items with a geolocation</td>
</tr>
<tr>
<td>4390875</td>
<td>Items with a PubMed Central ID</td>
</tr>
<tr>
<td>3516037</td>
<td>Links from items about works to items about their main subjects</td>
</tr>
<tr>
<td>2868187</td>
<td>Links from items about works to items about their authors</td>
</tr>
<tr>
<td>2519365</td>
<td>Items with a taxon name</td>
</tr>
<tr>
<td>186519</td>
<td>Items about authors with an ORCID profile that has public content</td>
</tr>
</tbody>
</table>

Wikidata statistics on WikiCite data. Currently presented on the main page of Scholia.

121 million citations.

17 million PubMed links.

14 million DOI links.

187 thousand ORCID links.
Jakob Voß’ WikiCite statistics

Jakob Voß’ Wikicite statistics that is update regularly.

http://wikicite.org/statistics.html

Number of publications and citations in Wikidata.

Note the staircase curve of the citations. My guess is that this shape is due to prolific James Hare using Europe PubMed Central initially and then switching to CrossRef for citations.
Scholia
Scholia shows Wiki-data data in *aspects*, author, work, organization (e.g., university, research group), venue (journal or conference), series, publisher, sponsor, location, event, award, topic, chemical, disease, etc.

For instance, the *Technical University of Denmark* may be viewed as a publisher, topic, organization, sponsor and location.
Author aspect: Co-author graph

The egocentric co-author graph in Scholia’s author aspect for the researcher Mikkel Wallentin, Aarhus University.

Colored according to gender.
Organization aspect: Citations

Co-author-normalized citations per year for Technical University of Denmark: Number of citations per year divided by number of co-authors on cited paper.
Work aspect: Retractions

Wikidata can specify retracted papers, retraction notices and their connection.

By combining citation and retraction information we can find papers citing another paper after it has been retracted.

Currently, Scholia visualizes such information in a timeline. Here Identification of Aurora-A as a direct target of E2F3 during G2/M cell cycle progression: “For example, silencing E2F3 prevented entry into G2/M in ovarian cancer cells [61].” (received April 2016, accepted August 2017)
Publisher aspects

Scatter plot of number of citations as a function of number of works published in journals published under the BioMed Central brand.

The top left one is Genome Biology, the lower right Critical Care.
Country aspect

Locations in Denmark that is the main subject of a work (Nielsen et al., 2018).

Example popup: Succession of phytoplankton in response to environmental factors in Lake Arresø, North Zealand, Denmark.

Similar maps can be created for narrative locations.
Research project aspect (Willighagen et al., 2018a).

If works are linked up to the project (by Wiki-data’s sponsored by property) we can make unusually statistics.

Here citations per million budget.

(The schema for projects and grants is not quite settled)
Use aspect

Bar chart for usage of SPM software (functional neuroimaging software) over time with different software versions indicated by color.

Uses the describes a project that uses property.

Such data is likely not available in directly machine readable format.
Comparison of multiple items

Multiple countries, e.g., some Southern and Eastern African countries or cheminformatics journals (here Willighagen’s citations to work ratio).
Scholia’s “subaspects”

Cocitation network for machine learning researchers in Denmark: /scholia/country/Q33/topic/Q2539.
Wikipedia researchers near Tübingen: Weight information in Wikidata by the geographical distance and topic of authored works (Nielsen et al., 2018).

/scholia/location/Q3806/-topic/Q52.

Nearby (in space and time) events also possible.

### Nearby researchers

<table>
<thead>
<tr>
<th>Score</th>
<th>Author</th>
<th>Example work</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.178268894199626</td>
<td>Ulrike Cress</td>
<td>A productive clash of perspectives? The interplay between articles' and authors' perspectives and their impact on Wikipedia edits in a controversial domain</td>
</tr>
<tr>
<td>9.818634462803981</td>
<td>Iassen Halatchliyski</td>
<td>A productive clash of perspectives? The interplay between articles' and authors' perspectives and their impact on Wikipedia edits in a controversial domain</td>
</tr>
<tr>
<td>1.604942154393766</td>
<td>Jason Weston</td>
<td>Reading Wikipedia to Answer Open-Domain Questions</td>
</tr>
<tr>
<td>0.16670001484301264</td>
<td>Denny Vrandečić</td>
<td>Revisiting reverts: accurate revert detection in Wikipedia</td>
</tr>
<tr>
<td>0.08335000742150632</td>
<td>Rudi Studer</td>
<td>Semantic Wikipedia</td>
</tr>
<tr>
<td>0.04167500371075316</td>
<td>Maria Koutraki</td>
<td>Wikipedia Infobox Type Prediction Using Embeddings</td>
</tr>
<tr>
<td>0.04167500371075316</td>
<td>Harald Sack</td>
<td>Wikipedia Infobox Type Prediction Using Embeddings</td>
</tr>
</tbody>
</table>
Finding related items
Related diseases with Wikidata Query Service

Count some form of co-occurences with a SPARQL query in the Wikidata Query service.

Scholia is doing this for diseases and proteins with tailor-made SPARQL. Here for the disease schizophrenia.

Shows genetically associated diseases via the P2293 (genetic association) property.
Finding related items based on word2vec-based knowledge graph embedding (Nielsen, 2017).

Here for a scientific article.

In this case, the similar articles found are (probably) mostly related to coauthorship relations.

But a newer embedding would probably be much affected by the citation relations between papers.
Related items by co-citations

<table>
<thead>
<tr>
<th>Count</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact</td>
</tr>
<tr>
<td>11</td>
<td>Twitter Predicts Citation Rates of Ecological Research.</td>
</tr>
<tr>
<td>10</td>
<td>How the scientific community reacts to newly submitted preprints: article downloads, Twitter mentions, and citations</td>
</tr>
<tr>
<td>9</td>
<td>Altmetrics: Value all research products</td>
</tr>
<tr>
<td>9</td>
<td>Characterizing social media metrics of scholarly papers: the effect of document properties and collaboration patterns</td>
</tr>
<tr>
<td>8</td>
<td>Tweeting birds: online mentions predict future citations in ornithology.</td>
</tr>
<tr>
<td>8</td>
<td>I Like, I Cite? Do Facebook Likes Predict the Impact of Scientific Work?</td>
</tr>
<tr>
<td>7</td>
<td>The differential impact of scientific quality, bibliometric factors, and social media activity on the influence of systematic reviews and meta-analyses about psoriasis.</td>
</tr>
<tr>
<td>7</td>
<td>A systematic identification and analysis of scientists on Twitter.</td>
</tr>
<tr>
<td>6</td>
<td>Social media release increases dissemination of original articles in the clinical pain sciences</td>
</tr>
</tbody>
</table>

Example with *Do alt-metrics work? Twitter and ten other social web services.*

Counts citations back and forth, one step and two step with the SPARQL fragment:

```
wd:Q21133507 (^wdt:P2860 | wdt:P2860) /
(^wdt:P2860 | wdt:P2860)?work .
```
How do we get data into Wikidata?
### Wikidata input

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>main subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Trapping the Tiger: Efficacy of the Novel BG-Sentinel 2 With Several Attractants and Carbon Dioxide for Collecting Aedes albopictus (Diptera: Culicidae) in Southern France Q22330695</td>
<td>Asian tiger mosquito</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Culicidae</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chikungunya Virus</td>
</tr>
<tr>
<td>2</td>
<td>New vascular plant records for the Canadian Arctic Archipelago Q22583137</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Demography of some non-native isopods (Crustacea, Isopoda, Oniscidea) in a Mid-Atlantic forest, USA Q22675943</td>
<td>demographics</td>
</tr>
<tr>
<td>4</td>
<td>An Asiatic Chironomid in Brazil: morphology, DNA barcode and bionomics Q22675958</td>
<td>Brazil</td>
</tr>
<tr>
<td>5</td>
<td>Occurrence of Diopatra marocensis (Annelida, Onuphidae) in the eastern Mediterranean Q22680870</td>
<td></td>
</tr>
</tbody>
</table>


Magnus Manske’s tools: Source-MD including its ORCIDator and resolver, Quickstatements, TABernacle (left screenshot). Relatively quick for each researcher if ORCID profile has DOI publications.

Other approaches: Fatameh, programmatic upload, e.g., with WikidataIntegrator.

Scholia has arXiv scraping.
Scientometrics limitations

PubMed bias: A large portion of the documents comes from PubMed.

DOI bias: Documents with DOIs are easier to setup than documents without.

I4OC bias: The citations we have (and that we are going to get) are primarily from open citation databases (CrossRef), i.e., citations from organizations such as IEEE and Elsevier are underrepresented.

Authors are not equally represented. One problem: Some author names are hard to resolve, e.g., Chinese and Korean names, cf. (Ioannidis et al., 2018).

Scholia bias: Chemoinformatics, Zika virus, etc.
Scholia usage statistics

Monthly pageview for Scholia has increased and has been over 300'000.

The latest increase is likely due to inclusion of link to Scholia from Wikimedia Commons templates. Whether page view coming this way are bots or users are not known.
Scholia/Wikidata promotions

How do we spread the word of Scholia and Wikidata?

Here Egon Willighagen uses the hashtag #icanhazwikidata to encourage researchers to tweet their ORCID iD so that we can “orcidator” their publication into Wikidata.

Deep links from Wikipedia and Wikimedia Commons to Scholia profiles, e.g., on Uta Frith.
Development

Development takes place on GitHub under GPL at https://github.com/-fniesen/scholia/.

Three developers: Egon Willighagen (almost all chemoinformatics aspects, biological pathways, etc., see also (Willighagen et al., 2018b)) and Daniel Mietchen.

Provided a Python development environment, you can download and run Scholia on your own computer.
Conclusion

Wikidata and its Wikidata Query Service yield an open corpus of metadata queryable in complex ways.

Scholia aggregates Wikidata data and present the data in an interactive environment.

Data in Wikidata is limited and there is biased coverage.

Wikidata input is somewhat cumbersome. We rely heavily on Magnus Manskes bespoke tools.

Ontology still not clear, e.g., preprints, postprints

WikiCite part of Wikidata continues to grow.
References


Copyright and license

Wikidata logo by Arun Ganesh (Planemad). It is a trademark of the Wikimedia Foundation.

WikiData UI mockup by Denny Vrandecic, CC0.

Jakob Voß’ statistics plot is by himself with an unknown license.

Screenshot from Magnus Manske webservice.

Map is CC BY-SA by OpenStreetMap contributors.

WikiCite logo by Dario Taraborelli, CC0.

Photo of Dario Taraborelli by Pax Ahimsa Gethen, CC BY-SA 4.0.