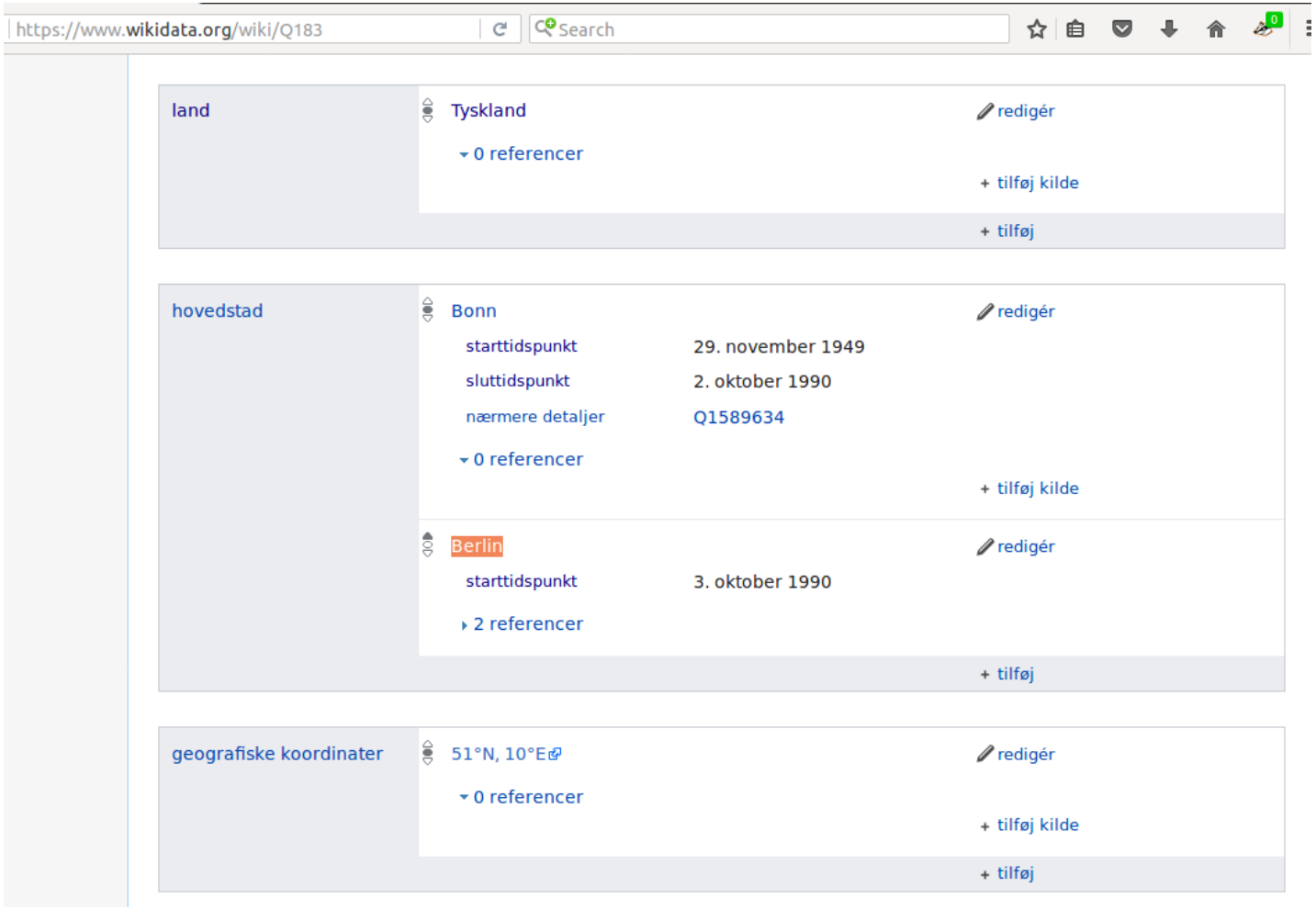
Scholia and scientometrics with Wikidata

Finn Årup Nielsen,[1] Daniel Mietchen[2] and Egon Willighagen[3]

[1]DTU Compute, Technical University of Denmark;
[2]EvoMRI Communications;
[3]Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University

May 28, 2017

Wikidata



Wikipedia sister project for structured data: http://www.wikidata.org

Wikidata

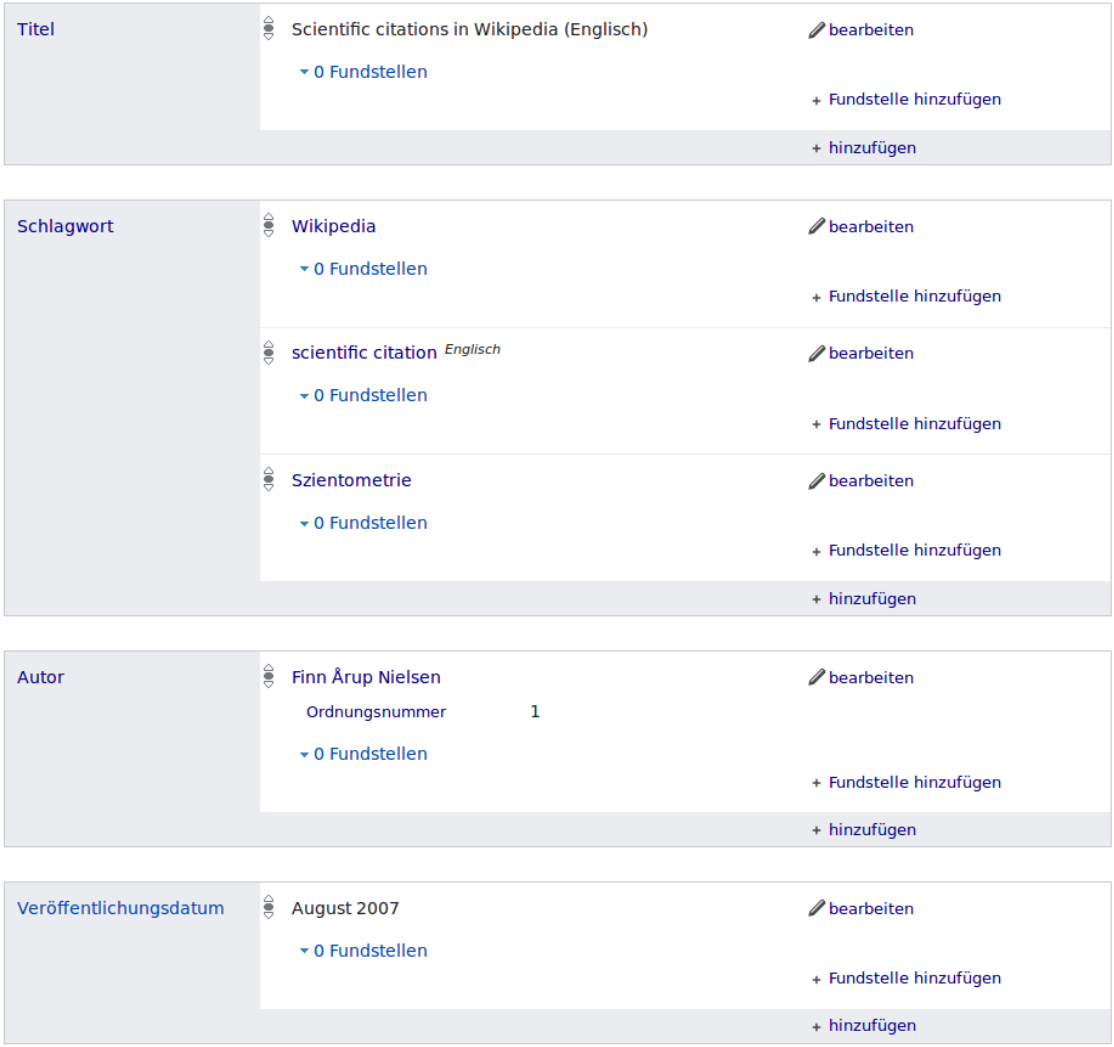
Wikidata = triples + qualifiers + references

Triples is the Semantic Web concept (Ressource Description Framework), e.g., (Germany, has_capital, Berlin)

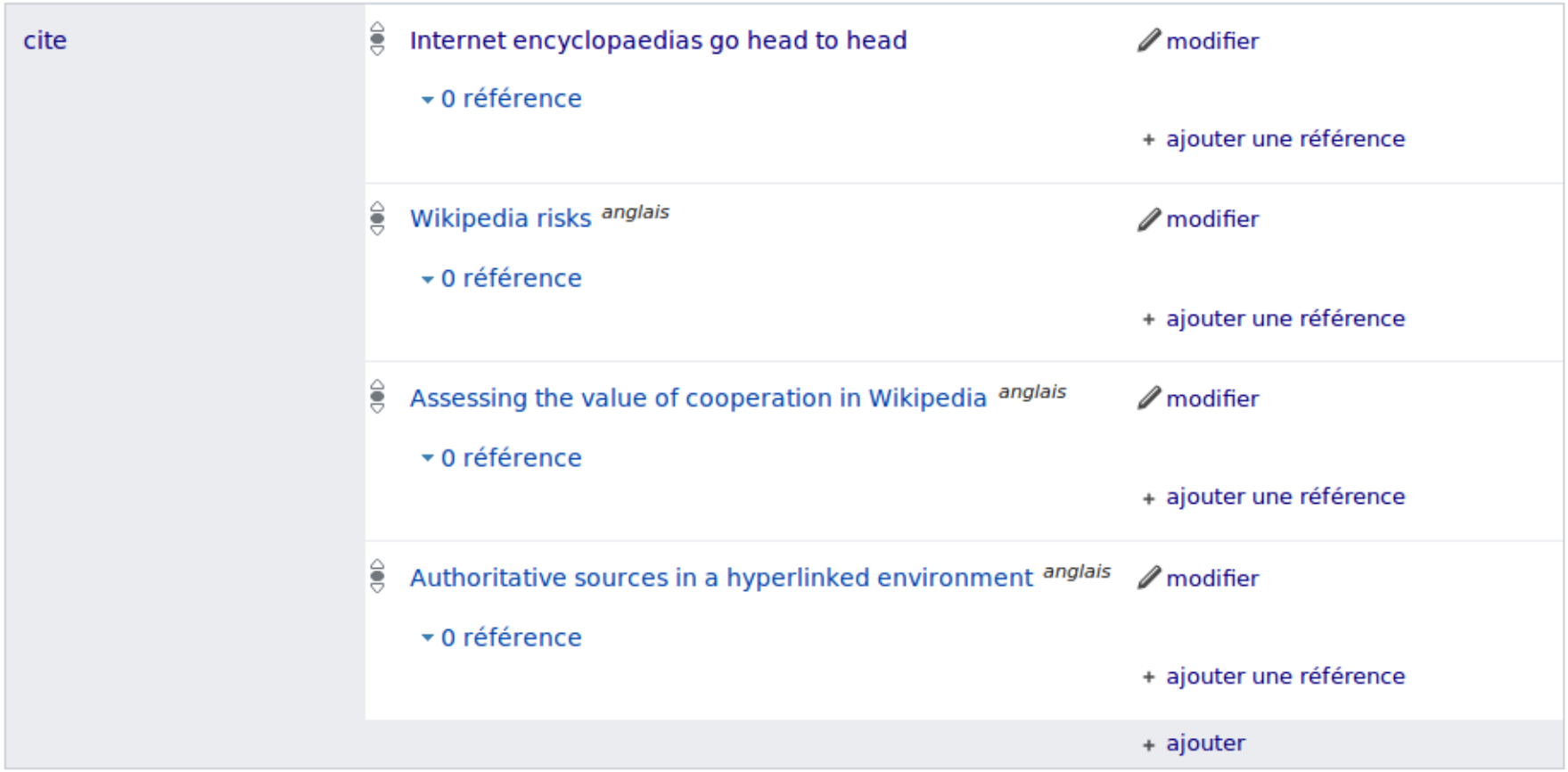
With qualifiers, e.g., (Germany, has_capital, Berlin, start_time, 1990-10-03)

With references, e.g., (Germany, has_capital, Berlin, start_time, 1990-10-03, url, http://www.bundestag.de/bundestag/aufgaben/rechtsgrundlagen-/grundgesetz/gg_02.html)

Bibliographic data in Wikidata



and citation information



Here Wikidata describes that (Nielsen, 2007) cites (Giles, 2005; Denning et al., 2005; Wilkinson and Huberman, 2007; Kleinberg, 1999).

Data entry

Wikidata's bibliographic information (Wikicite) data relies heavily on individuals and a bioinformatics group:

Magnus Manske: Tools, such as quickstatement and resolver

James Hare: Upload of scientific bibliographic data

Daniel Mietchen: Upload of scientific bibliographic data

San Diego et al. bioinformatics group: Genes, proteins, drugs, diseases, etc. (Mitraka et al., 2015; Burgstaller-Muehlbacher et al., 2016; Putman et al., 2017)

But so far we got

671'892 scientific articles according to WDQS as of 8 May 2017.

9633 scientific authors as Wikidata items according to WDQS.

1'791'391 unique scientific author strings according to WDQS.

And the number of citations:

> "The @Wikidata Citation Graph hit 3 million connections earlier this morning. @Wikicite"
>
> — James Hare announcing on Twitter 30 April 2017

Wikidata

Wikidata was first used to capture the language links between Wikipedias.

Now it is being used to fill Wikipedia infoboxes.

Some Wikipedias are using the Wikidata bibliographic items.

Wikidata

But Wikidata has the potential to do more than that.

Wikidata

But Wikidata has the potential to do more than that.

Scientometrics?

Wikidata

But Wikidata has the potential to do more than that.

Scientometrics?

Bibliography reference management?

Wikidata

But Wikidata has the potential to do more than that.

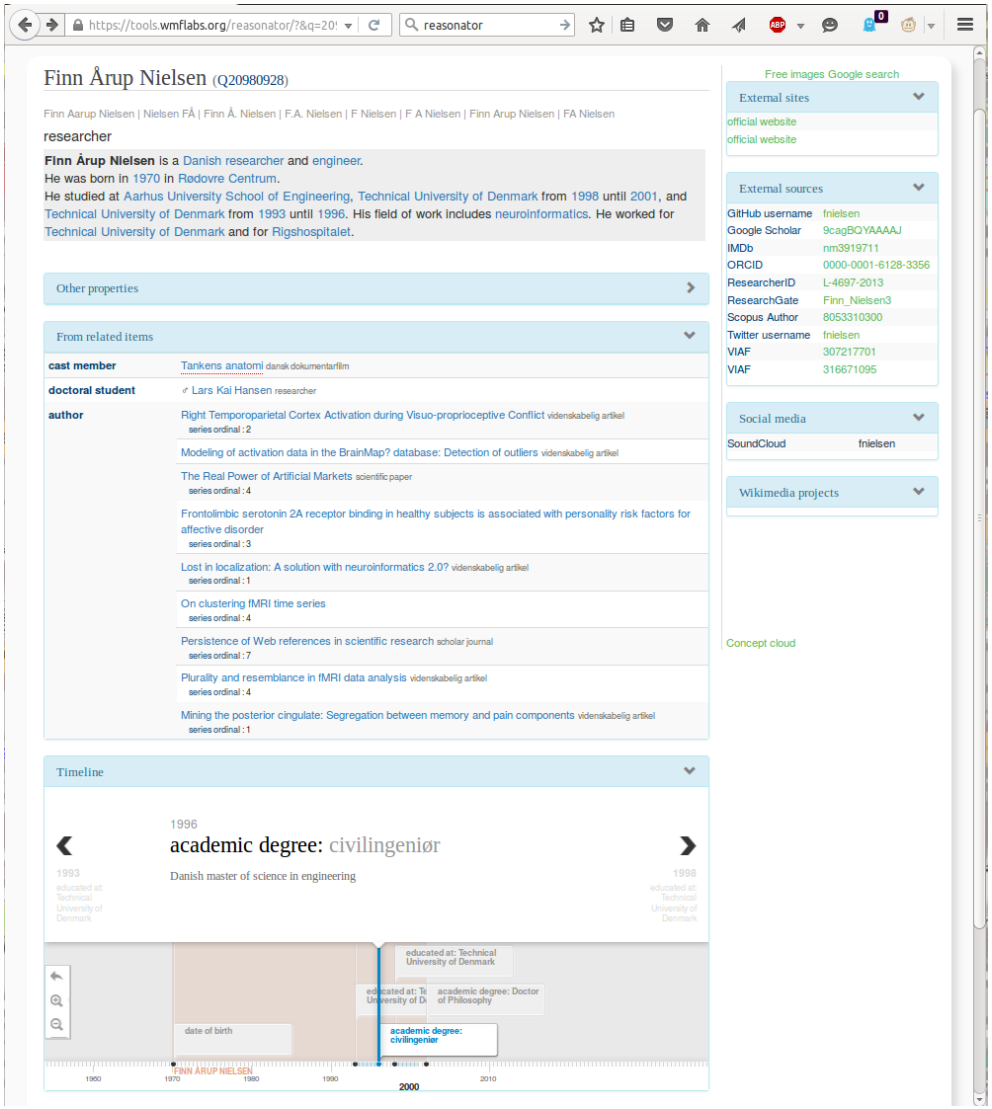
Scientometrics?

Bibliography reference management?

. . .

How can we present data from Wikidata?

Presenting Wikidata: Reasonator

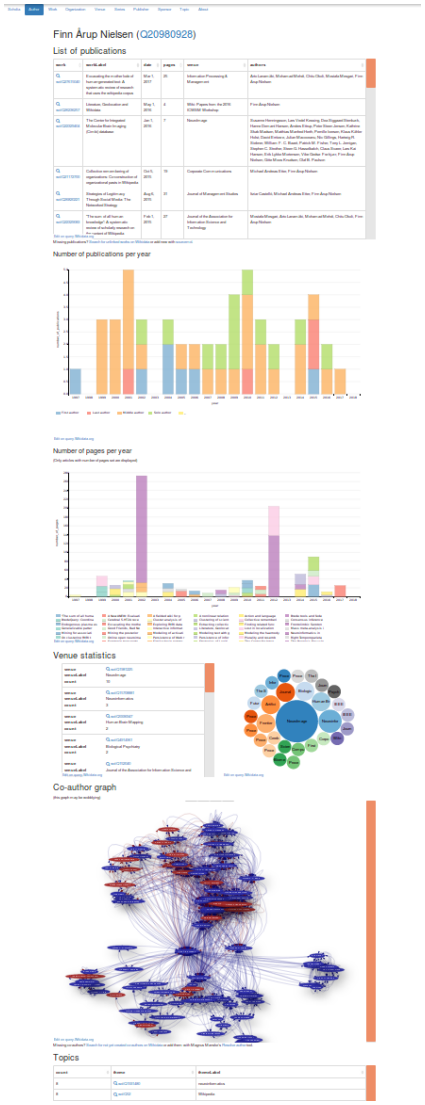


Magnus Manske's Reasonator, https://tools.wmflabs.org/reasonator/

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 Tool Labs*

Scholia

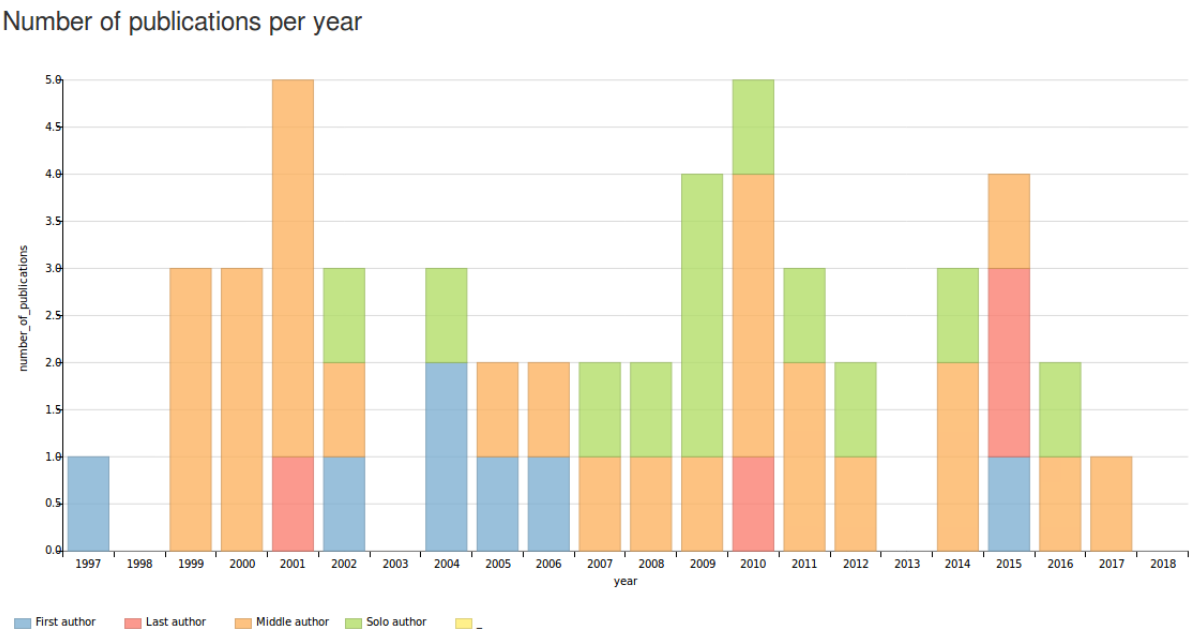


Web site with scholarly information extracted from Wikidata running from https://tools.wmflabs.org/scholia/.

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

Almost entirely built by using Wikidata Query Service, — an 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.

Scholia: Author aspect publications per year



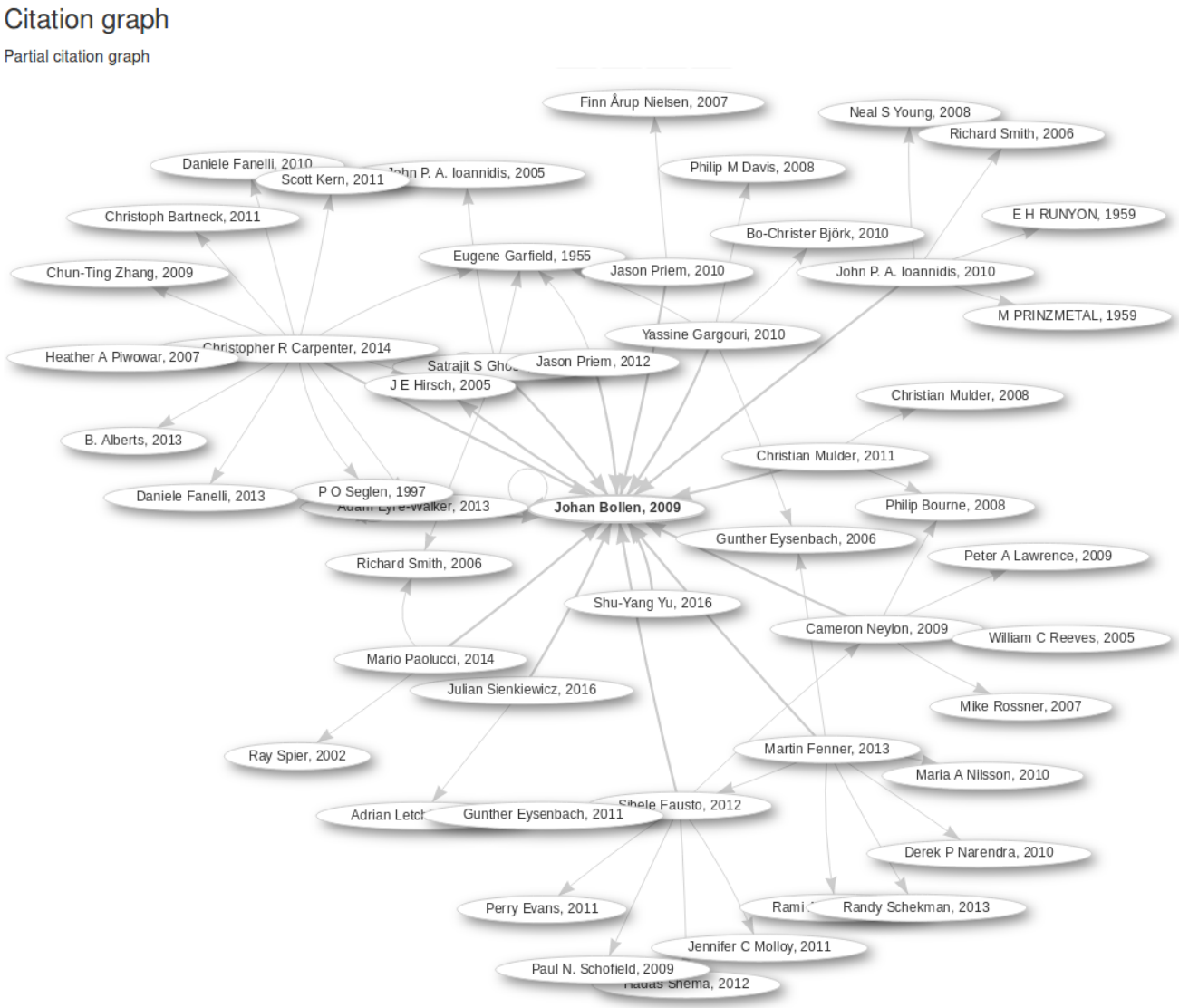
Inspired by Shubhanshu Mishra's and Vetle I. Torvik's LEGOLAS visualization.

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...

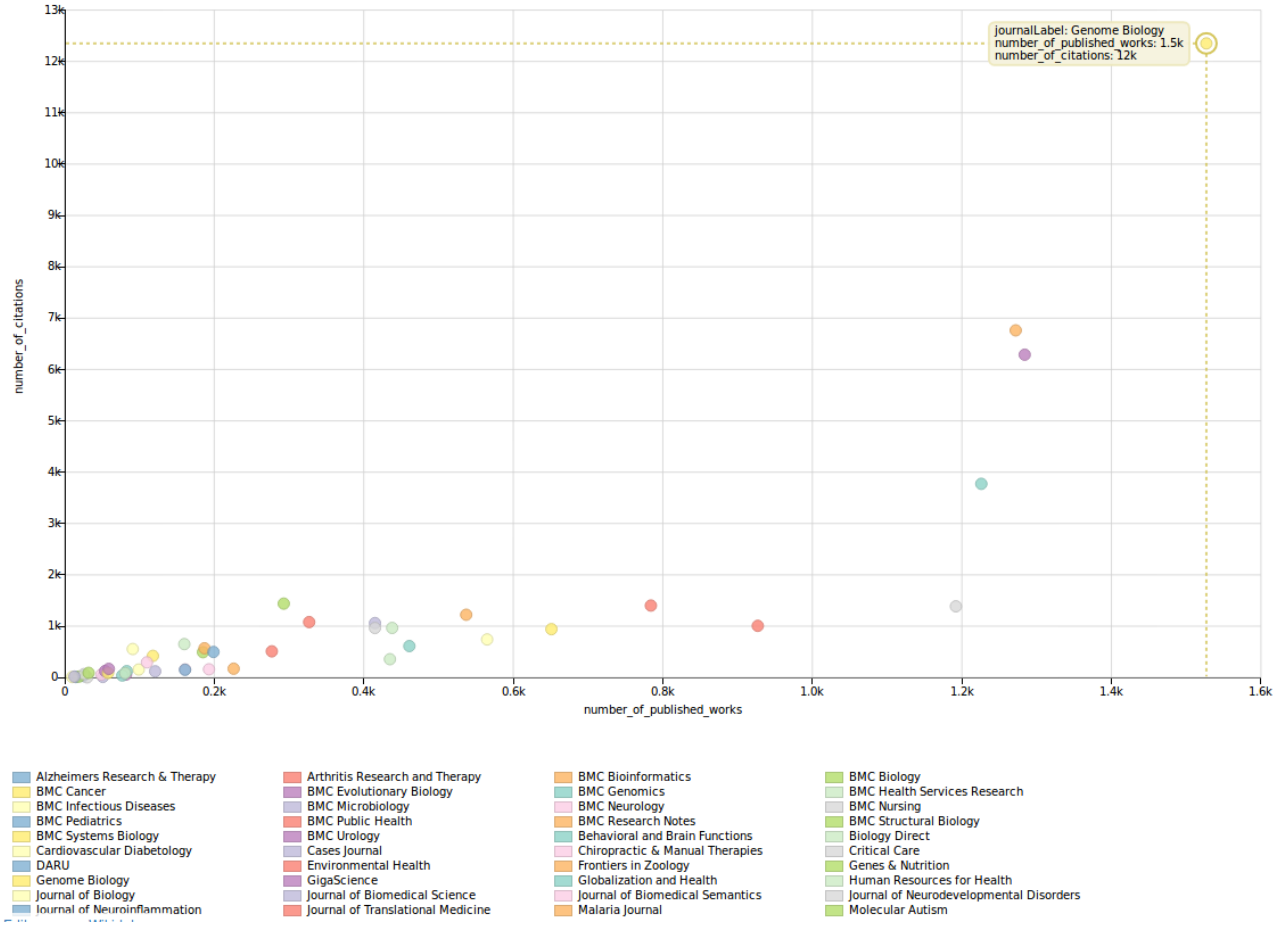
Scholia: Work aspect citation graph



Citation panel on *work* aspect for partial citation graph.

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

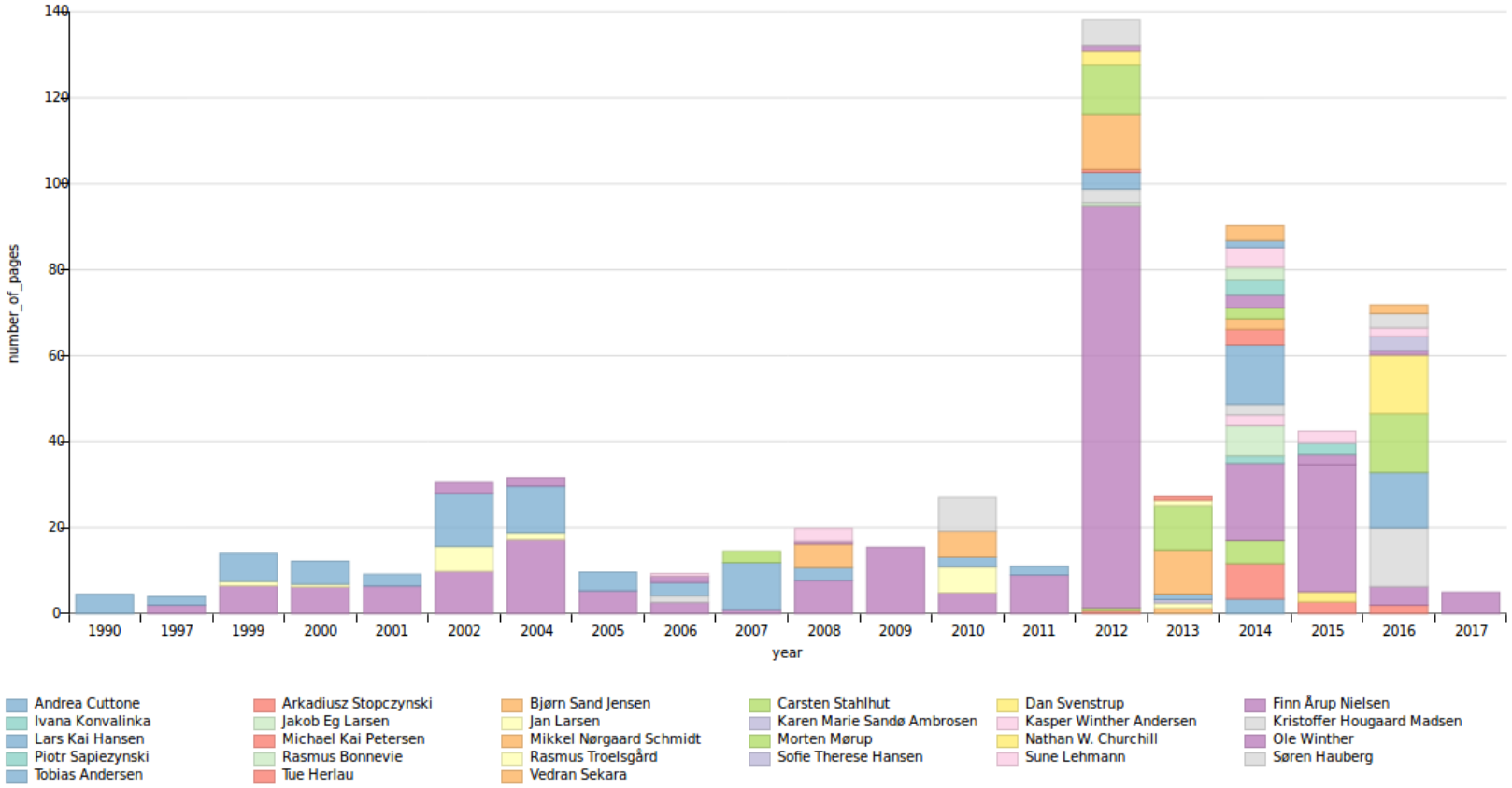
Scholia: Publisher aspect



Overview of number of papers published and their citations across journals published by the publisher.

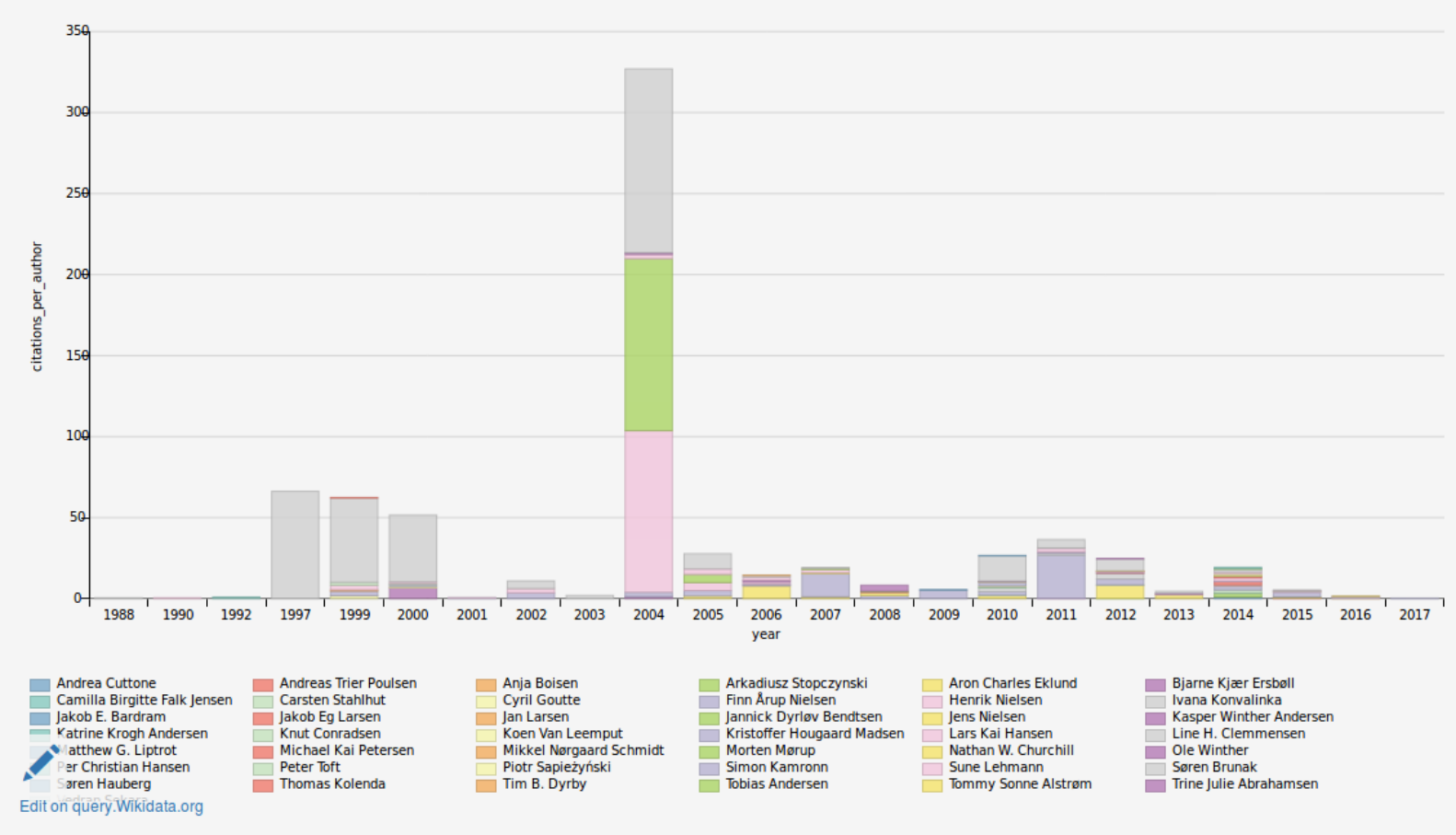
Here for BioMedCentral (which may be an imprint)

Scholia: Organization aspect



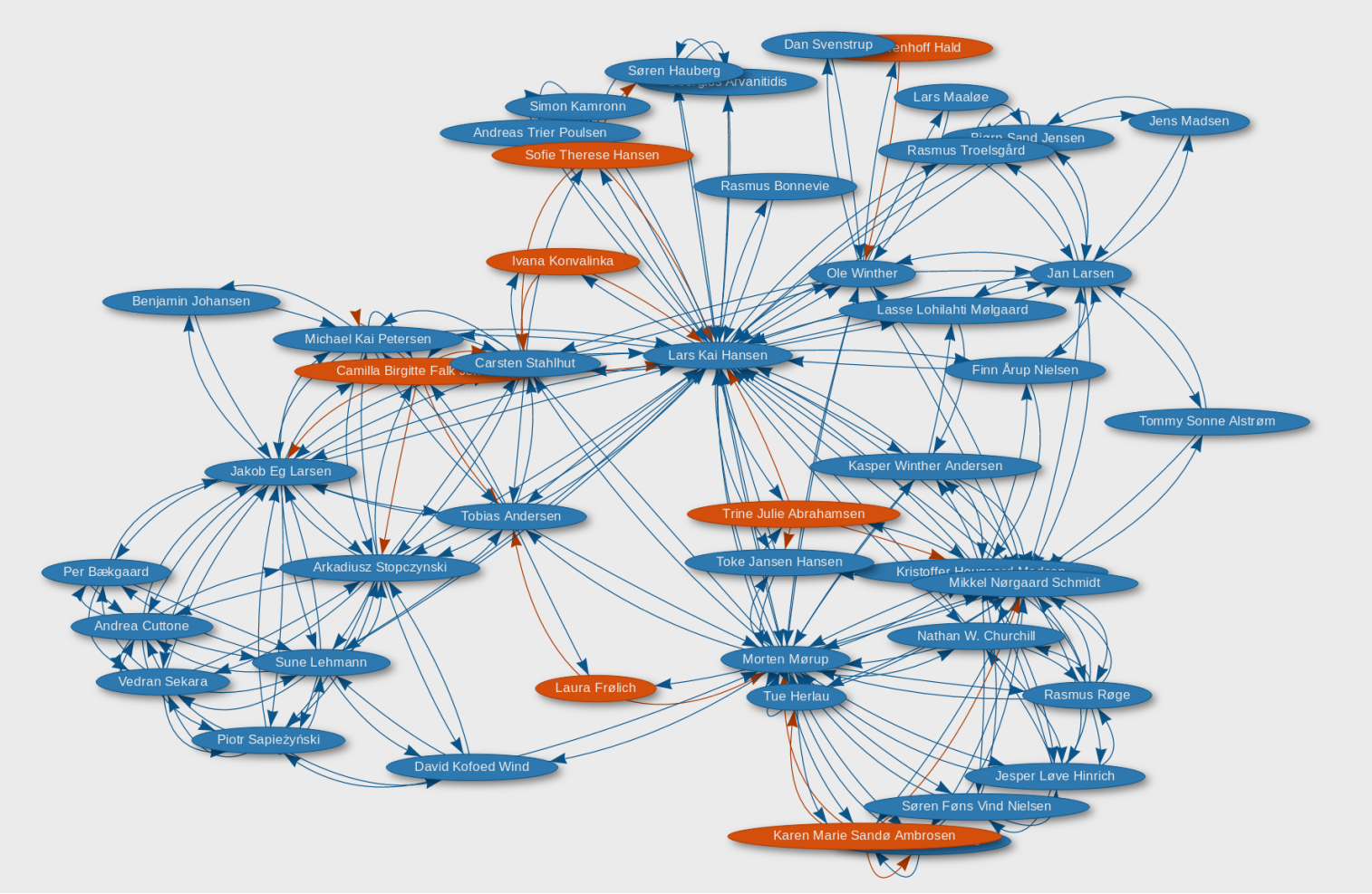
Incomplete statistics on page production per year for DTU Cognitive Systems.

Scholia: Organization aspect



Co-author-normalized citations per year for Technical University of Denmark.

Scholia: Organization aspect



Co-author graph for DTU Cognitive Systems.

What questions from real life can Scholia answer?

Top 10 researchers with most Nature/Science articles on University of Copenhagen

Top 10 researchers with most Nature/Science articles on University of Copenhagen

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

Top 10 researchers with most Nature/Science articles on University of Copenhagen

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

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

Missing: Torben Hansen (27), Oluf Borbye Pedersen (24), Guo-jie 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

count	cited_work	cited_workLabel
3	wd:Q24731579	Distributed Representations of Words and Phrases and their Compositionality
3	wd:Q24699014	Efficient Estimation of Word Representations in Vector Space
1	wd:Q28646033	Two Step CCA: A new spectral method for estimating vector models of words

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{Nielsen2017Scholia_slides}
```

Commands:

```
latex Nielsen2017Scholia_slides.tex
python -m scholia.tex write-bib-from-aux Nielsen2017Scholia_slides.aux
bibtex Nielsen2017Scholia_slides
latex Nielsen2017Scholia_slides.tex
latex Nielsen2017Scholia_slides.tex
```

Wikicite issues :(

Wikidata far from complete.

Citation data lacking, but some released with I4OC.

Paper affiliations are not made, thus scientometrics with precise affiliation resolving is not possible at the moment.

Large-scale analysis is difficult with WDQS because of time-out.

Wikicite 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?

Continued upload of data available from APIs to Wikidata.

Building scrapers, e.g., in Scholia.

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, . . . ?

Thanks

References

Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T. E., Leong, J., Naik, C., Pavlidis, P., Schriml, L., Good, B. M., and Su, A. I. (2016). Wikidata as a semantic framework for the Gene Wiki initiative. *Database*, 2016:baw015. DOI: 10.1093/DATABASE/BAW015.

Denning, P., Horning, J., Parnas, D., and Weinstein, L. (2005). Wikipedia risks. 48:152. DOI: 10.1145/1101779.1101804.

Dhillon, P. S., Rodu, J., Foster, D. P., and Ungar, L. H. (2012). Two Step CCA: A new spectral method for estimating vector models of words.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438:900–901. DOI: 10.1038/438900A.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632. DOI: 10.1145/324133.324140.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space.

Mikolov, T., Dean, J., and Corrado, G. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119.

Mitraka, E., Waagmeester, A., Burgstaller-Muehlbacher, S., Schriml, L. M., Su, A. I., and Good, B. M. (2015). Wikidata: A platform for data integration and dissemination for the life sciences and beyond. DOI: 10.1101/031971.

Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *First Monday*, 12. DOI: 10.5210/FM.V12I8.1997.

Putman, T. E., Lelong, S., Burgstaller-Muehlbacher, S., Burgstaller-Muelhbacher, S., Waagmeester, A., Diesh, C., Dunn, N., Munoz-Torres, M., Stupp, G., Su, A. I., Wu, C., Su, A. I., Good, B. M., and Good, B. M. (2017). WikiGenomes: an open Web application for community consumption and curation of gene annotation data in Wikidata. *Database*, 2017. DOI: 10.1101/102046.

Wilkinson, D. M. and Huberman, B. A. (2007). Assessing the value of cooperation in Wikipedia. *First Monday*, 12. DOI: 10.5210/FM.V12I4.1763.