

Scholia and scientometrics with Wikidata

Finn Årup Nielsen,¹ Daniel Mietchen² and Egon Willighagen³

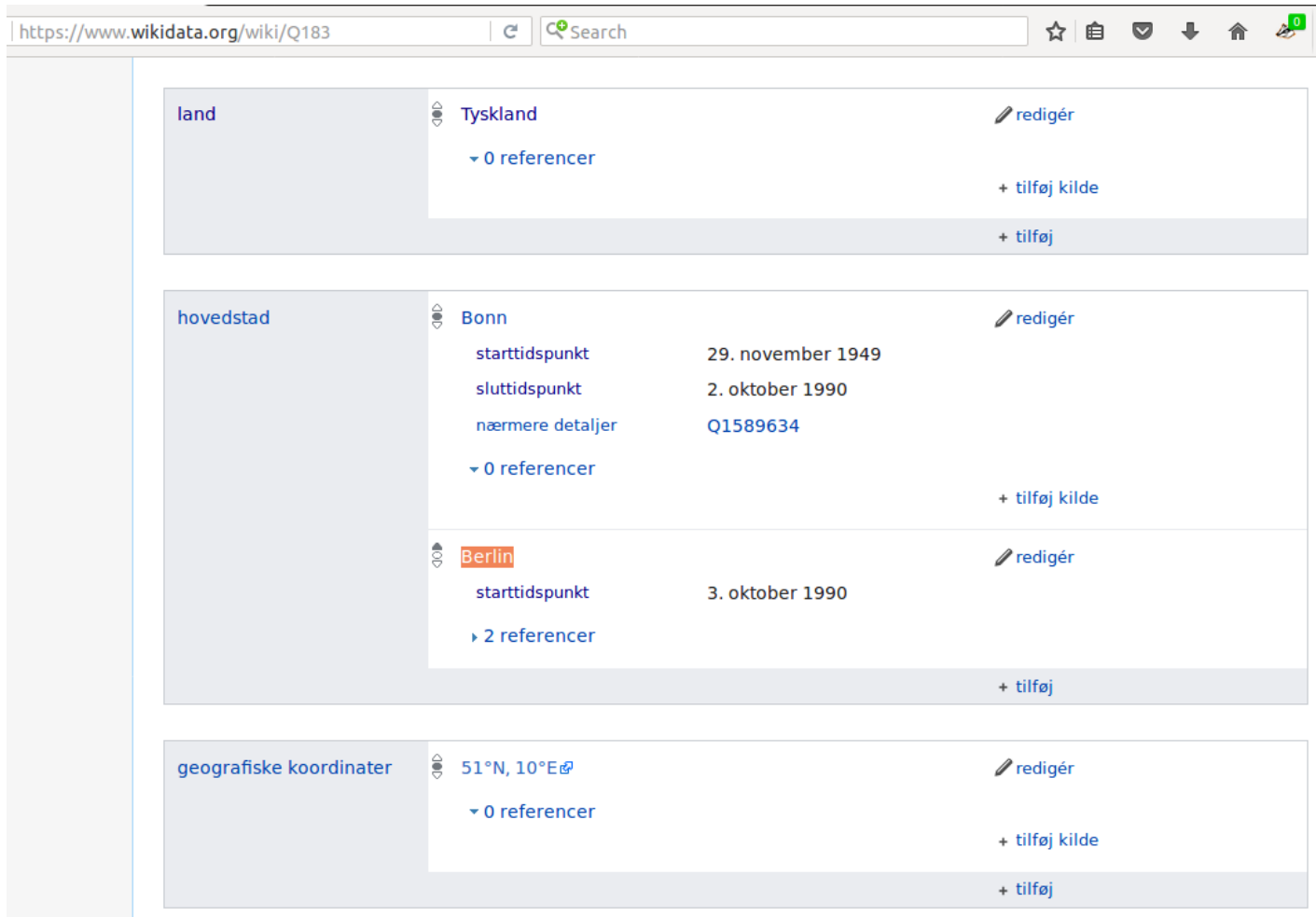
¹DTU Compute, Technical University of Denmark;

²EvoMRI Communications;

³Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University

May 28, 2017

Wikidata



The screenshot shows the Wikidata page for Q183 (Berlin). The browser address bar displays <https://www.wikidata.org/wiki/Q183>. The page content is organized into several sections, each with a label on the left and data on the right. Each section includes a 'redigér' (edit) button and options to '+ tilføj kilde' (add source) and '+ tilføj' (add).

Property	Value	References
land	Tyskland	0 referencer
hovedstad	Bonn	0 referencer
	starttidspunkt: 29. november 1949	
	sluttidspunkt: 2. oktober 1990	
	nærmere detaljer: Q1589634	
	Berlin	2 referencer
	starttidspunkt: 3. oktober 1990	
geografiske koordinater	51°N, 10°E	0 referencer

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

Titel	Scientific citations in Wikipedia (English) ▾ 0 Fundstellen	✎ bearbeiten + Fundstelle hinzufügen + hinzufügen
Schlagwort	Wikipedia ▾ 0 Fundstellen scientific citation <i>Englisch</i> ▾ 0 Fundstellen Szientometrie ▾ 0 Fundstellen	✎ bearbeiten + Fundstelle hinzufügen ✎ bearbeiten + Fundstelle hinzufügen ✎ bearbeiten + Fundstelle hinzufügen + hinzufügen
Autor	Finn Årup Nielsen Ordnungsnummer 1 ▾ 0 Fundstellen	✎ bearbeiten + Fundstelle hinzufügen + hinzufügen
Veröffentlichungsdatum	August 2007 ▾ 0 Fundstellen	✎ bearbeiten + Fundstelle hinzufügen + hinzufügen

and citation information

cite	  Internet encyclopaedias go head to head  modifier	
	▼ 0 référence	+ ajouter une référence
	  Wikipedia risks <i>anglais</i>  modifier	
	▼ 0 référence	+ ajouter une référence
  Assessing the value of cooperation in Wikipedia <i>anglais</i>  modifier		
▼ 0 référence	+ ajouter une référence	
  Authoritative sources in a hyperlinked environment <i>anglais</i>  modifier		
▼ 0 référence	+ ajouter une référence	
	+ ajouter	

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 Mitchen: 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

The screenshot shows the Reasonator tool interface for Finn Årup Nielsen (Q20980928). The page is divided into several sections:

- Header:** Displays the name "Finn Årup Nielsen (Q20980928)" and the role "researcher".
- Description:** A paragraph stating: "Finn Årup Nielsen is a Danish researcher and engineer. He was born in 1970 in Rødovre Centrum. He studied at Aarhus University School of Engineering, Technical University of Denmark from 1998 until 2001, and Technical University of Denmark from 1993 until 1996. His field of work includes neuroinformatics. He worked for Technical University of Denmark and for Rigshospitalet."
- Other properties:** A section for additional information.
- From related items:** A list of related items including:
 - cast member: Tankens anatomi dansk dokumentarfilm
 - doctoral student: Lars Kai Hansen researcher
 - author: Multiple scientific articles such as "Right Temporoparietal Cortex Activation during Visuo-proprioceptive Conflict", "Modeling of activation data in the BrainMap? database: Detection of outliers", "The Real Power of Artificial Markets", "Frontolimbic serotonin 2A receptor binding in healthy subjects is associated with personality risk factors for affective disorder", "Lost in localization: A solution with neuroinformatics 2.0?", "On clustering fMRI time series", "Persistence of Web references in scientific research", "Plurality and resemblance in fMRI data analysis", and "Mining the posterior cingulate: Segregation between memory and pain components".
- Timeline:** A horizontal timeline showing key events:
 - 1993: educated at: Technical University of Denmark
 - 1996: academic degree: civilingeniør
 - 1998: Danish master of science in engineering
 - 1998: educated at: Technical University of Denmark
 - 2001: educated at: Technical University of Denmark
 - 2001: academic degree: Doctor of Philosophy
- External sites:** Links for "official website" and "official website".
- External sources:** A table of identifiers:

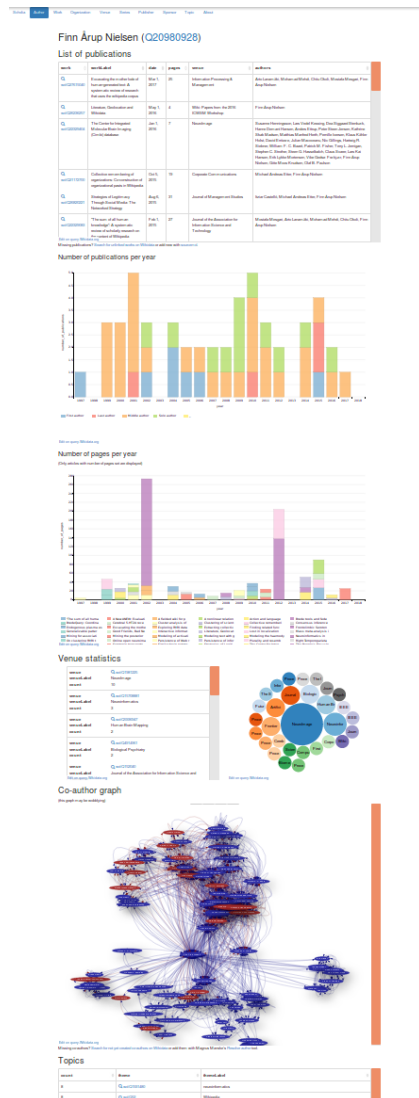
GitHub username	Inielson
Google Scholar	9cagBQYAAAAJ
IMDb	nm3919711
ORCID	0000-0001-6128-3356
ResearcherID	L-4697-2013
ResearchGate	Finn_Nielsen3
Scopus Author	8053310300
Twitter username	Inielson
VIAF	307217701
VIAF	316671095
- Social media:** A link for "SoundCloud: Inielson".
- Wikimedia projects:** A section for related Wikimedia projects.
- Concept cloud:** A section for a concept cloud.

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

Number of publications per year



Inspired by [Shubhanshu Mishra's](#) and [Vetle I. Torvik's](#) LEGOLAS visualization.

Number of publications per year.

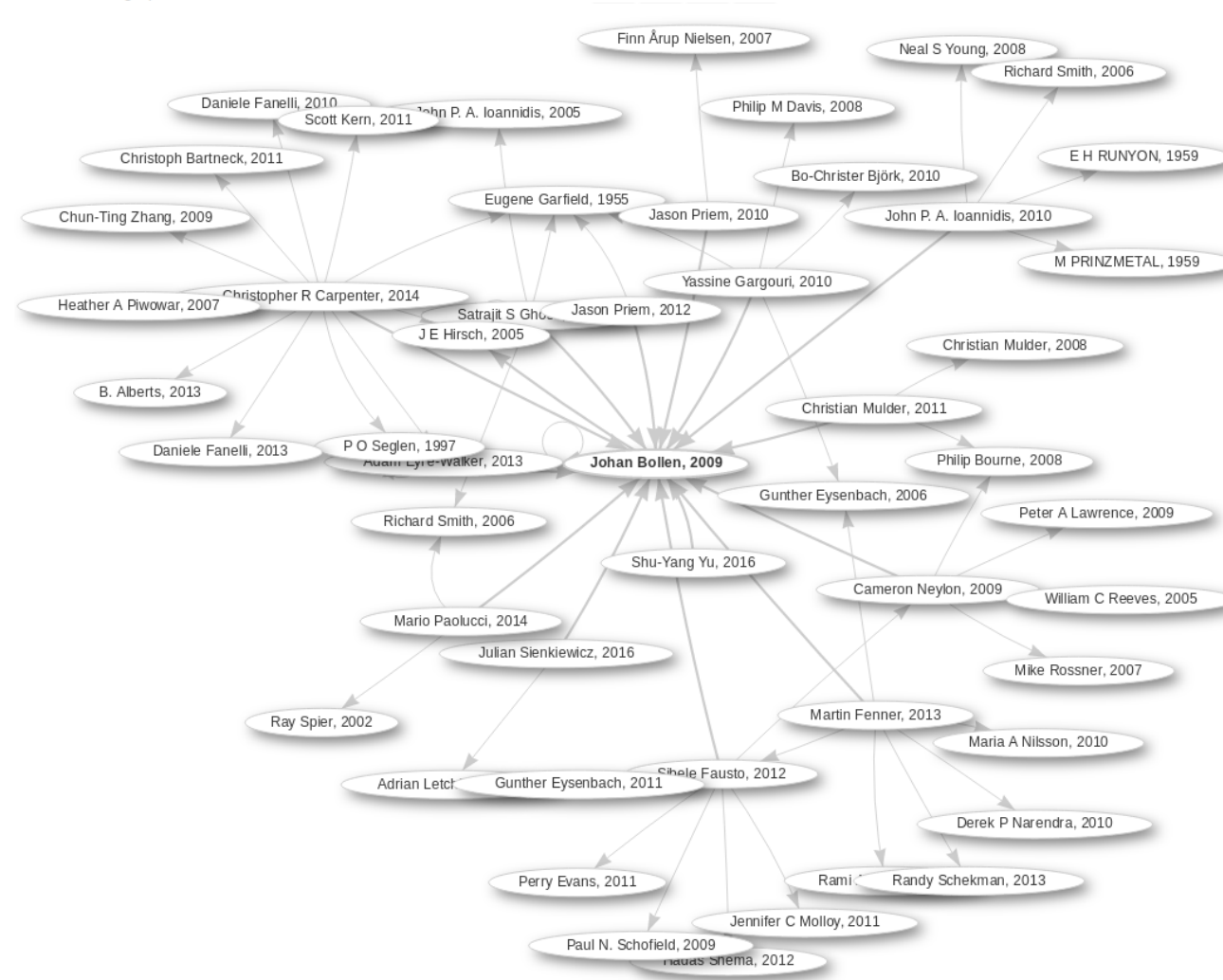
Color-coding based on author-role (first author, last author, middle author, solo author)

Using default “BarChart” <https://query.wikidata.org/#%23defaultView...>

Scholia: Work aspect citation graph

Citation graph

Partial citation graph



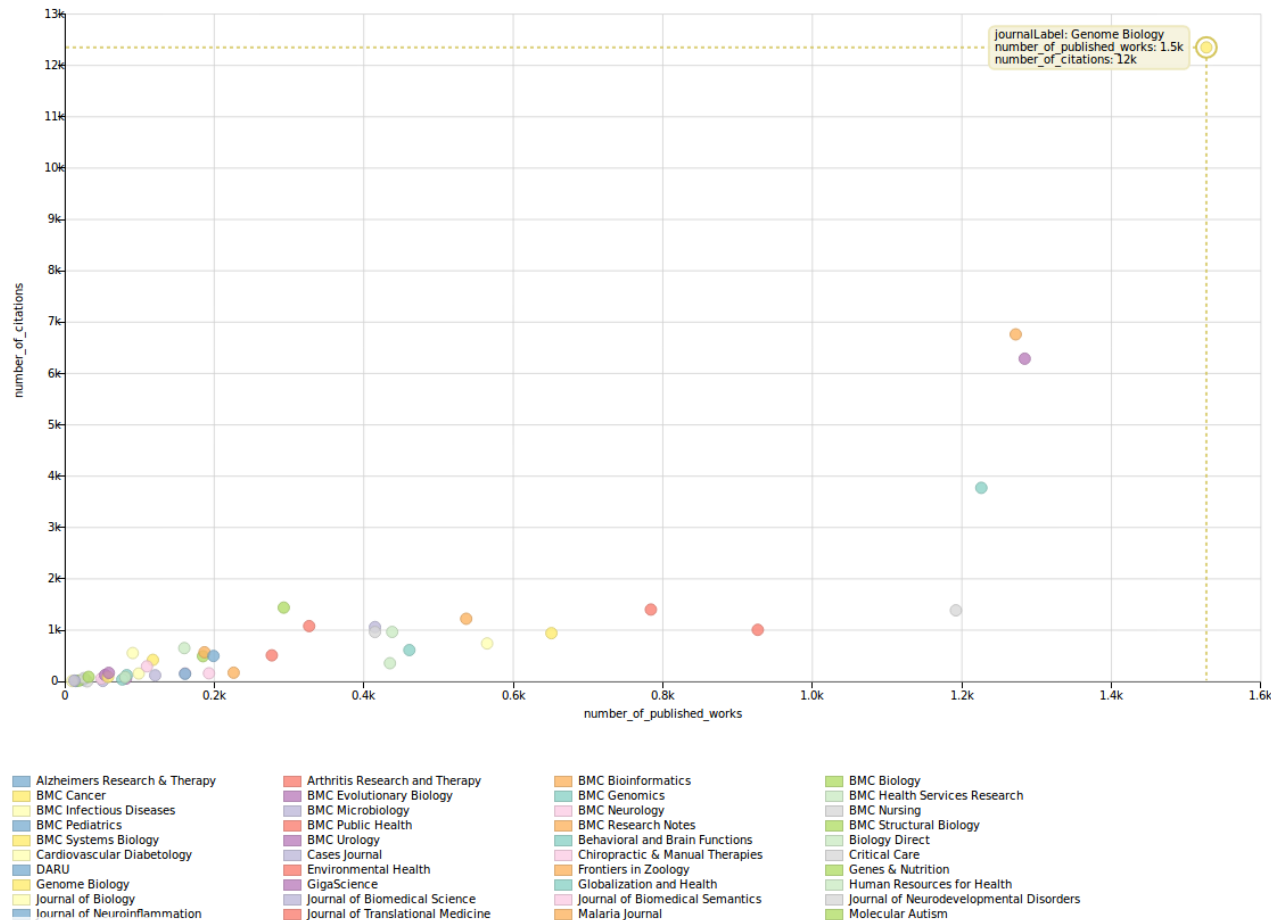
Citation panel on *work* aspect for partial citation graph.

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

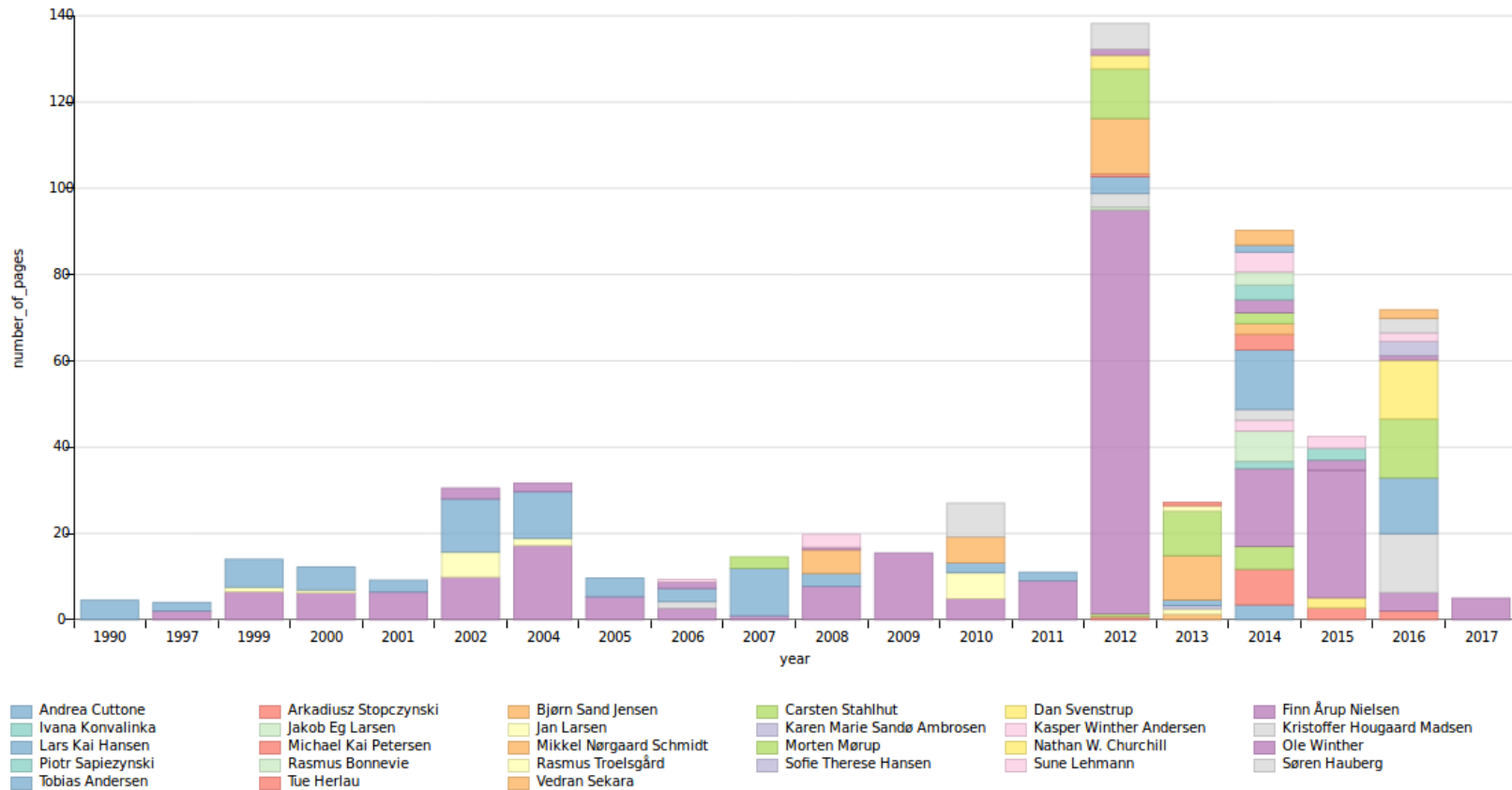
Scholia: Publisher aspect

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

Here for BioMedCentral (which may be an imprint)

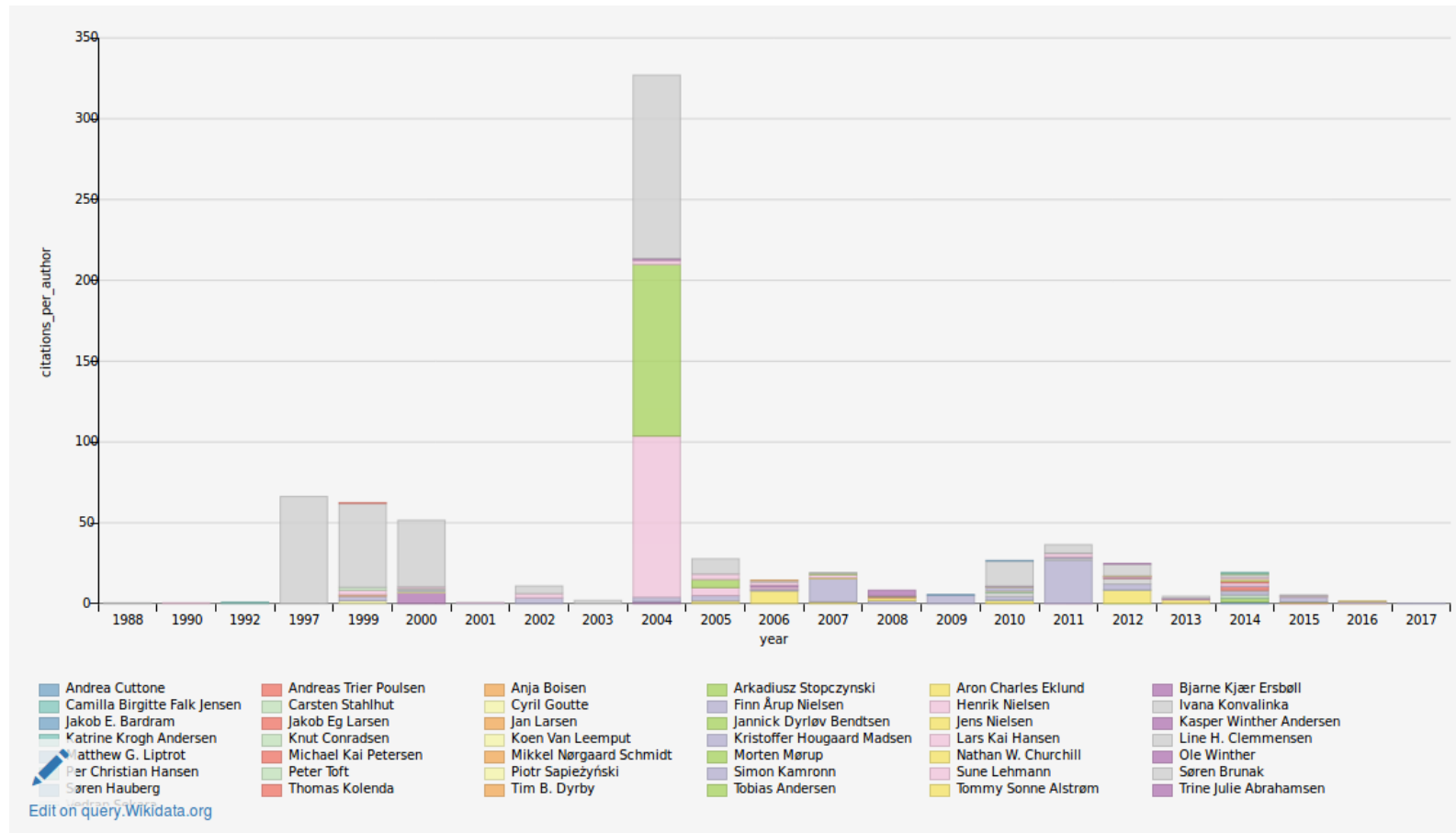


Scholia: Organization aspect



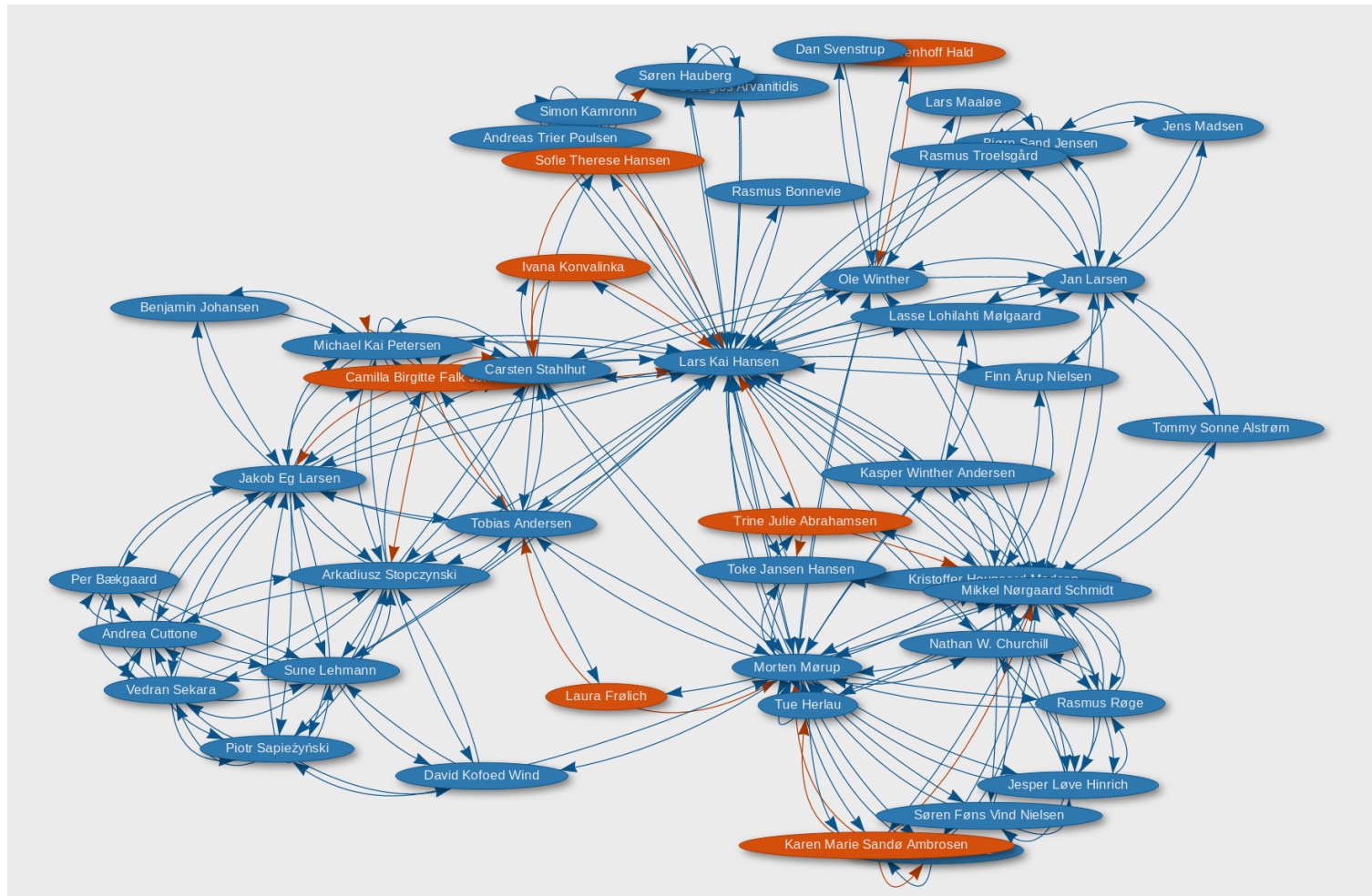
Incomplete statistics on page production per year for **DTU Cognitive Systems**.

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), 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

count	cited_work	cited_workLabel
3	Q24731579	Distributed Representations of Words and Phrases and their Compositionality
3	Q24699014	Efficient Estimation of Word Representations in Vector Space
1	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., Mitraga, 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](https://doi.org/10.1093/DATABASE/BAW015).
- Denning, P., Horning, J., Parnas, D., and Weinstein, L. (2005). Wikipedia risks. 48:152. DOI: [10.1145/1101779.1101804](https://doi.org/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](https://doi.org/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](https://doi.org/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.
- Mitraga, 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](https://doi.org/10.1101/031971).
- Nielsen, F. Å. (2007). [Scientific citations in Wikipedia](#). *First Monday*, 12. DOI: [10.5210/FM.V12I8.1997](https://doi.org/10.5210/FM.V12I8.1997).
- Putman, T. E., Lelong, S., Burgstaller-Muehlbacher, S., Burgstaller-Muehlbacher, 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](https://doi.org/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](https://doi.org/10.5210/FM.V12I4.1763).