An overview of Scholia

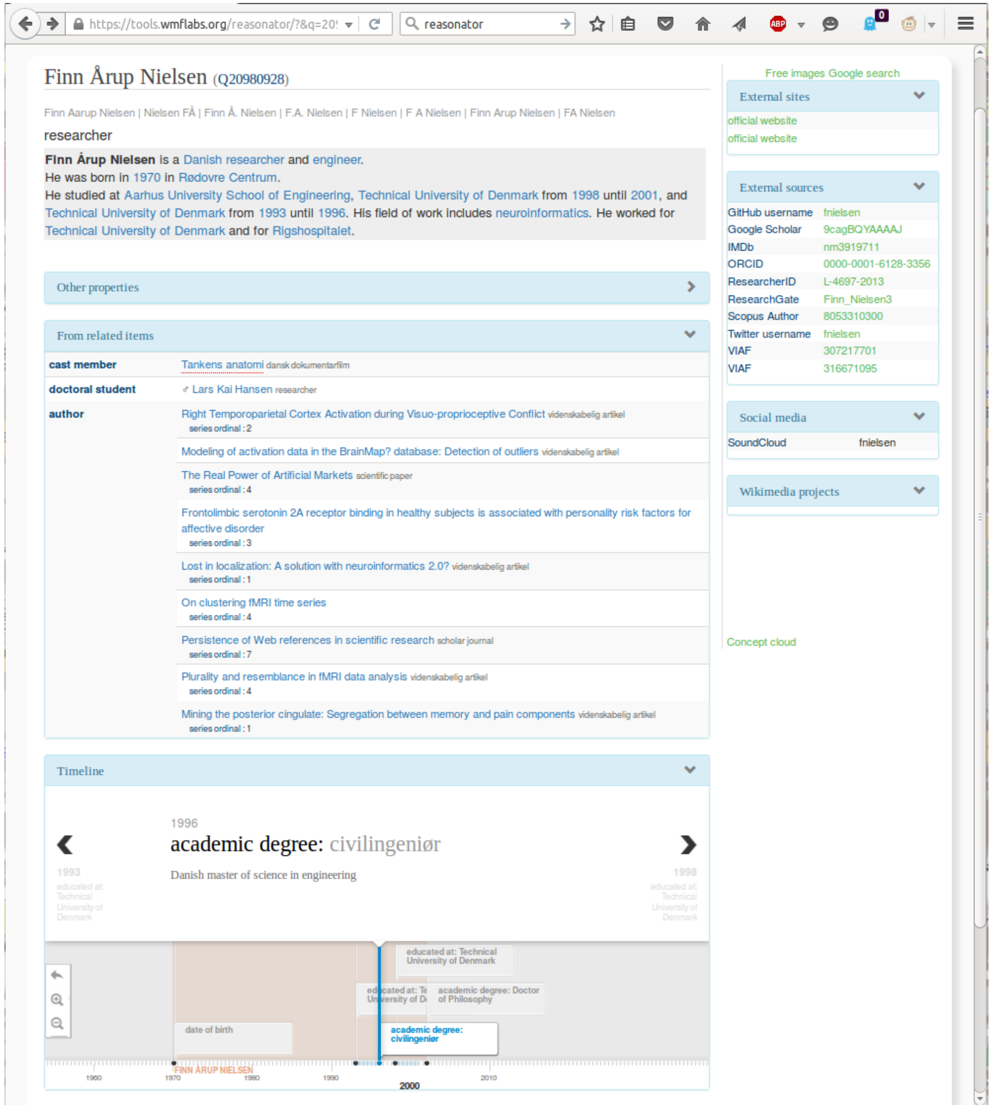
Finn Årup Nielsen

DTU Compute
Technical University of Denmark

May 18, 2017

How do we show data from Wikidata?

Presenting Wikidata: Reasonator



https://tools.wmflabs.org/reasonator/?&q=20! | reasonator

Finn Årup Nielsen (Q20980928)

Finn Aarup Nielsen | Nielsen FÅ | Finn Å. Nielsen | F.A. Nielsen | F Nielsen | F A Nielsen | Finn Arup Nielsen | FA Nielsen

researcher

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

From related items

cast member	Tankens anatomi dansk dokumentarfilm
doctoral student	♂ Lars Kai Hansen researcher
author	Right Temporoparietal Cortex Activation during Visuo-proprioceptive Conflict videnskabelig artikel series ordinal : 2
	Modeling of activation data in the BrainMap? database: Detection of outliers videnskabelig artikel
	The Real Power of Artificial Markets scientific paper series ordinal : 4
	Frontolimbic serotonin 2A receptor binding in healthy subjects is associated with personality risk factors for affective disorder series ordinal : 3
	Lost in localization: A solution with neuroinformatics 2.0? videnskabelig artikel series ordinal : 1
	On clustering fMRI time series series ordinal : 4
	Persistence of Web references in scientific research scholar journal series ordinal : 7
	Plurality and resemblance in fMRI data analysis videnskabelig artikel series ordinal : 4
	Mining the posterior cingulate: Segregation between memory and pain components videnskabelig artikel series ordinal : 1

Timeline

1996

academic degree: civilingeniør

Danish master of science in engineering

1993
educated at: Technical University of Denmark
1998
educated at: Technical University of Denmark
educated at: Technical University of Denmark
educated at: Te University of D
academic degree: Doctor of Philosophy
date of birth
academic degree: civilingeniør
FINN ÅRUP NIELSEN
1960
1970
1980
1990
2000
2010

Free images Google search

External sites

official website
official website

External sources

GitHub username	fnielsen
Google Scholar	9cagBQYAAAAJ
IMDb	nm3919711
ORCID	0000-0001-6128-3356
ResearcherID	L-4697-2013
ResearchGate	Finn_Nielsen3
Scopus Author	8053310300
Twitter username	fnielsen
VIAF	307217701
VIAF	316671095

Social media

SoundCloud	fnielsen

Wikimedia projects

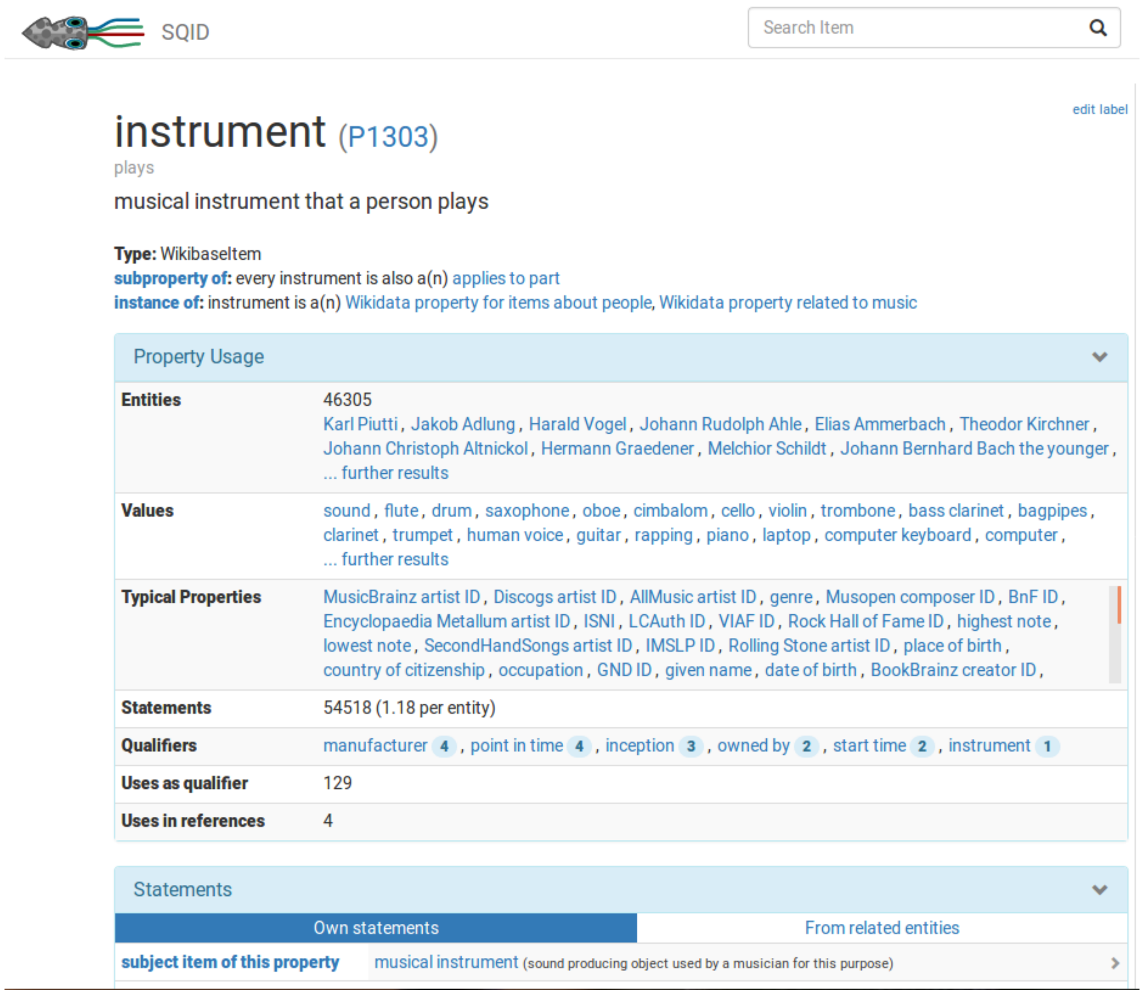
Concept cloud

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*

Presenting Wikidata: SQID



instrument (P1303)

plays

musical instrument that a person plays

Type: WikibaseItem
subproperty of: every instrument is also a(n) applies to part
instance of: instrument is a(n) Wikidata property for items about people, Wikidata property related to music

Property Usage

Entities	46305 Karl Piutti, Jakob Adlung, Harald Vogel, Johann Rudolph Ahle, Elias Ammerbach, Theodor Kirchner, Johann Christoph Altnickol, Hermann Graedener, Melchior Schildt, Johann Bernhard Bach the younger, ... further results
Values	sound, flute, drum, saxophone, oboe, cimbalom, cello, violin, trombone, bass clarinet, bagpipes, clarinet, trumpet, human voice, guitar, rapping, piano, laptop, computer keyboard, computer, ... further results
Typical Properties	MusicBrainz artist ID, Discogs artist ID, AllMusic artist ID, genre, Musopen composer ID, BnF ID, Encyclopaedia Metallum artist ID, ISNI, LCAuth ID, VIAF ID, Rock Hall of Fame ID, highest note, lowest note, SecondHandSongs artist ID, IMSLP ID, Rolling Stone artist ID, place of birth, country of citizenship, occupation, GND ID, given name, date of birth, BookBrainz creator ID,
Statements	54518 (1.18 per entity)
Qualifiers	manufacturer 4, point in time 4, inception 3, owned by 2, start time 2, instrument 1
Uses as qualifier	129
Uses in references	4

Statements

Own statements | From related entities

subject item of this property musical instrument (sound producing object used by a musician for this purpose)

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

Wikidata class browser.

Displays typical properties

Runs from *Wikimedia Tool Labs*

Scholia

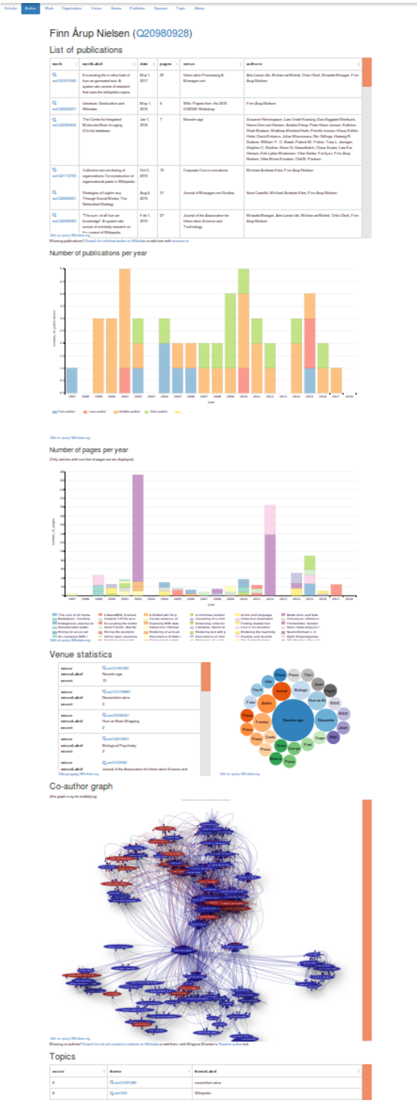
How can we show scientific (bibliographic) data from Wikidata?

Scholia

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



Finn Årup Nielsen (Q20980928)

List of publications

Number of publications per year

Number of pages per year

Venue statistics

Co-author graph

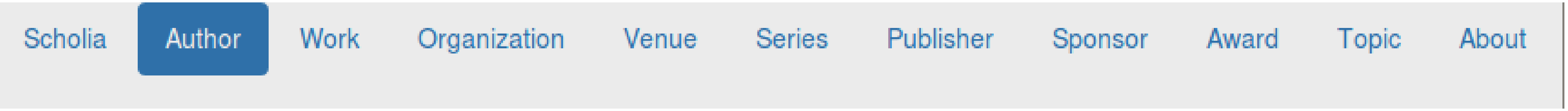
Topics

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".

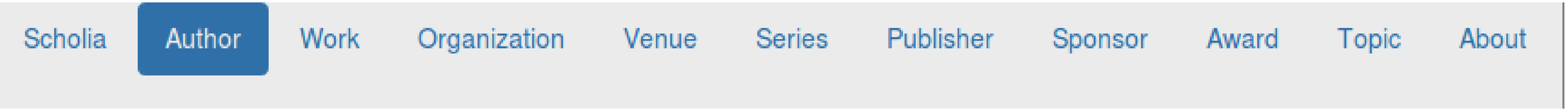
“Aspects”



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"

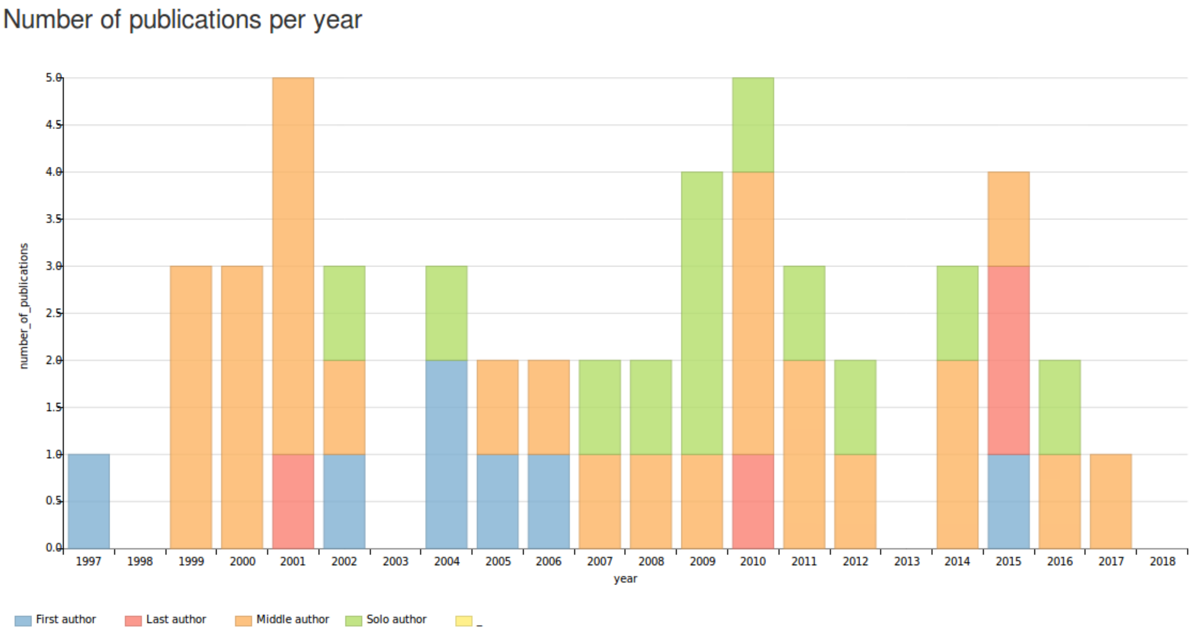


Scholia presents the data in different "aspects": author, work, organization (e.g., university, research group), venue (journal or conference), series (e.g., conference proceedings series), publisher, sponsor, award, topic.

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



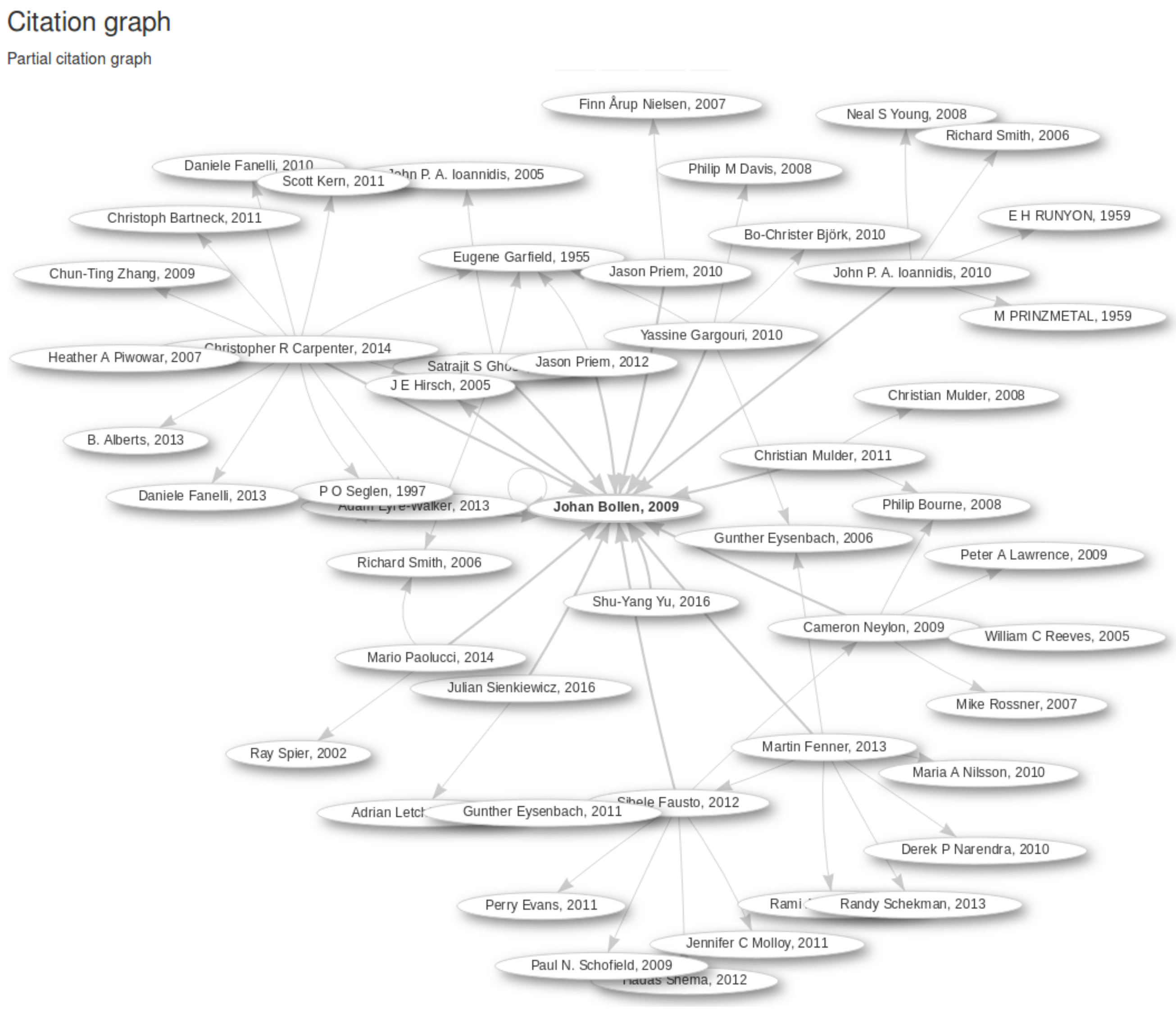
Inspired by Shubhanshu Mishra's and Vetle I. Torvik's LEGOLAS visualization.

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



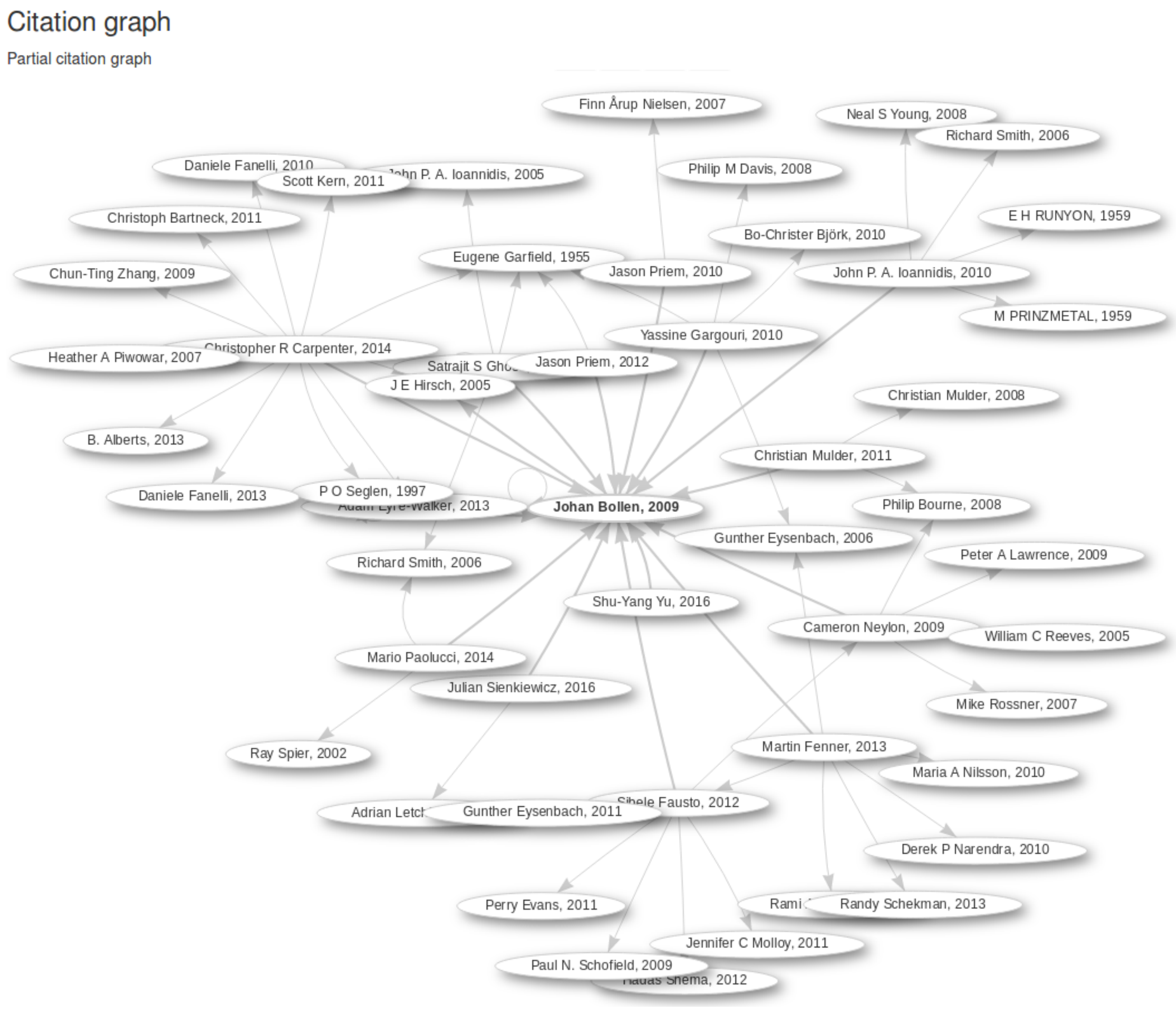
Citation graph

Partial 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 graph

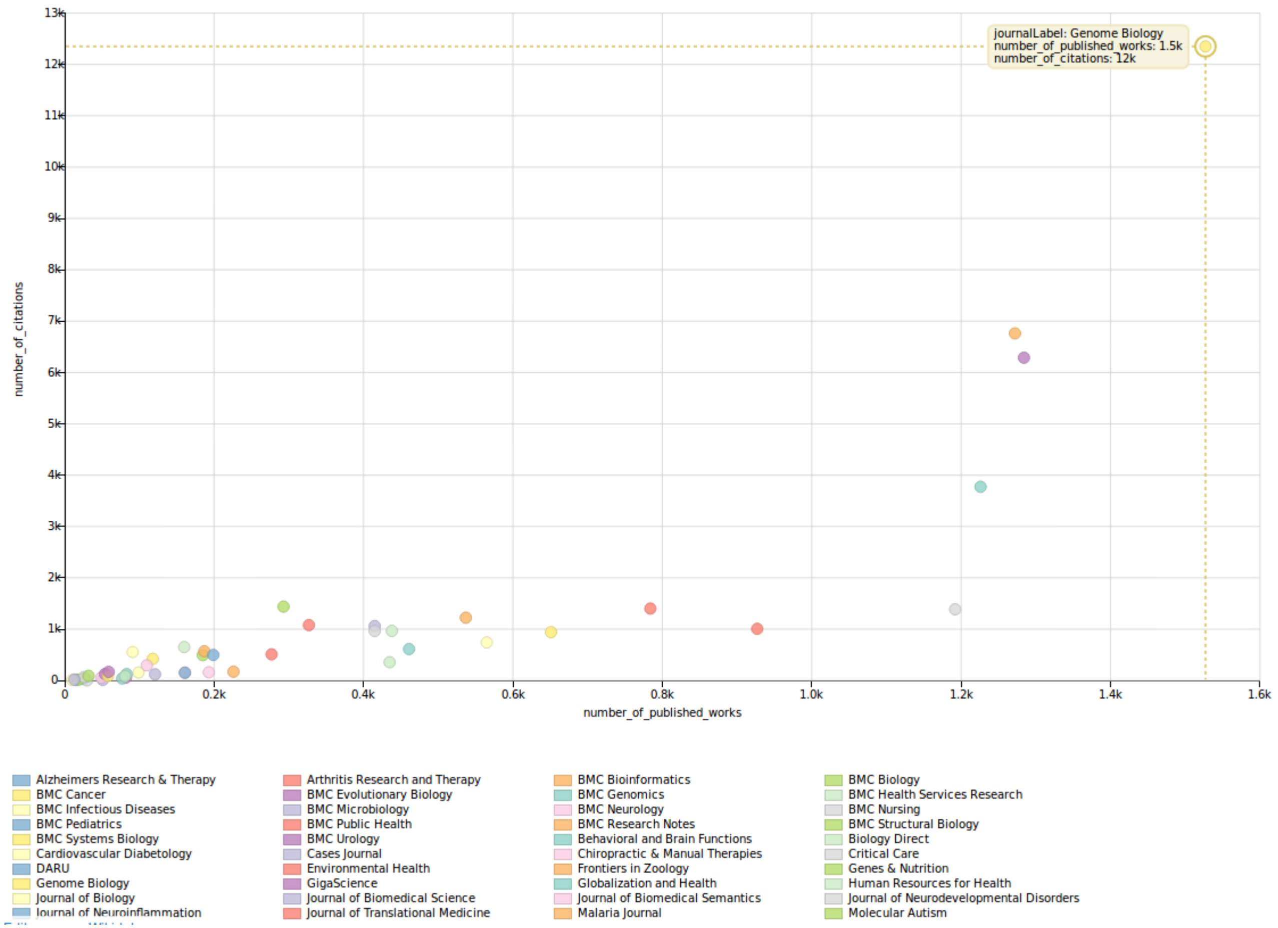
Partial citation graph

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.

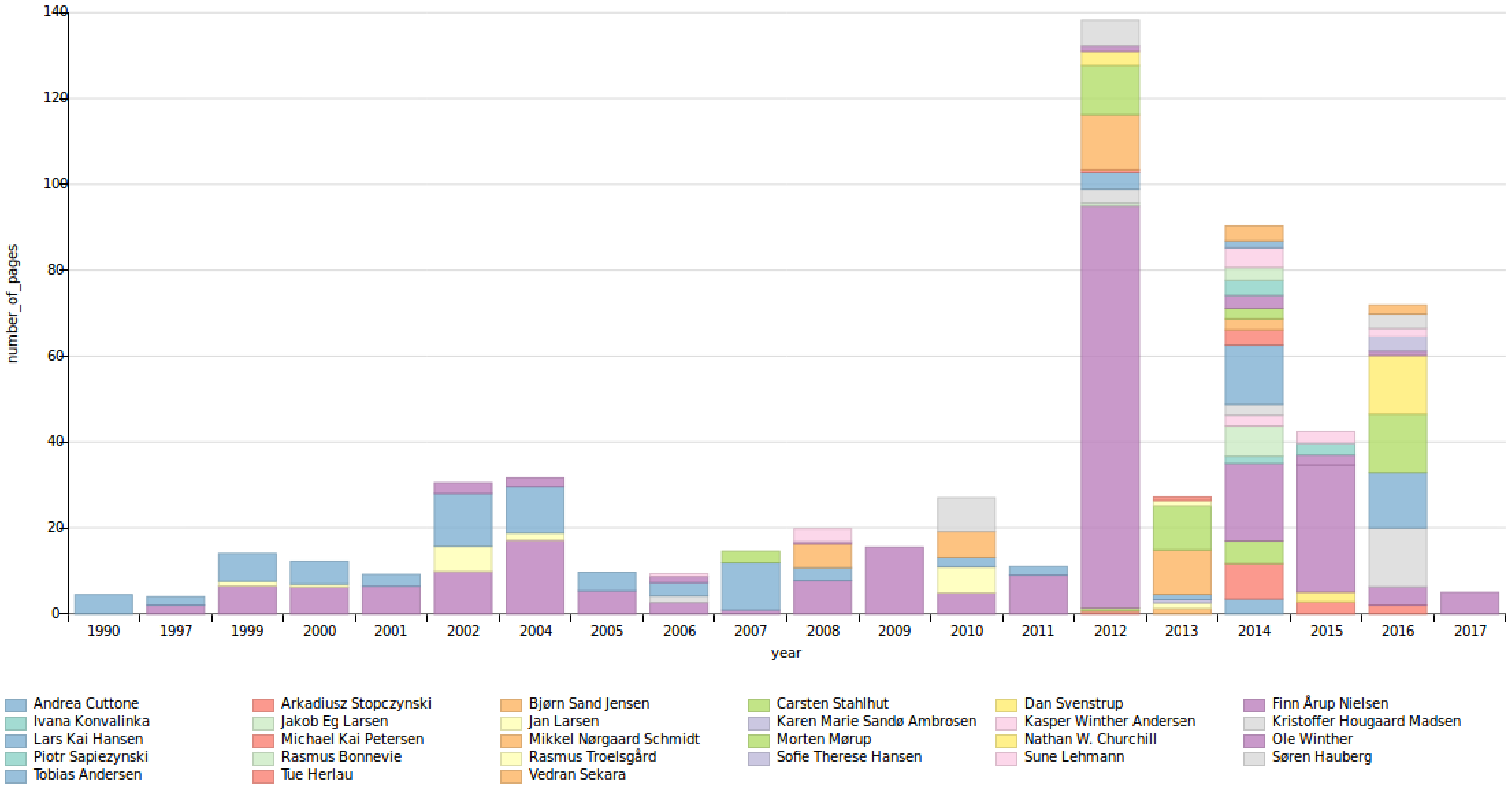
Scholia: Publisher aspect



Panel on publisher aspect with an overview of number of papers published and their citations across journals published by the publisher.

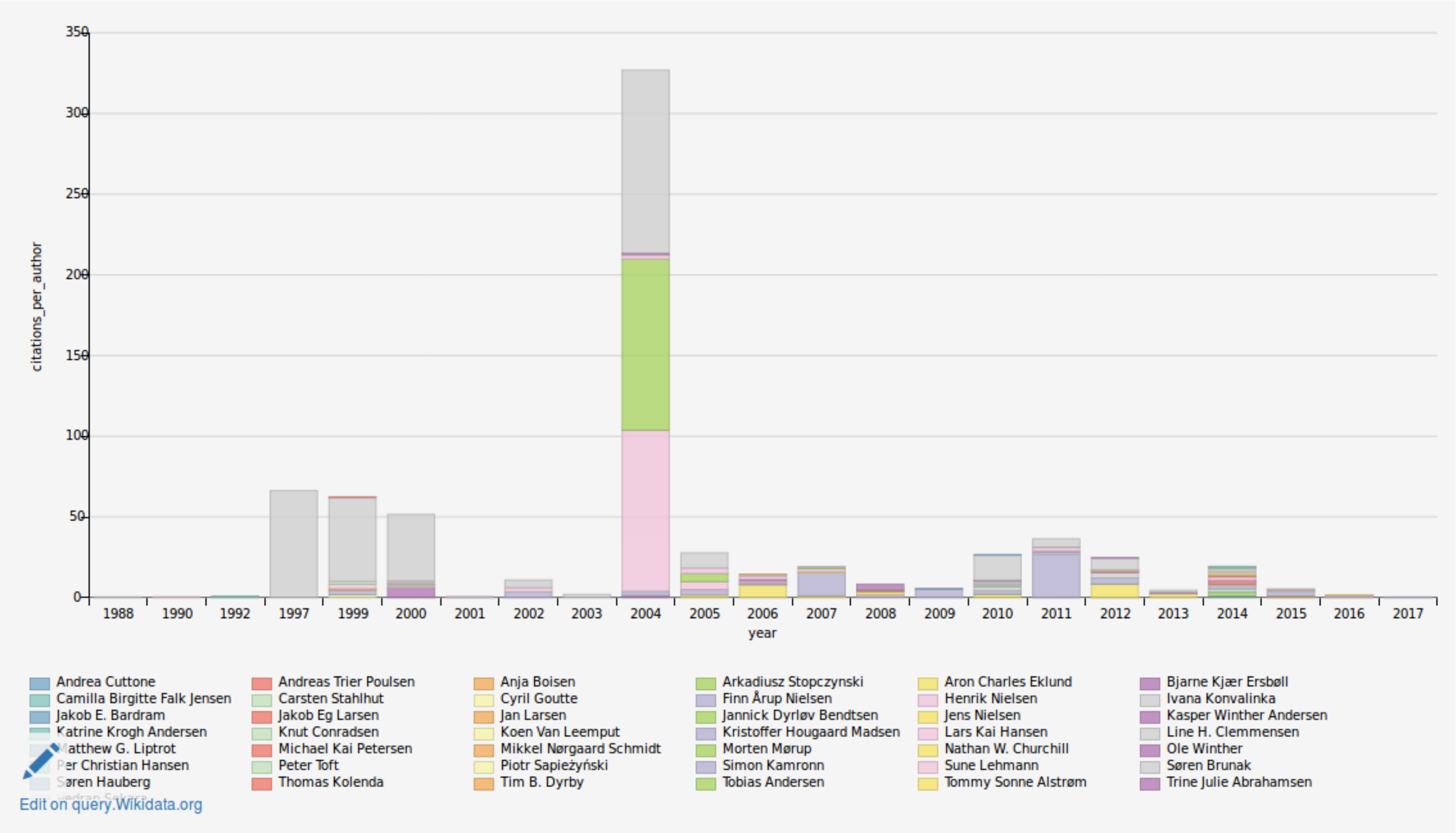
Here for BioMedCentral (which may be an imprint)

Scholia: Organization aspect



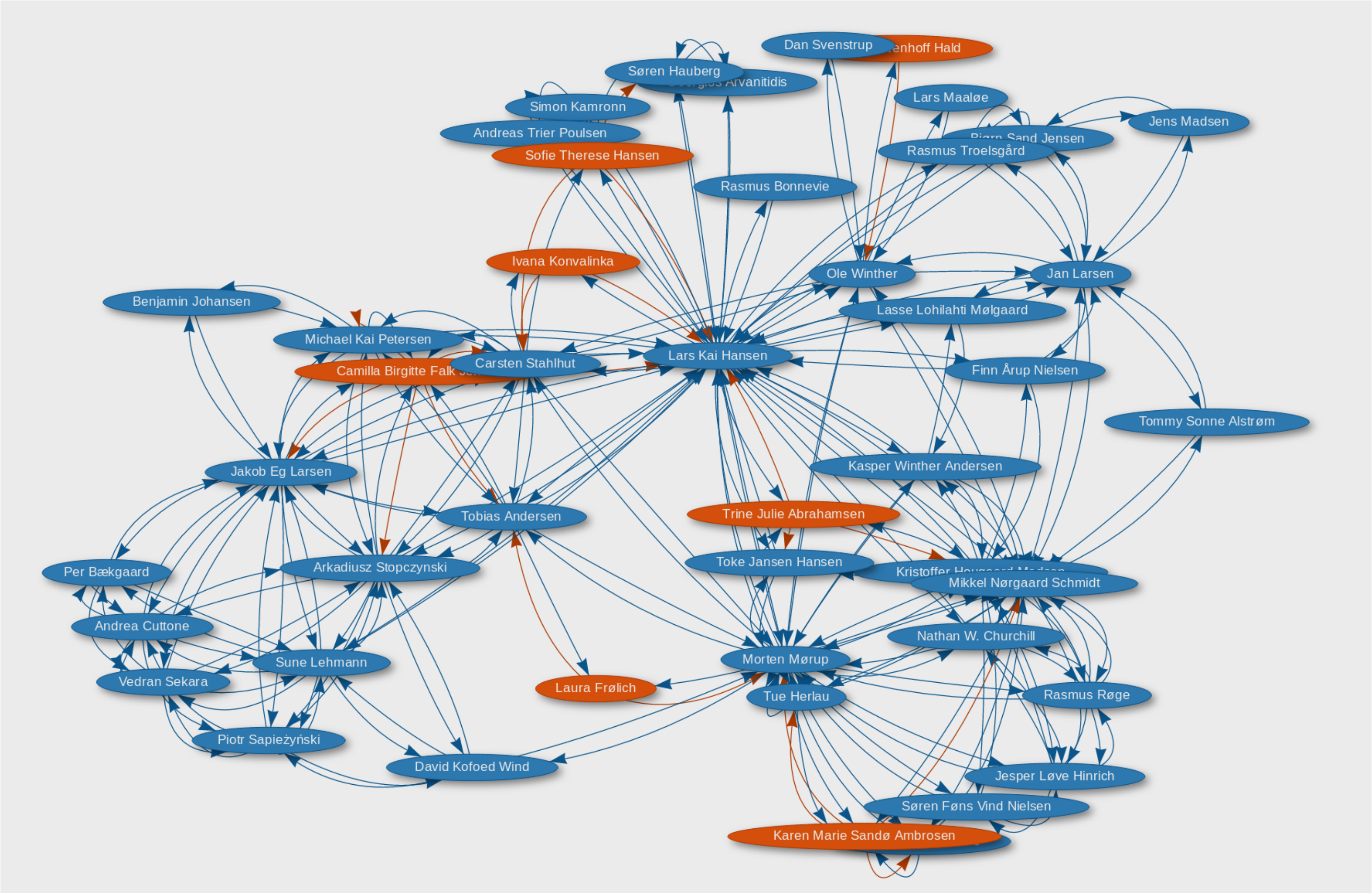
Incomplete statistics on page production per year for DTU Cognitive Systems.

Scholia: Organization aspect



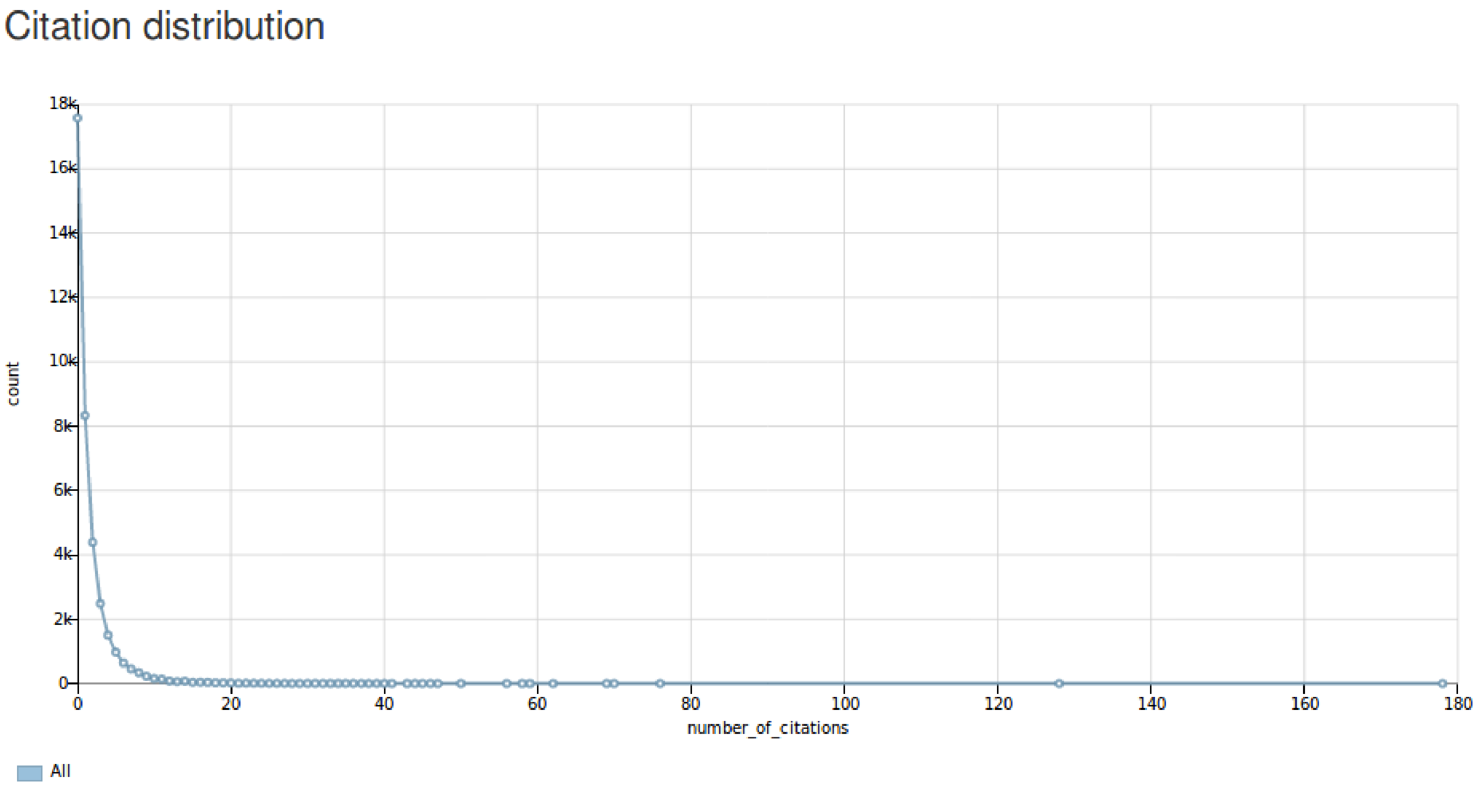
Co-author-normalized citations per year for Technical University of Denmark.

Scholia: Organization aspect



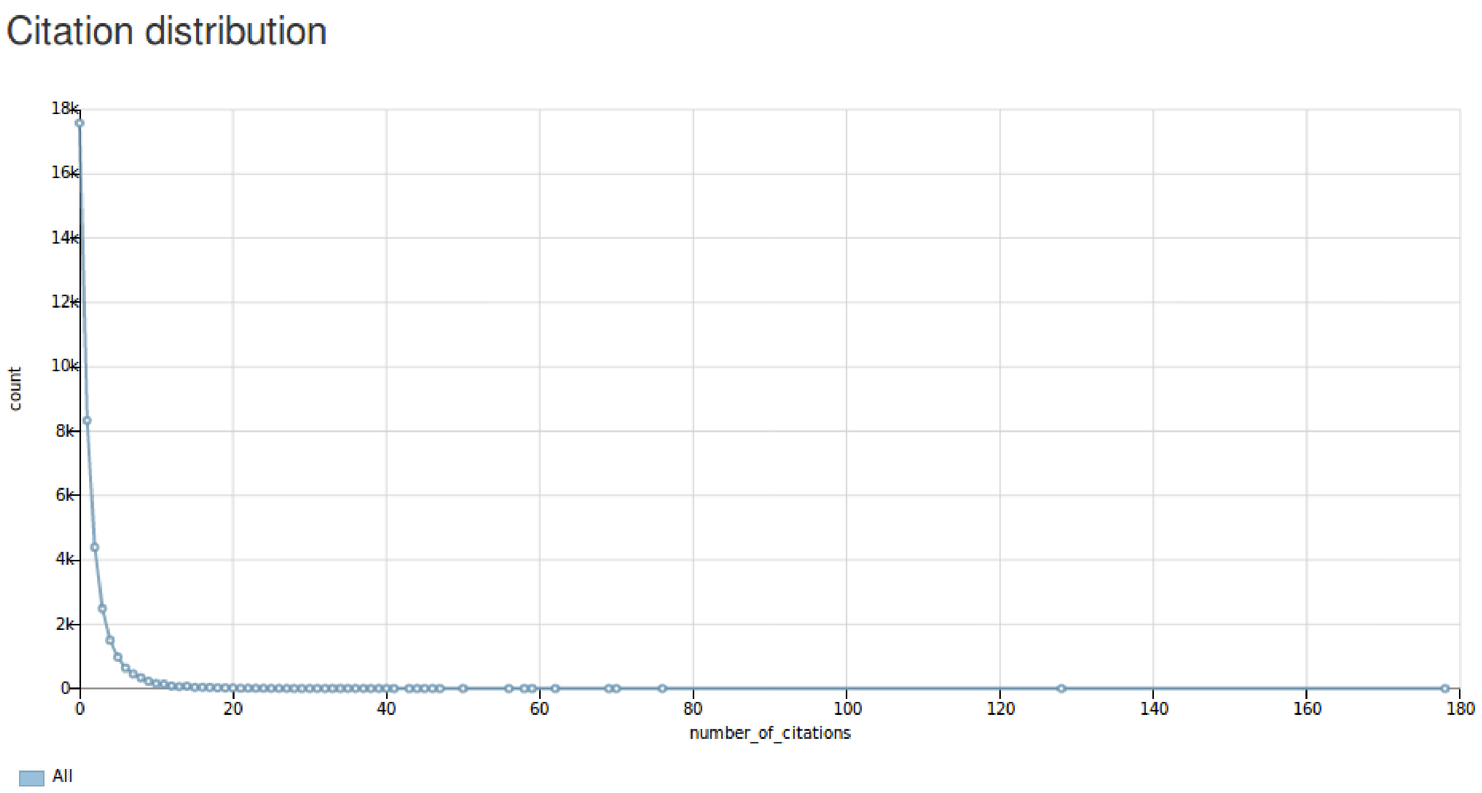
Co-author graph for DTU Cognitive Systems.

Citation distribution



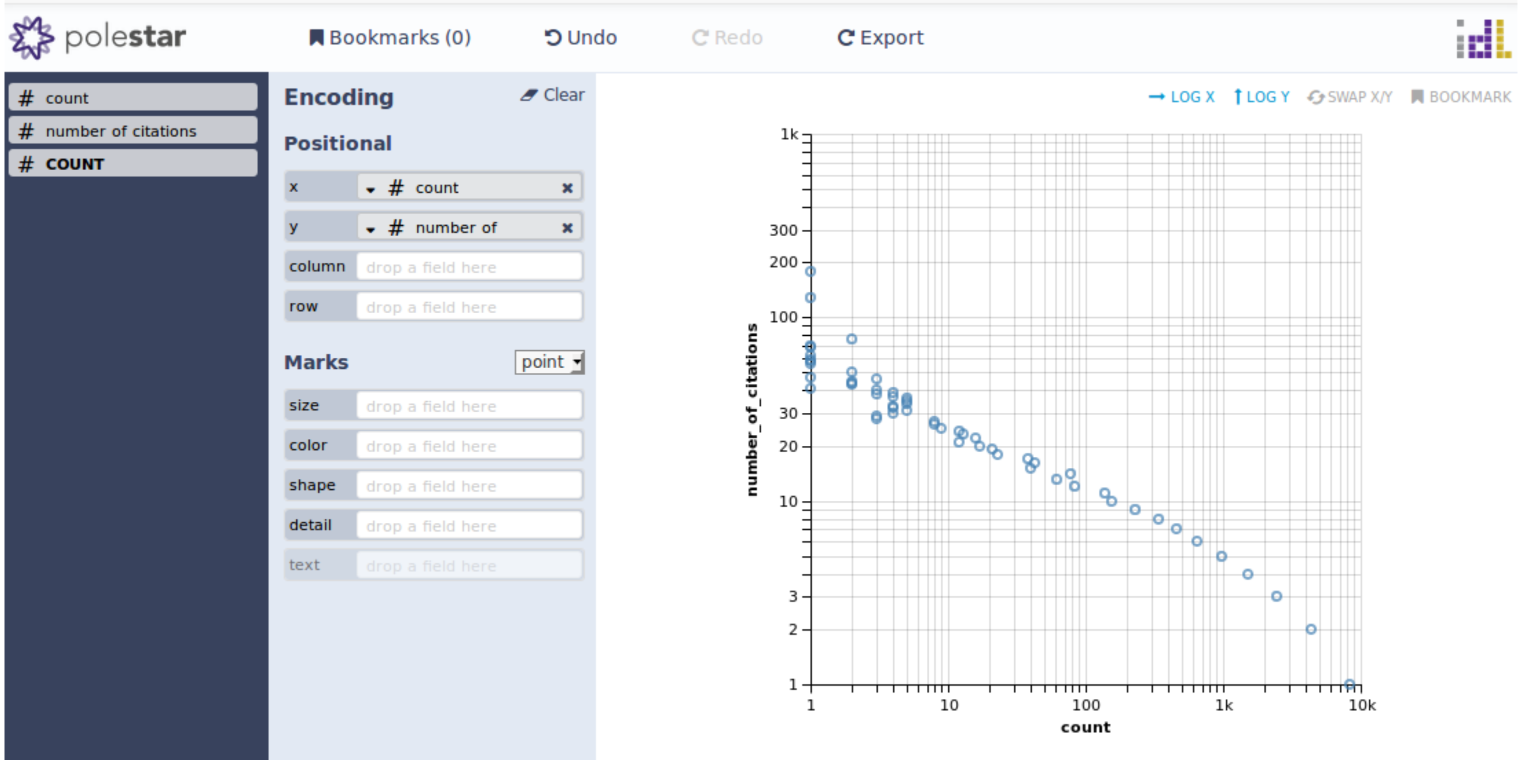
Citation distribution for PLOS ONE.

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.

Scholia

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

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.

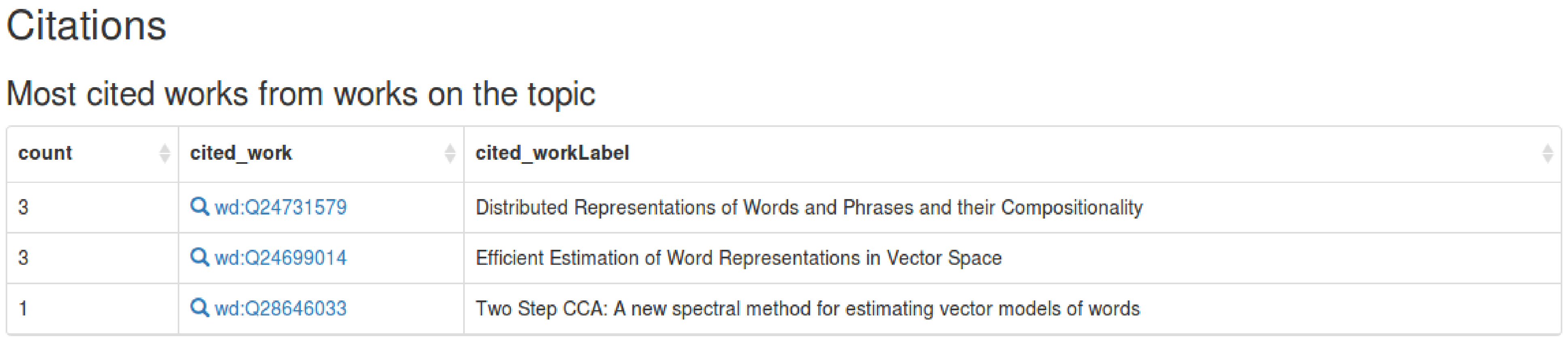
Give me an introductory paper

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.

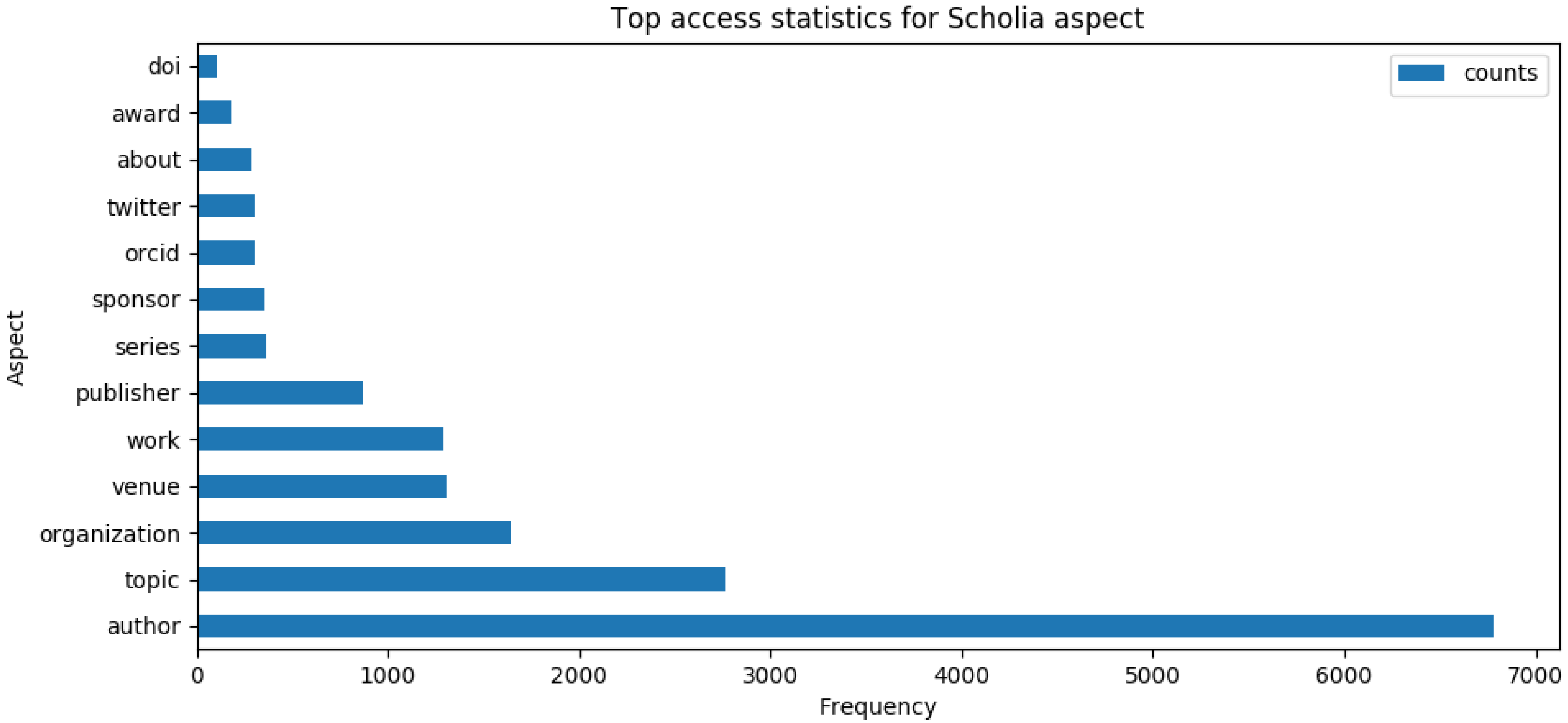


Citations

Most cited works from works on the topic

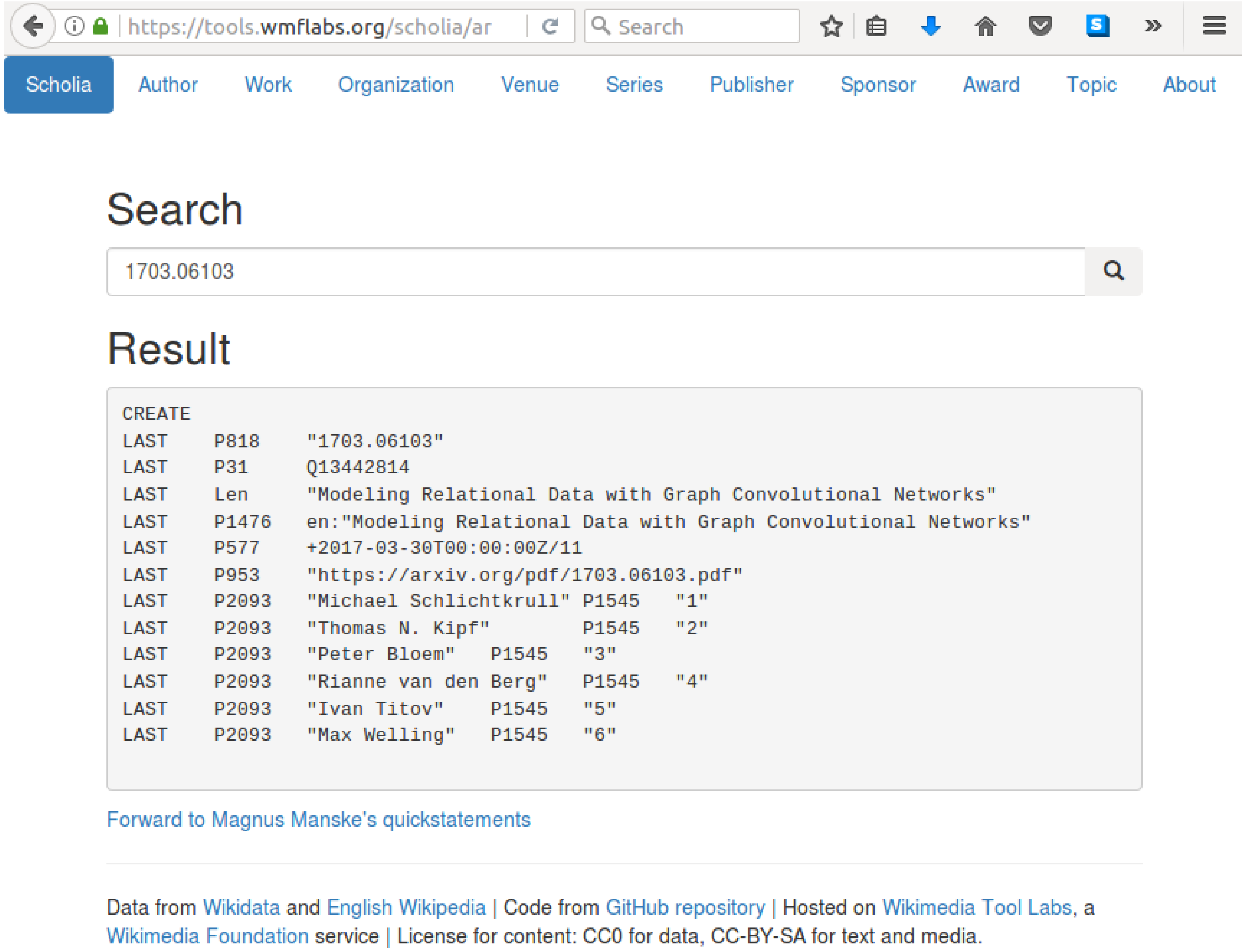
count	cited_work	cited_workLabel
3	wd:Q24731579	Distributed Representations of Words and Phrases and their Compositionality
3	wd:Q24699014	Efficient Estimation of Word Representations in Vector Space
1	wd:Q28646033	Two Step CCA: A new spectral method for estimating vector models of words

Scholia access statistics



Based on WMF toollabs' `uwsgi.log` log file with anonymized IP address.

Data entry: arxiv-to-quickstatements



Search

1703.06103

Result

```
CREATE
LAST    P818    "1703.06103"
LAST    P31     Q13442814
LAST    Len     "Modeling Relational Data with Graph Convolutional Networks"
LAST    P1476   en:"Modeling Relational Data with Graph Convolutional Networks"
LAST    P577    +2017-03-30T00:00:00Z/11
LAST    P953    "https://arxiv.org/pdf/1703.06103.pdf"
LAST    P2093   "Michael Schlichtkrull" P1545   "1"
LAST    P2093   "Thomas N. Kipf"        P1545   "2"
LAST    P2093   "Peter Bloem"   P1545   "3"
LAST    P2093   "Rianne van den Berg"   P1545   "4"
LAST    P2093   "Ivan Titov"    P1545   "5"
LAST    P2093   "Max Welling"   P1545   "6"
```

Forward to Magnus Manske's quickstatements

Data from Wikidata and English Wikipedia | Code from GitHub repository | Hosted on Wikimedia Tool Labs, a Wikimedia Foundation service | License for content: CC0 for data, CC-BY-SA for text and media.

Lookup ID on arXiv homepage, extract metadata and format it for Magnus Manske's quickstatement webservice.

Wikidata-based BIBTeX generation

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

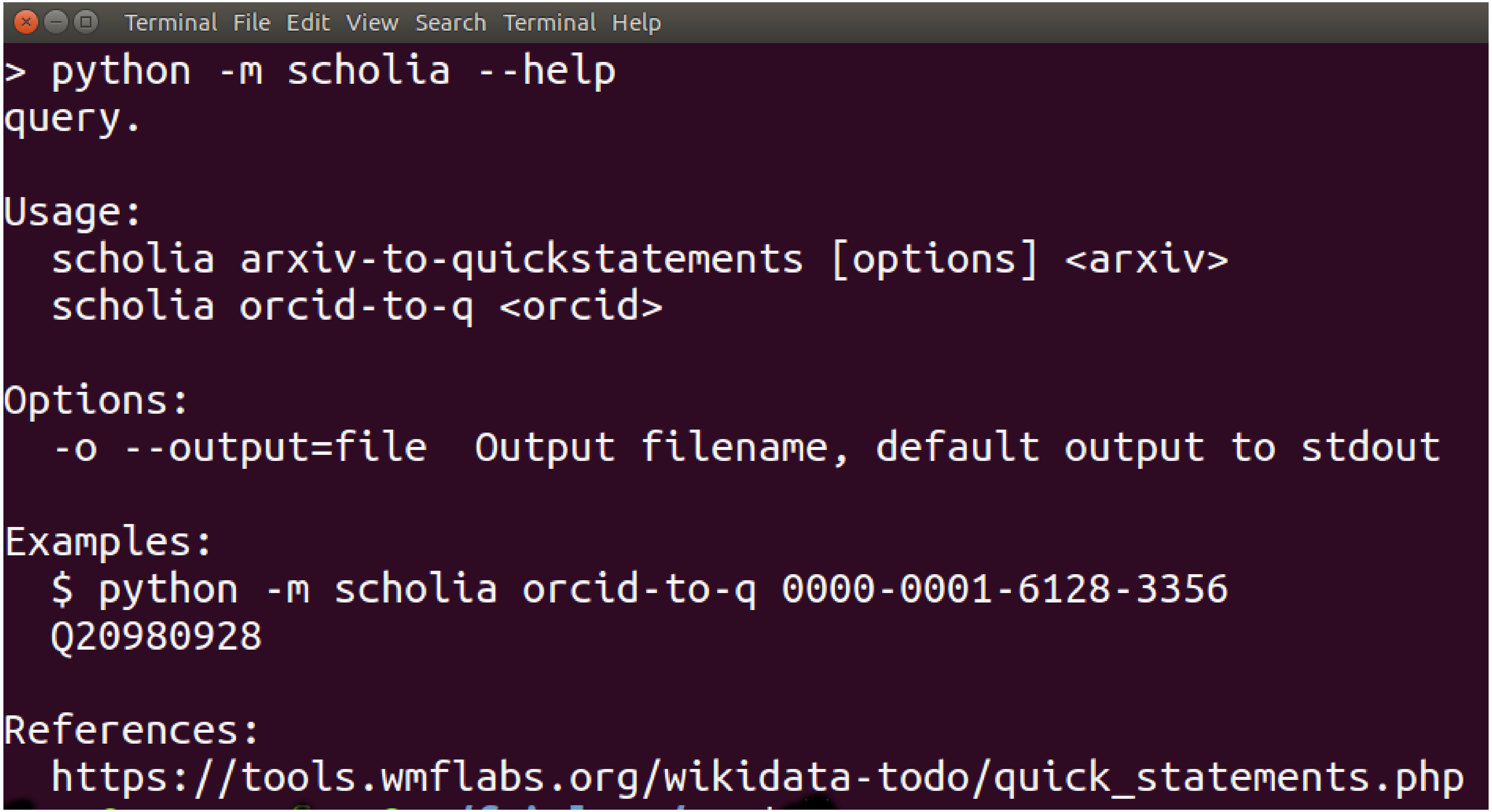
My .tex file:

```
\bibliographystyle{Nielsen2012Slides}
\bibliography{Nielsen2017Overview_slides}
```

Commands:

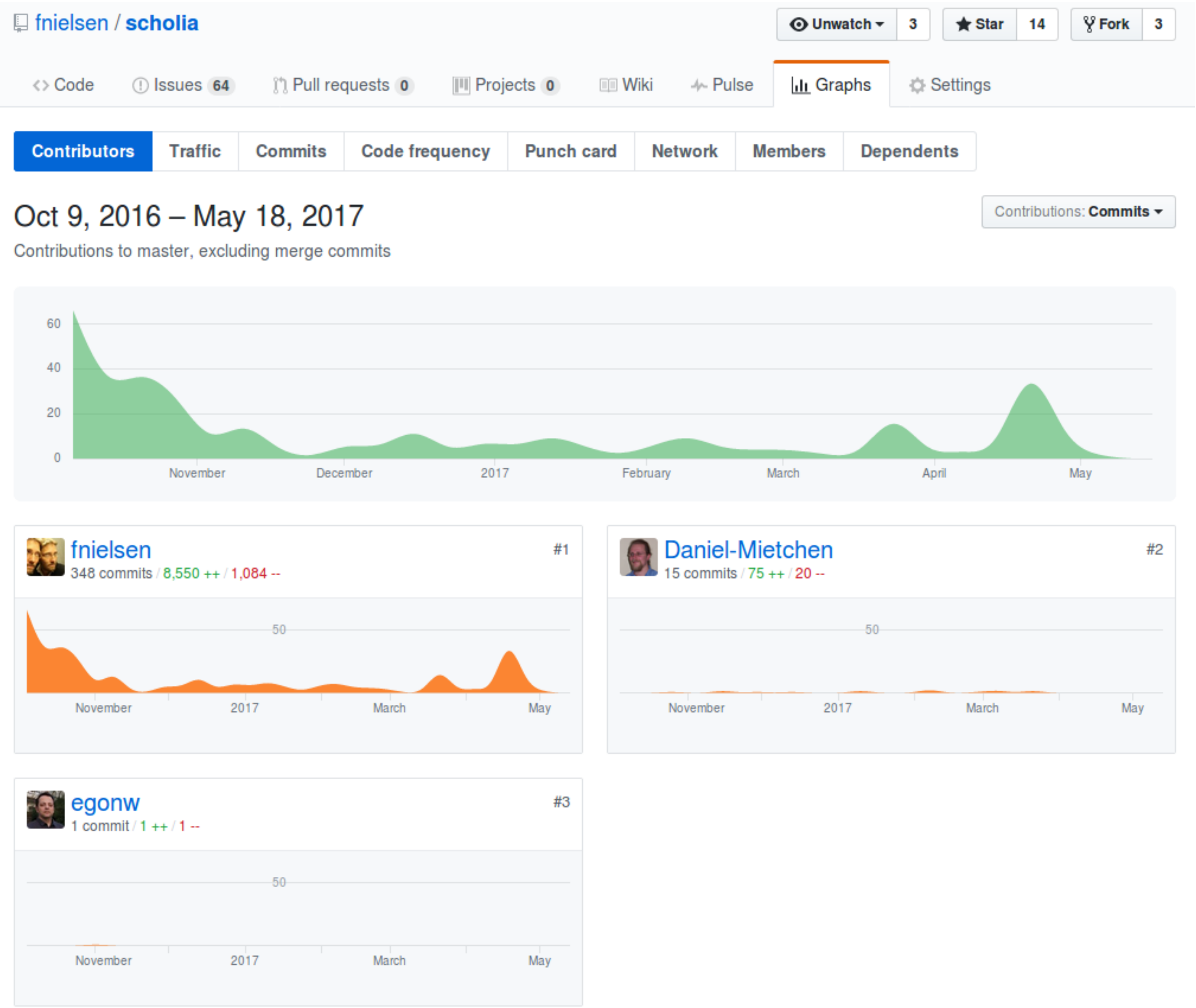
```
latex Nielsen2017Overview_slides.tex
python -m scholia.tex write-bib-from-aux Nielsen2017Overview_slides.aux
bibtex Nielsen2017Overview_slides
latex Nielsen2017Overview_slides.tex
latex Nielsen2017Overview_slides.tex
```

More command-line interfacing



```
> python -m scholia --help
query.

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



Oct 9, 2016 – May 18, 2017

Contributions to master, excluding merge commits

fnielsen
348 commits / 8,550 ++ / 1,084 --

Daniel-Mietchen
15 commits / 75 ++ / 20 --

egonw
1 commit / 1 ++ / 1 --

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

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.

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.

Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2017). Scholia and scientometrics with Wikidata.