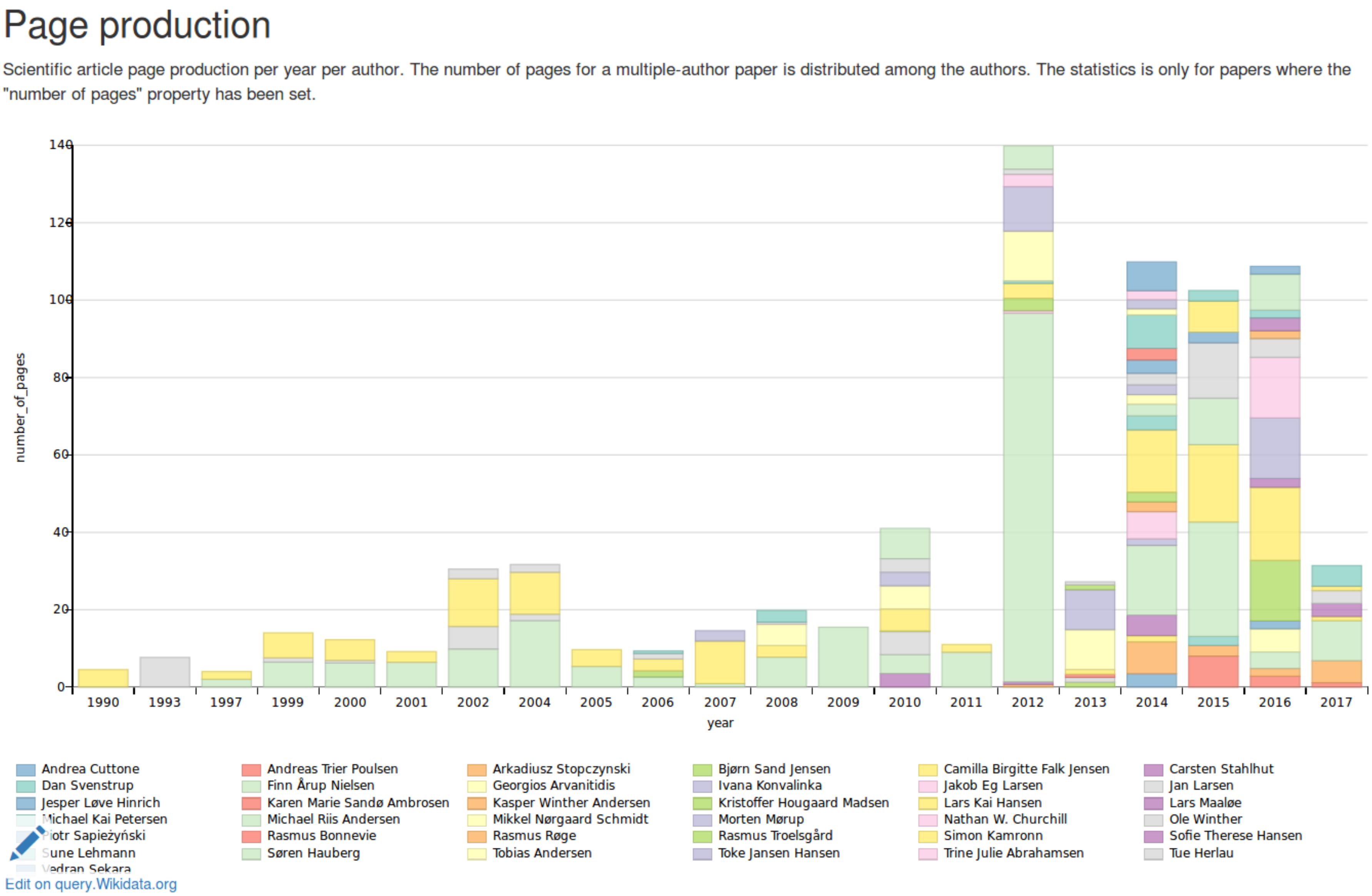
Scholia: A Wikidata-based site for analytics and visualization of science

Finn Årup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

3 oktober 2018

Scholia

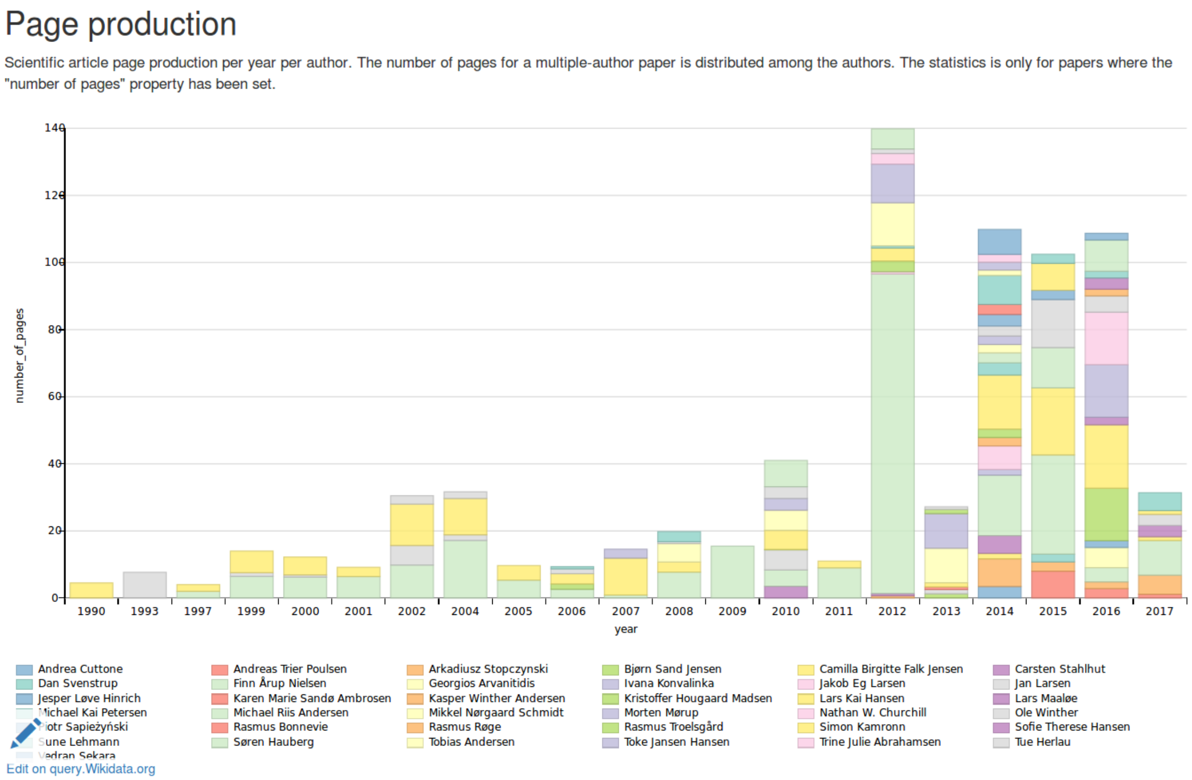


Page production

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.

Edit on query.Wikidata.org

Scholia



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.

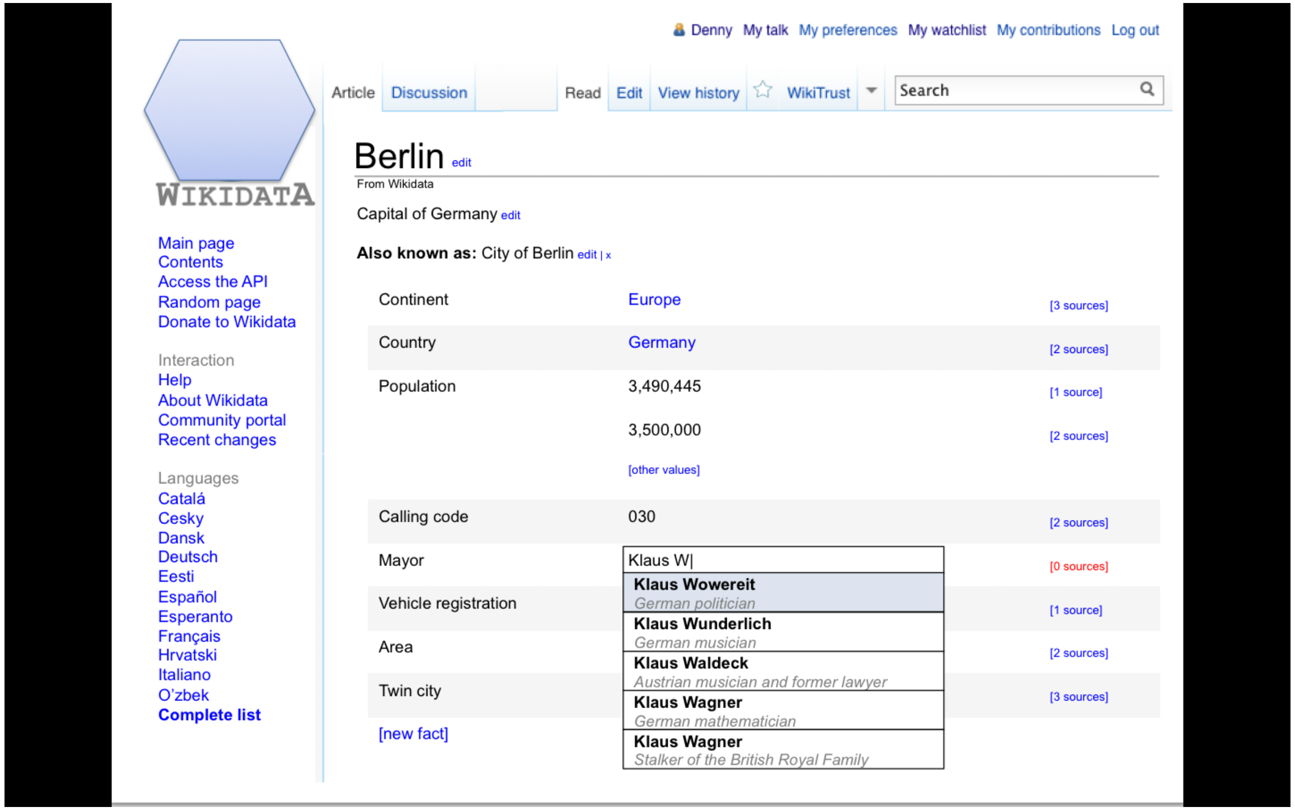
The webservice generates overview of science with *Wikidata Query Service* and is built with the Flask web framework, HTML, Bootstrap, Javascript and templated SPARQL.

For researcher profiles, scientometrics, bibliographic reference management, information discovery (find relevant papers, scientific meetings, researchers, funding opportunities, . . .).

Where does the data comes from?

WIKIDATA

Wikidata



Berlin edit

From Wikidata

Capital of Germany edit

Also known as: City of Berlin edit | x

Continent	Europe	[3 sources]
Country	Germany	[2 sources]
Population	3,490,445	[1 source]
	3,500,000	[2 sources]
	[other values]	
Calling code	030	[2 sources]
Mayor	Klaus W|	[0 sources]
Vehicle registration		[1 source]
Area		[2 sources]
Twin city		[3 sources]
[new fact]		

Klaus Wowereit
German politician
Klaus Wunderlich
German musician
Klaus Waldeck
Austrian musician and former lawyer
Klaus Wagner
German mathematician
Klaus Wagner
Stalker of the British Royal Family

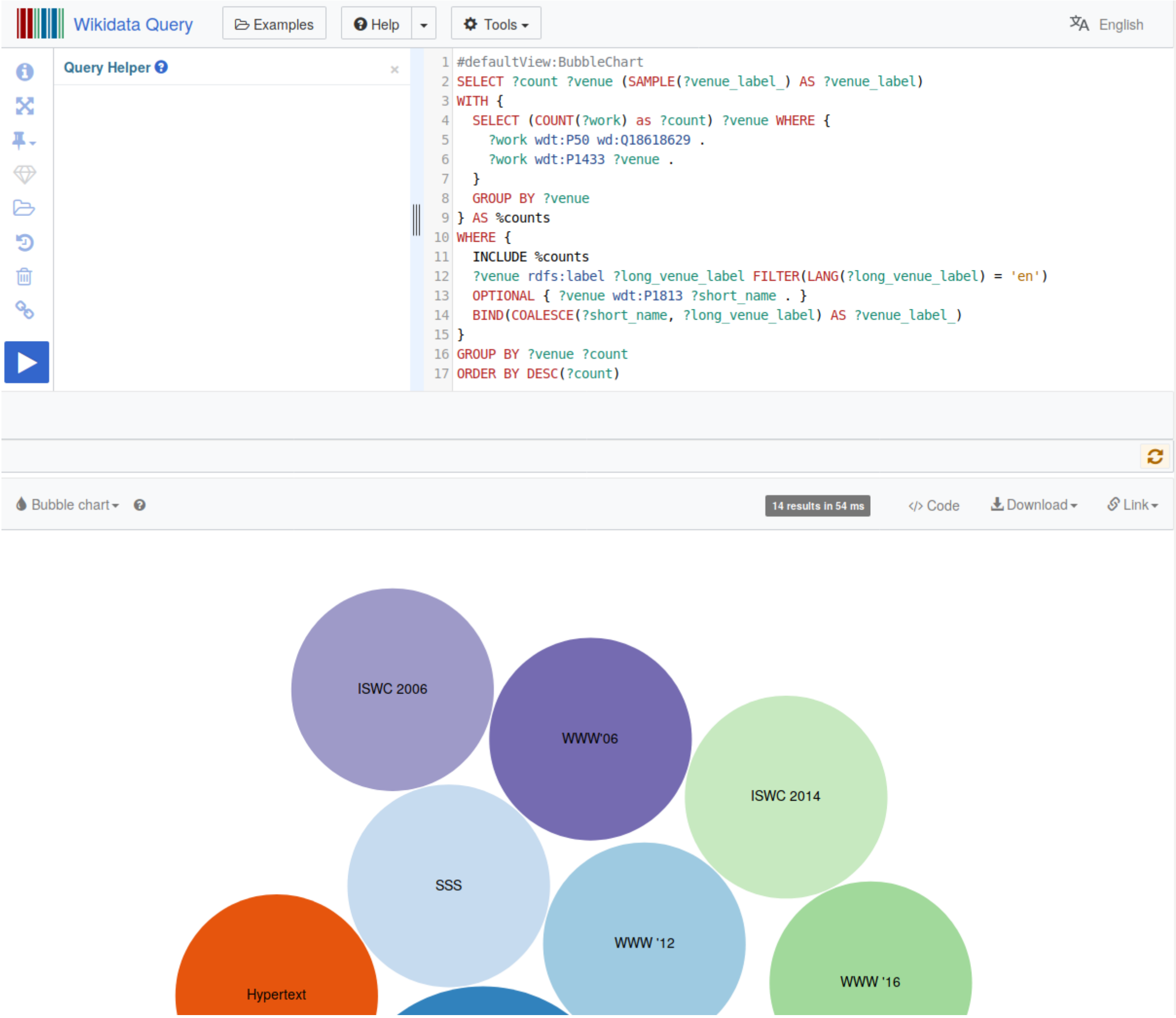
"Wikidata: Verifiable, Linked Open Knowledge That Anyone Can edit" (Dario Taraborelli)

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



```
#defaultView:BubbleChart
SELECT ?count ?venue (SAMPLE(?venue_label_) AS ?venue_label)
WITH {
  SELECT (COUNT(?work) as ?count) ?venue WHERE {
    ?work wdt:P50 wd:Q18618629 .
    ?work wdt:P1433 ?venue .
  }
  GROUP BY ?venue
} AS %counts
WHERE {
  INCLUDE %counts
  ?venue rdfs:label ?long_venue_label FILTER(LANG(?long_venue_label) = 'en')
  OPTIONAL { ?venue wdt:P1813 ?short_name . }
  BIND(COALESCE(?short_name, ?long_venue_label) AS ?venue_label_)
}
GROUP BY ?venue ?count
ORDER BY DESC(?count)
```

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.

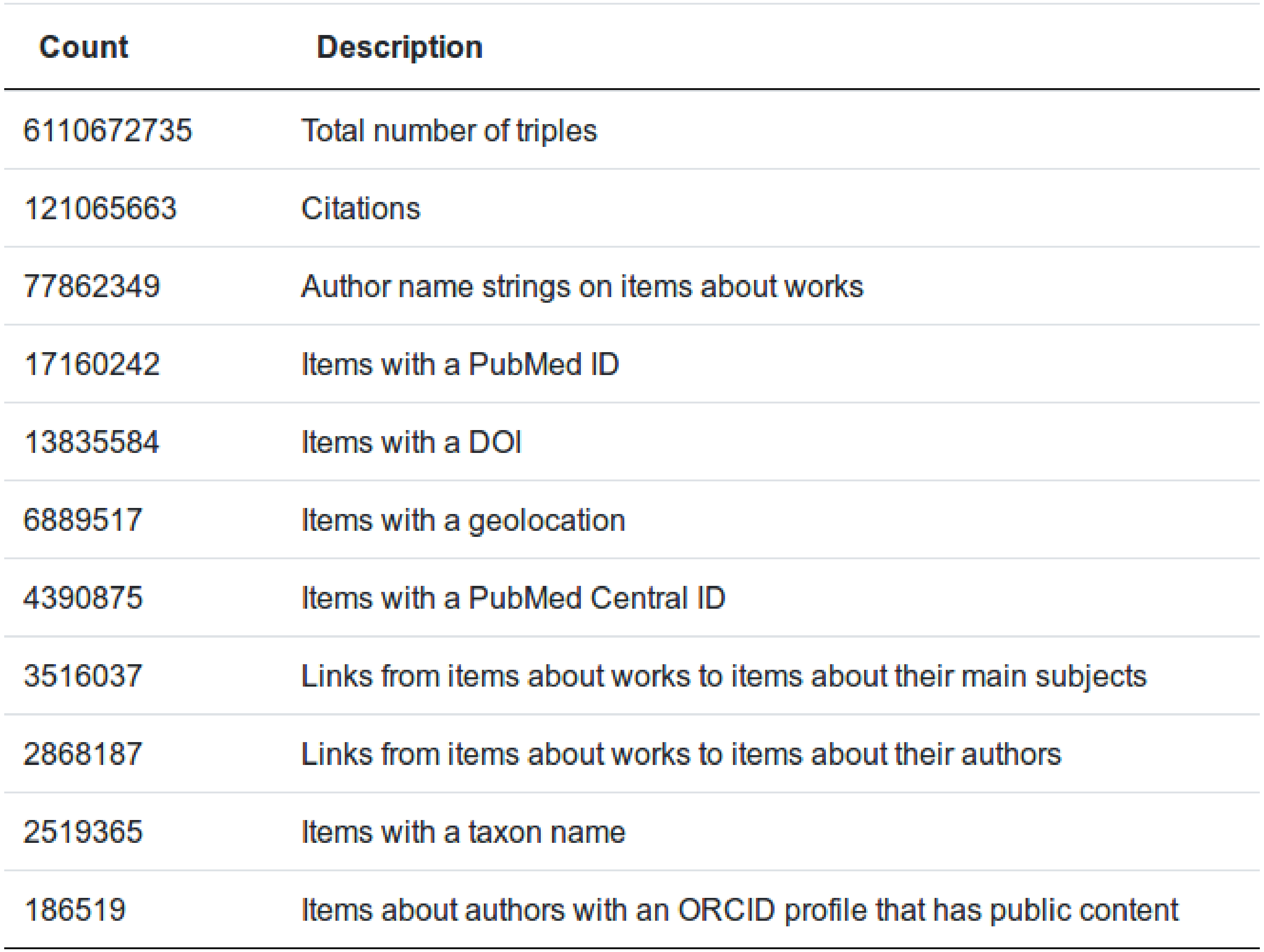
WikiCite

“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



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

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

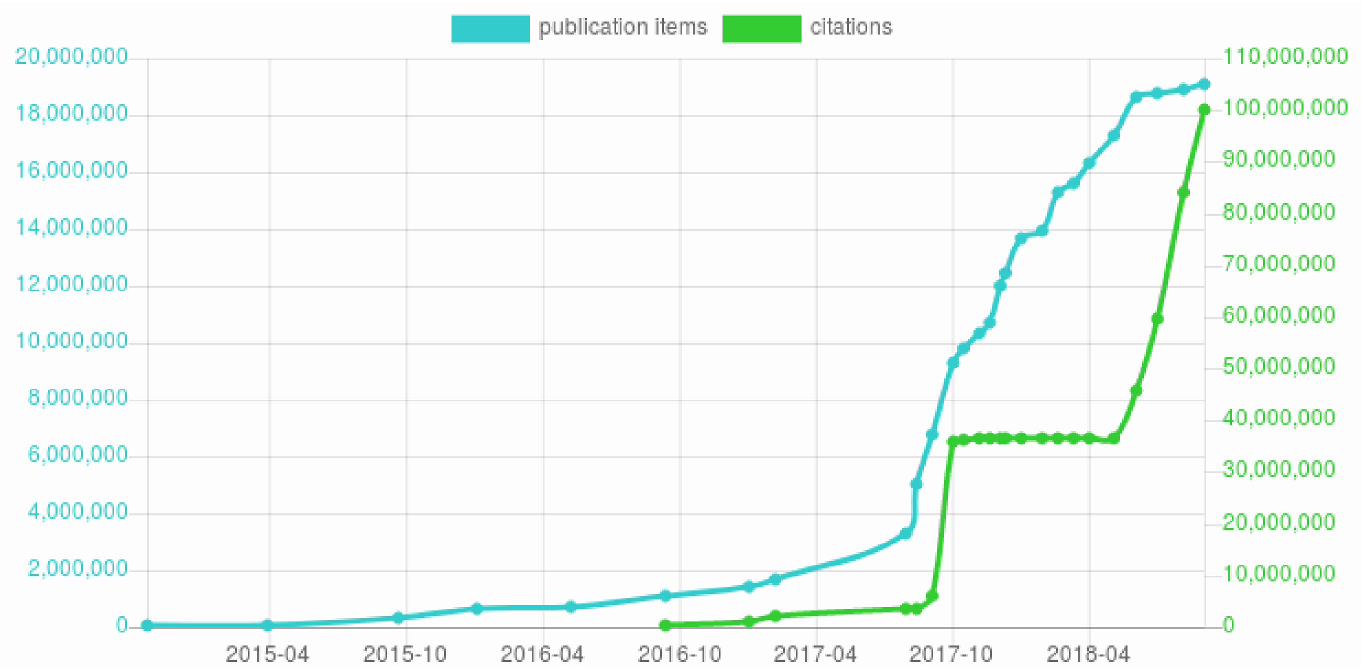
121 million citations.

17 million PubMed links.

14 million DOI links.

187 thousand ORCID links.

Jakob Voß' WikiCite statistics



Jakob Voß' Wikicite statistics that is update regularly.

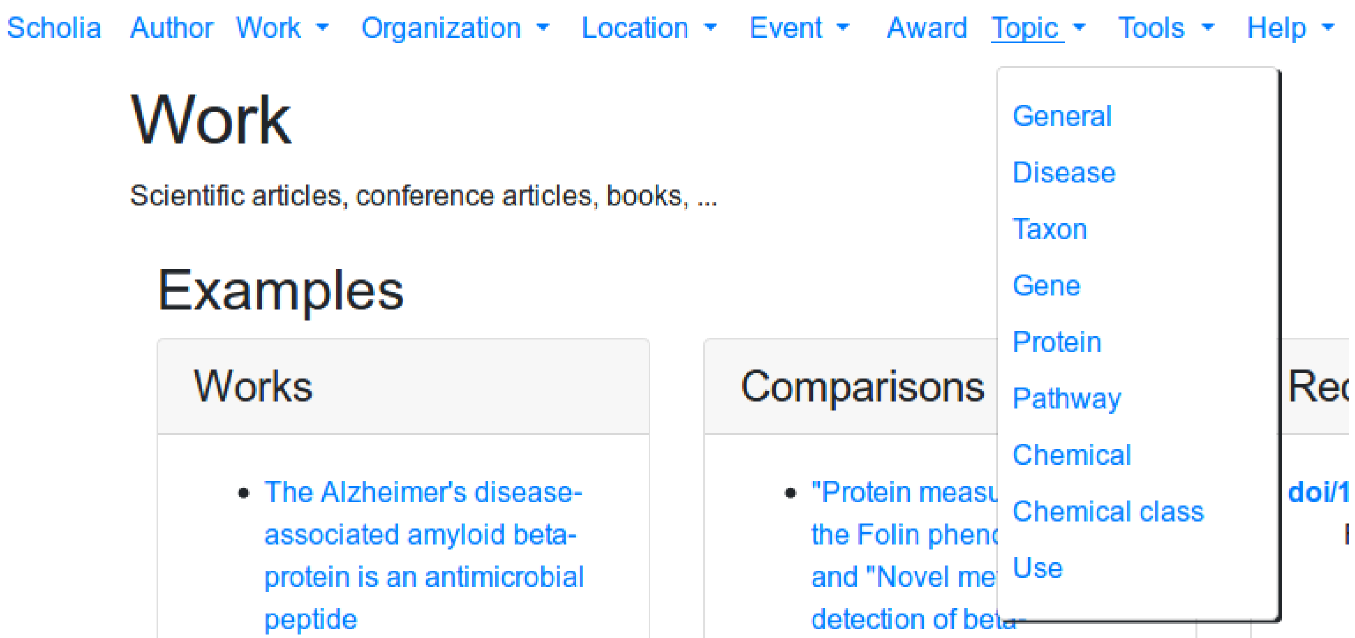
http://wikicite.org/statistics.html

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's aspects



Work

Scientific articles, conference articles, books, ...

Examples

Works

- The Alzheimer's disease-associated amyloid beta-protein is an antimicrobial peptide

Comparisons

- "Protein measu the Folin phen and "Novel me detection of beta-

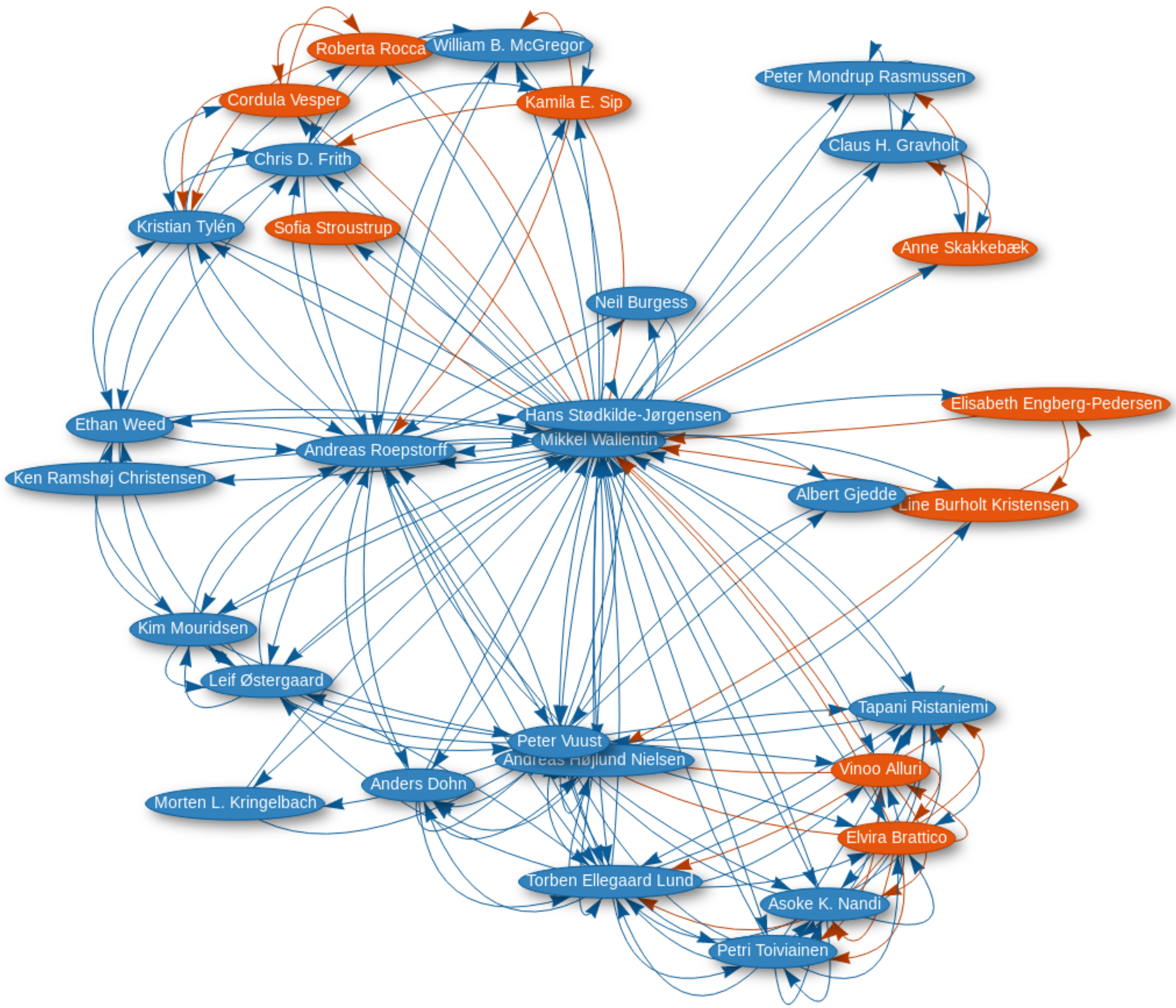
Rec

doi/1

Scholia shows Wikidata 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.

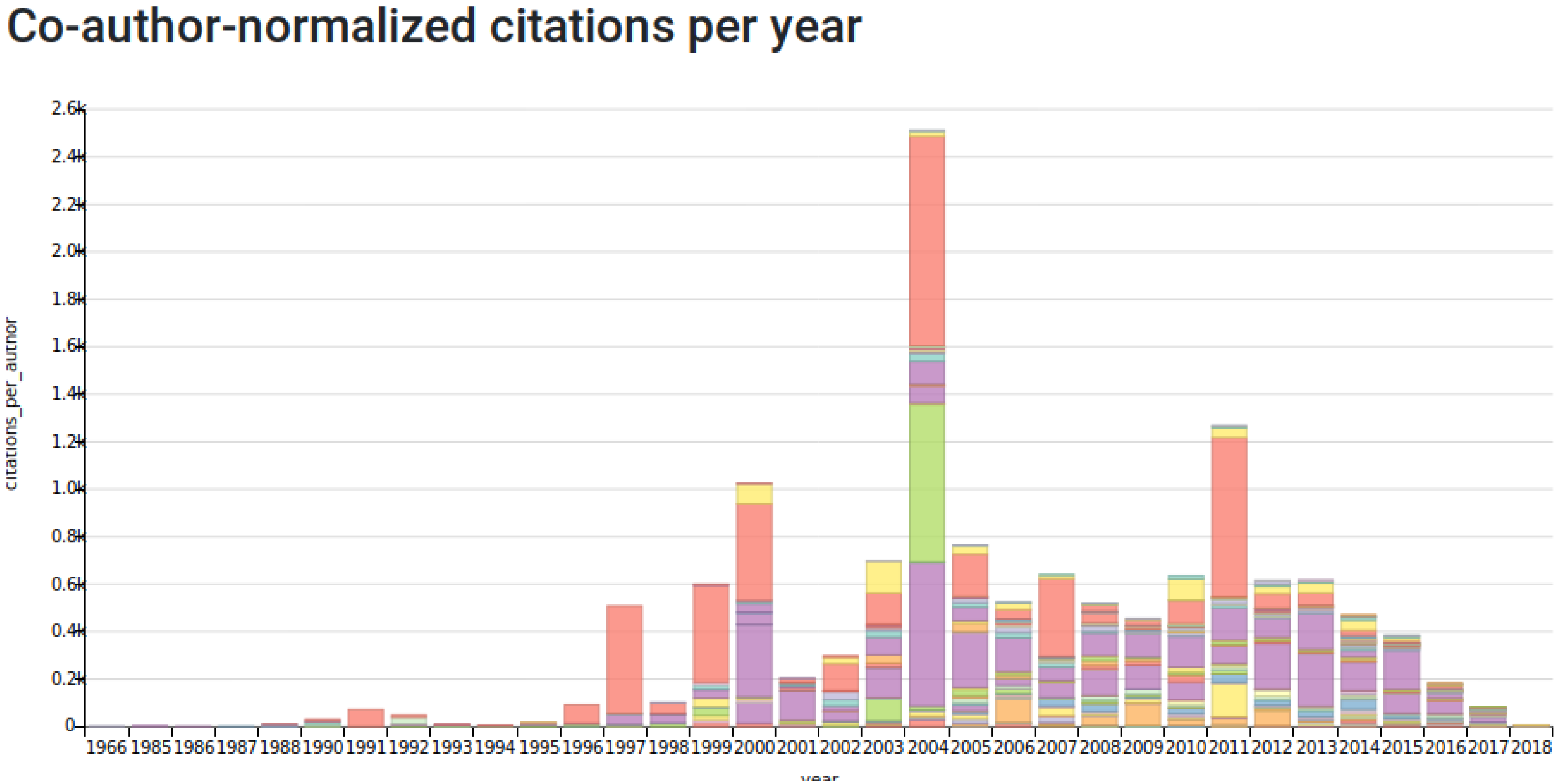
Author aspect: Co-author graph



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

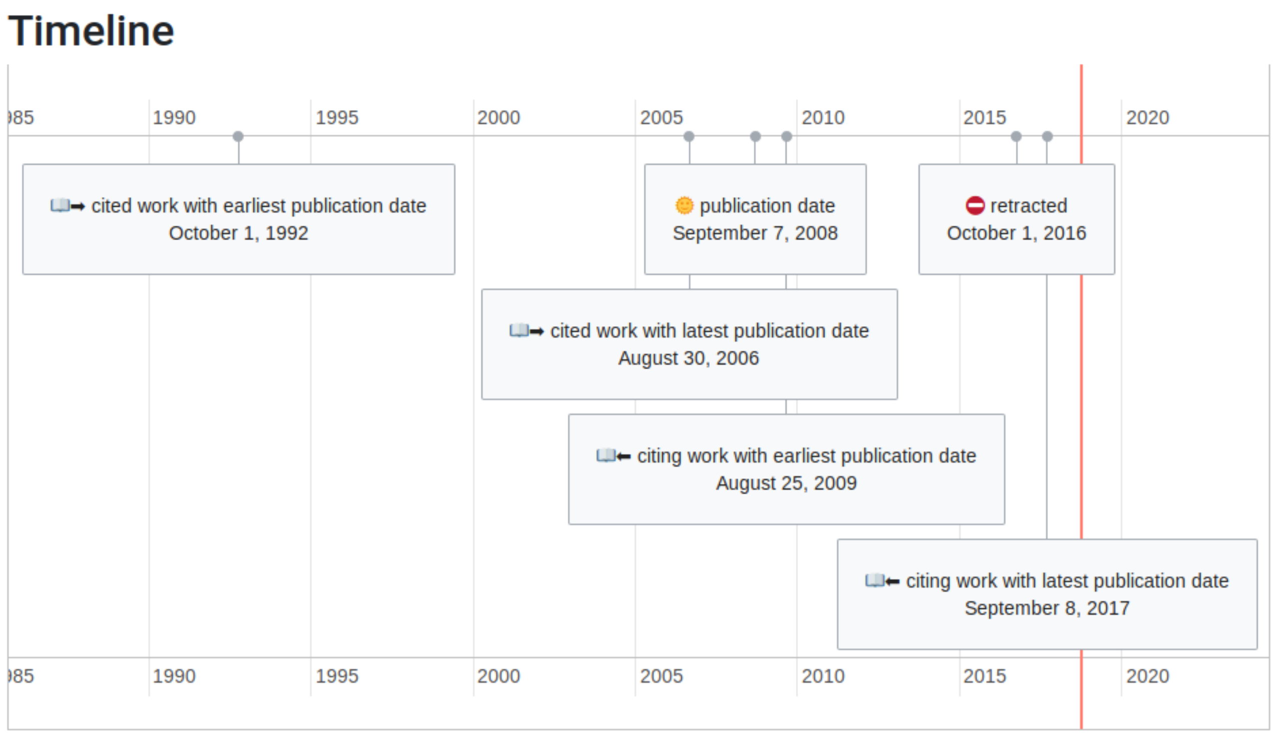
Colored according to gender.

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



Timeline

📖→ cited work with earliest publication date
October 1, 1992

😊 publication date
September 7, 2008

⛔ retracted
October 1, 2016

📖→ cited work with latest publication date
August 30, 2006

📖← citing work with earliest publication date
August 25, 2009

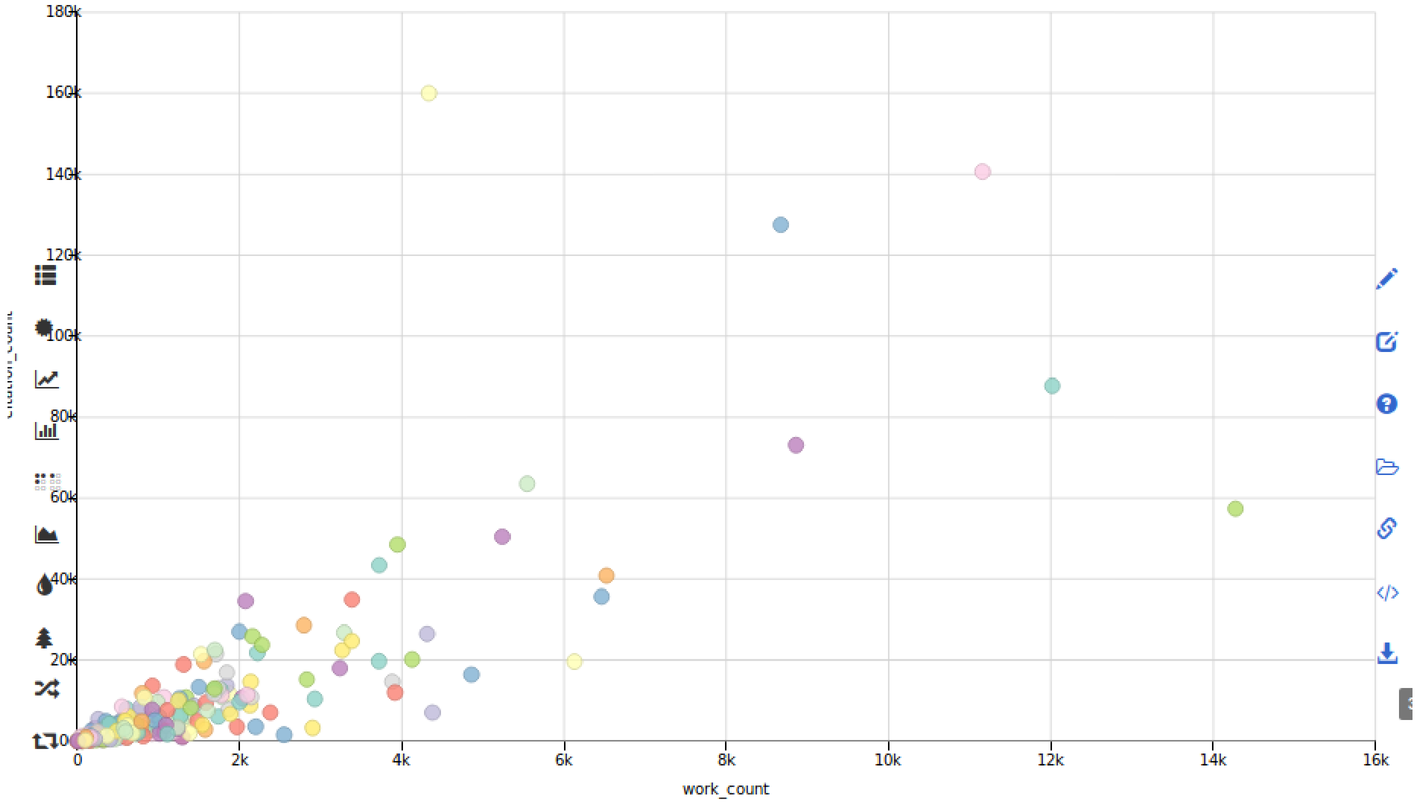
📖← citing work with latest publication date
September 8, 2017

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)

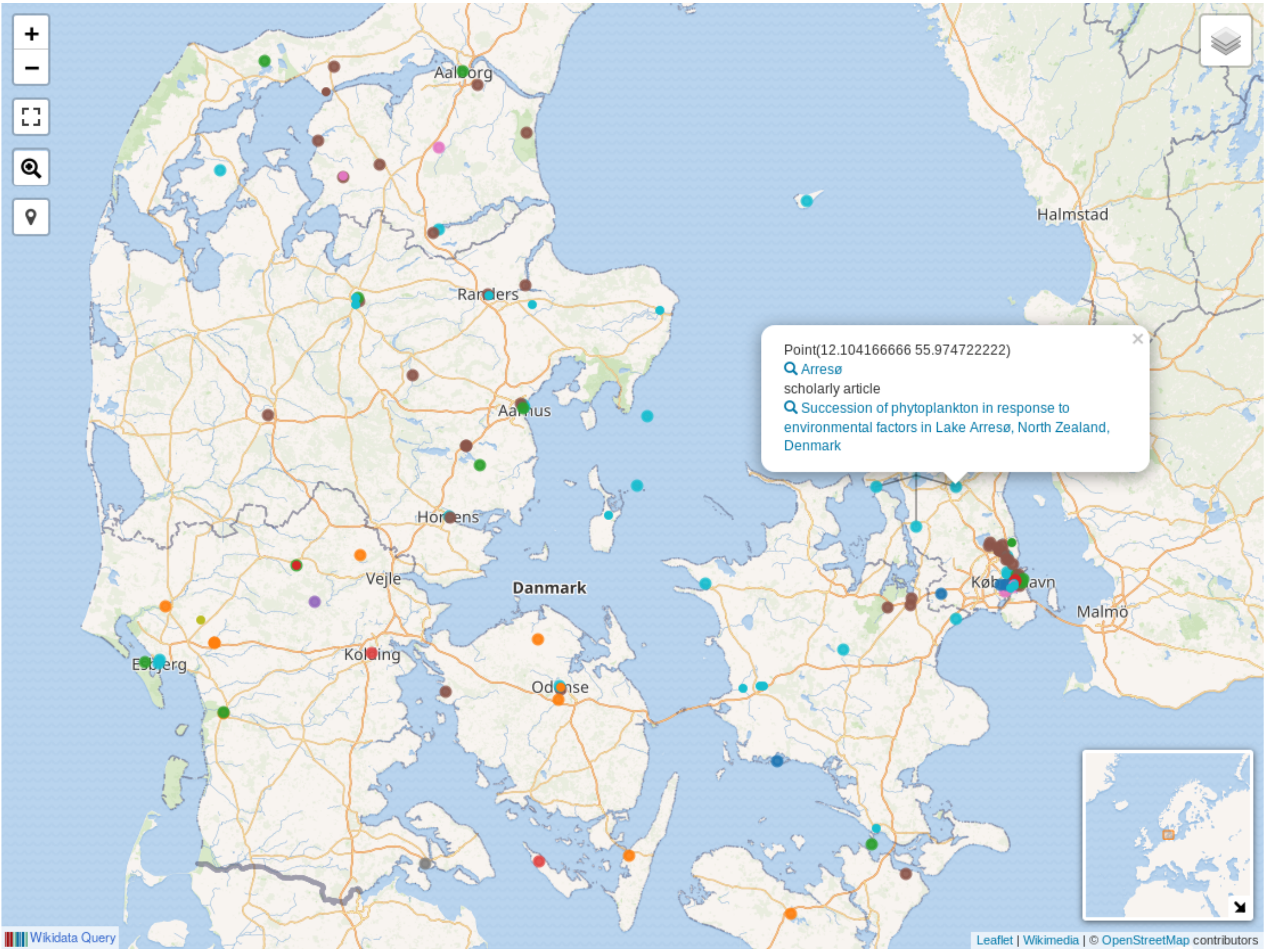
Publisher aspects



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



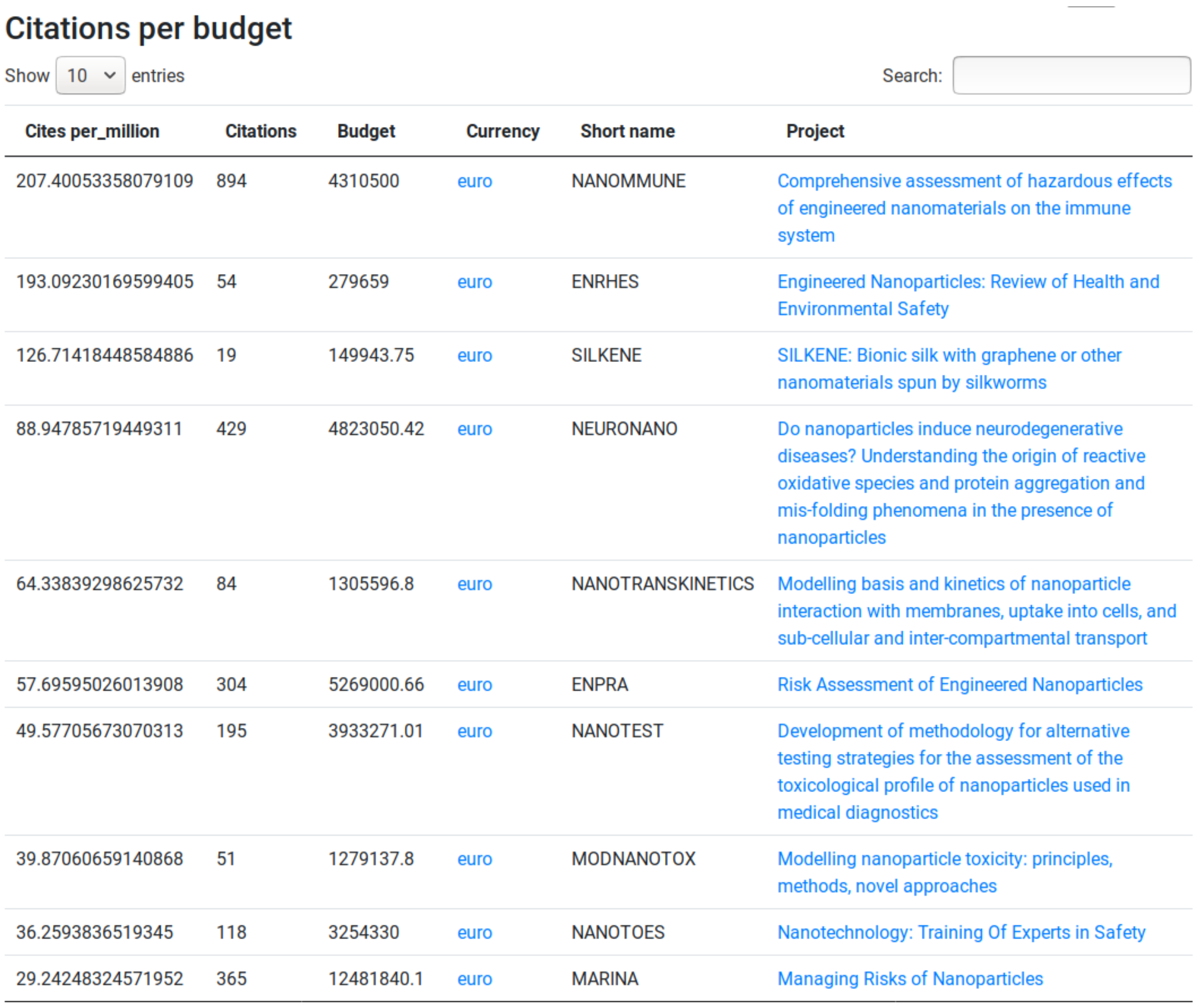
Point(12.104166666 55.974722222)
Arresø
scholarly article
Succession of phytoplankton in response to environmental factors in Lake Arresø, North Zealand, Denmark

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.

Project aspect: Research projects in Scholia



Citations per budget

Show 10 entries

Search:

Cites per_million	Citations	Budget	Currency	Short name	Project
207.40053358079109	894	4310500	euro	NANOMMUNE	Comprehensive assessment of hazardous effects of engineered nanomaterials on the immune system
193.09230169599405	54	279659	euro	ENRHES	Engineered Nanoparticles: Review of Health and Environmental Safety
126.71418448584886	19	149943.75	euro	SILKENE	SILKENE: Bionic silk with graphene or other nanomaterials spun by silkworms
88.94785719449311	429	4823050.42	euro	NEURONANO	Do nanoparticles induce neurodegenerative diseases? Understanding the origin of reactive oxidative species and protein aggregation and mis-folding phenomena in the presence of nanoparticles
64.33839298625732	84	1305596.8	euro	NANOTRANSKINETICS	Modelling basis and kinetics of nanoparticle interaction with membranes, uptake into cells, and sub-cellular and inter-compartmental transport
57.69595026013908	304	5269000.66	euro	ENPRA	Risk Assessment of Engineered Nanoparticles
49.57705673070313	195	3933271.01	euro	NANOTEST	Development of methodology for alternative testing strategies for the assessment of the toxicological profile of nanoparticles used in medical diagnostics
39.87060659140868	51	1279137.8	euro	MODNANOTOX	Modelling nanoparticle toxicity: principles, methods, novel approaches
36.2593836519345	118	3254330	euro	NANOTOES	Nanotechnology: Training Of Experts in Safety
29.24248324571952	365	12481840.1	euro	MARINA	Managing Risks of Nanoparticles

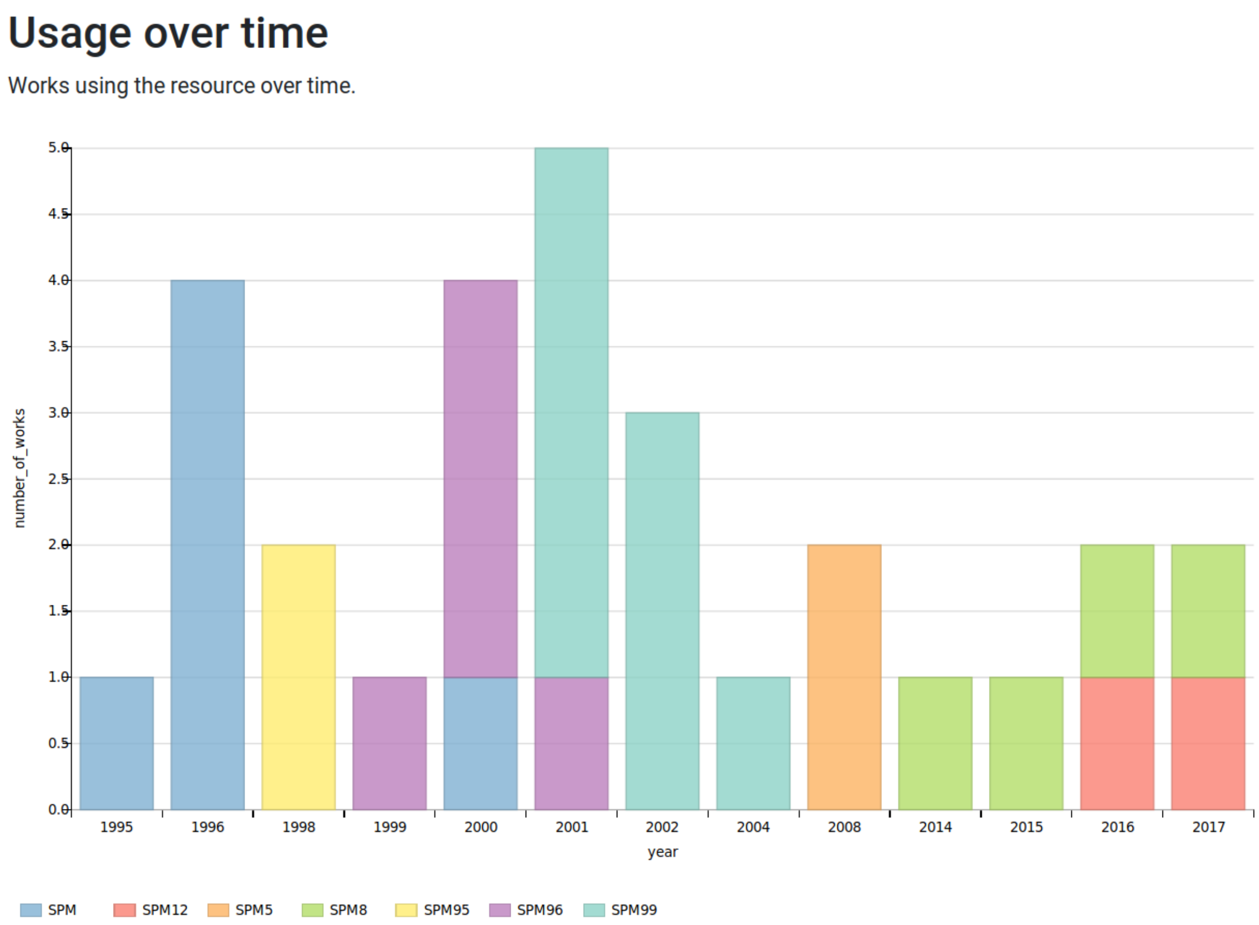
Research project aspect (Willighagen et al., 2018a).

If works are linked up to the project (by Wikidata'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



Usage over time

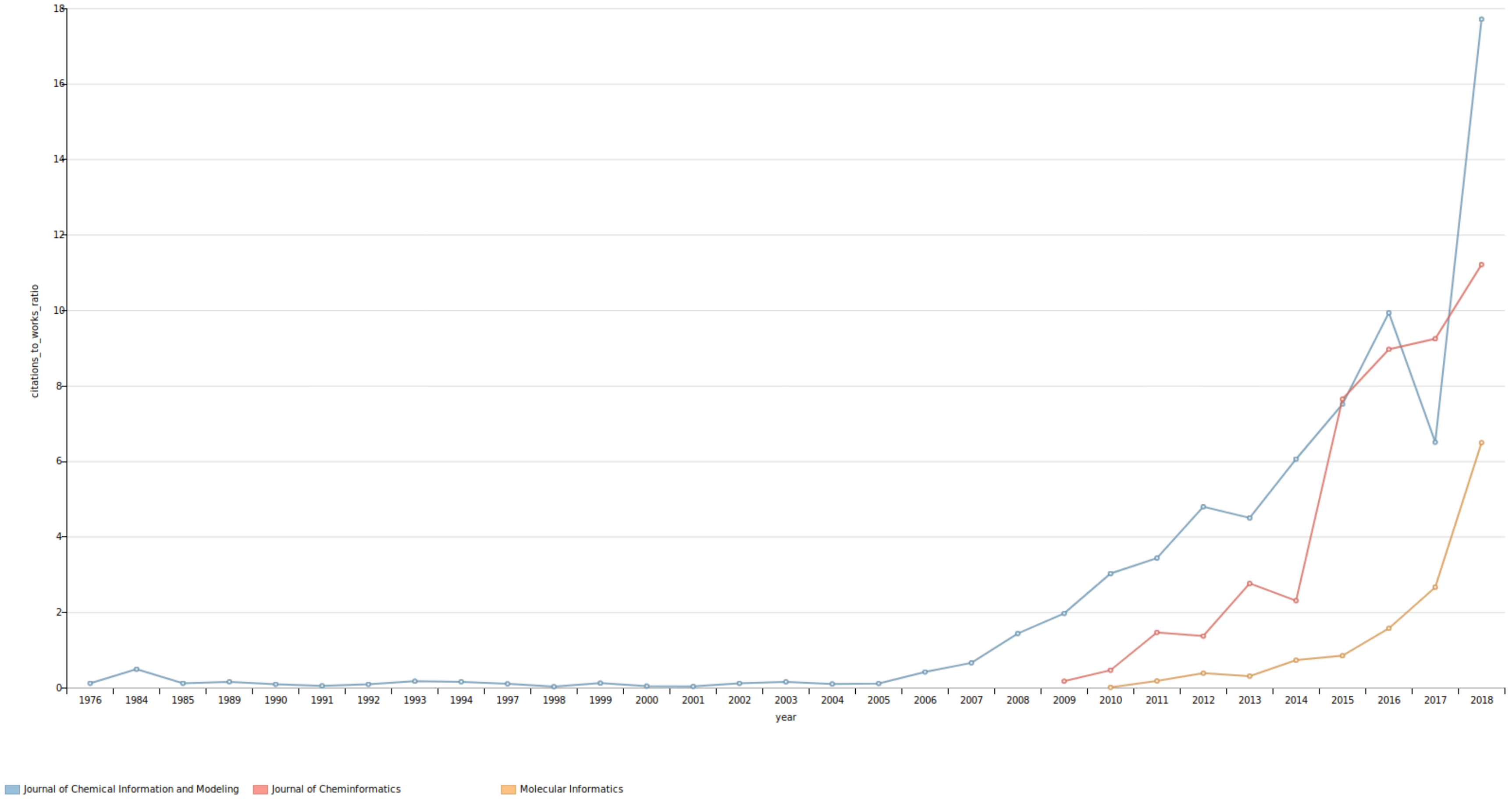
Works using the resource over time.

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.

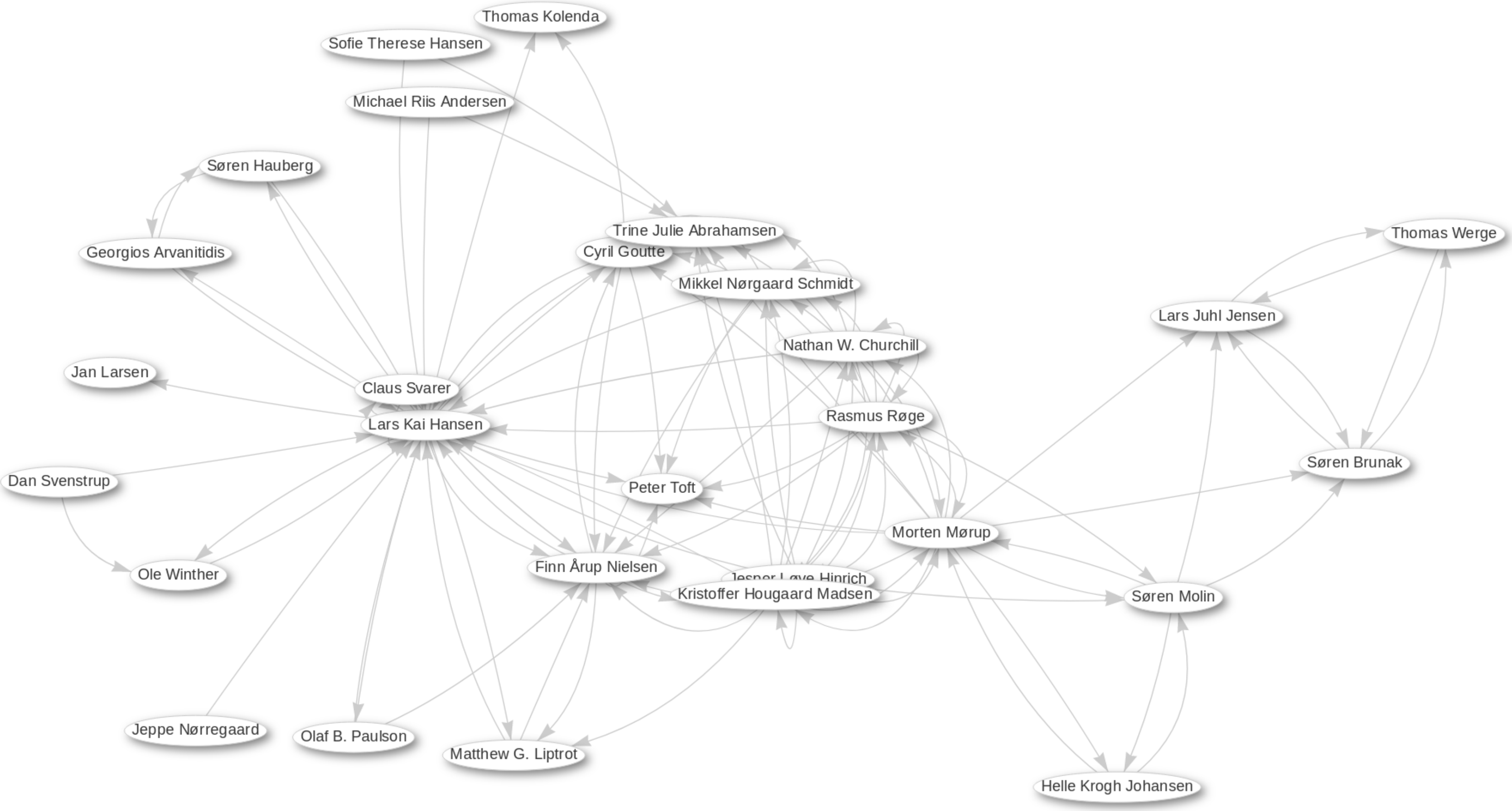
Such data is likely not available in directly machine readable format.

Comparison of multiple items



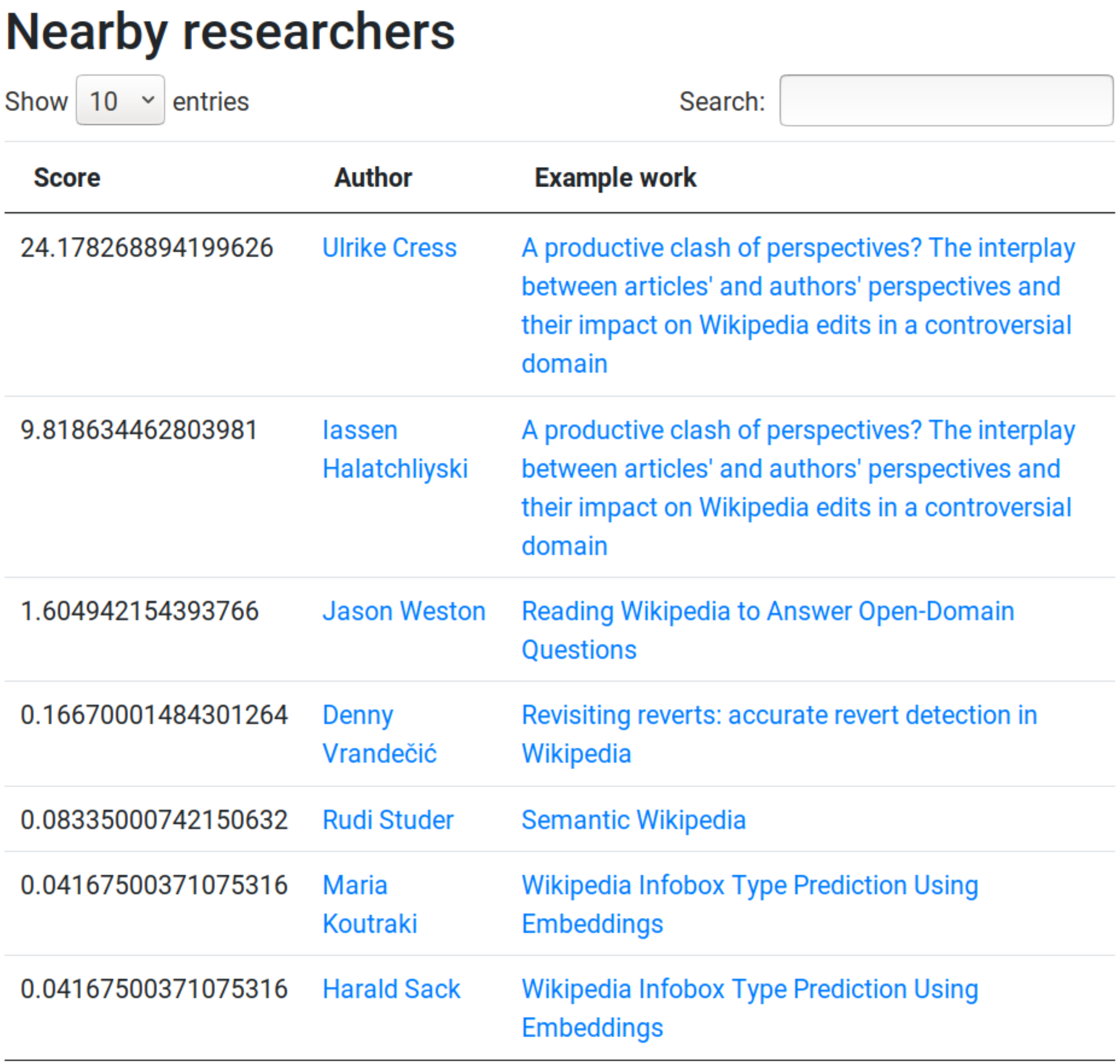
Multiple countries, e.g., some Southern and Eastern African countries or cheminformatics journals (here Willighagen's *citations to work ratio*).

Scholia's "subaspects"



Cocitation network for machine learning researchers in Denmark: /scholia/country/Q33/topic/Q2539.

Geodata and Scholia



Nearby researchers

Show 10 entries Search:

Score	Author	Example work
24.178268894199626	Ulrike Cress	A productive clash of perspectives? The interplay between articles' and authors' perspectives and their impact on Wikipedia edits in a controversial domain
9.818634462803981	Iassen Halatchliyski	A productive clash of perspectives? The interplay between articles' and authors' perspectives and their impact on Wikipedia edits in a controversial domain
1.604942154393766	Jason Weston	Reading Wikipedia to Answer Open-Domain Questions
0.16670001484301264	Denny Vrandečić	Revisiting reverts: accurate revert detection in Wikipedia
0.08335000742150632	Rudi Studer	Semantic Wikipedia
0.04167500371075316	Maria Koutraki	Wikipedia Infobox Type Prediction Using Embeddings
0.04167500371075316	Harald Sack	Wikipedia Infobox Type Prediction Using Embeddings

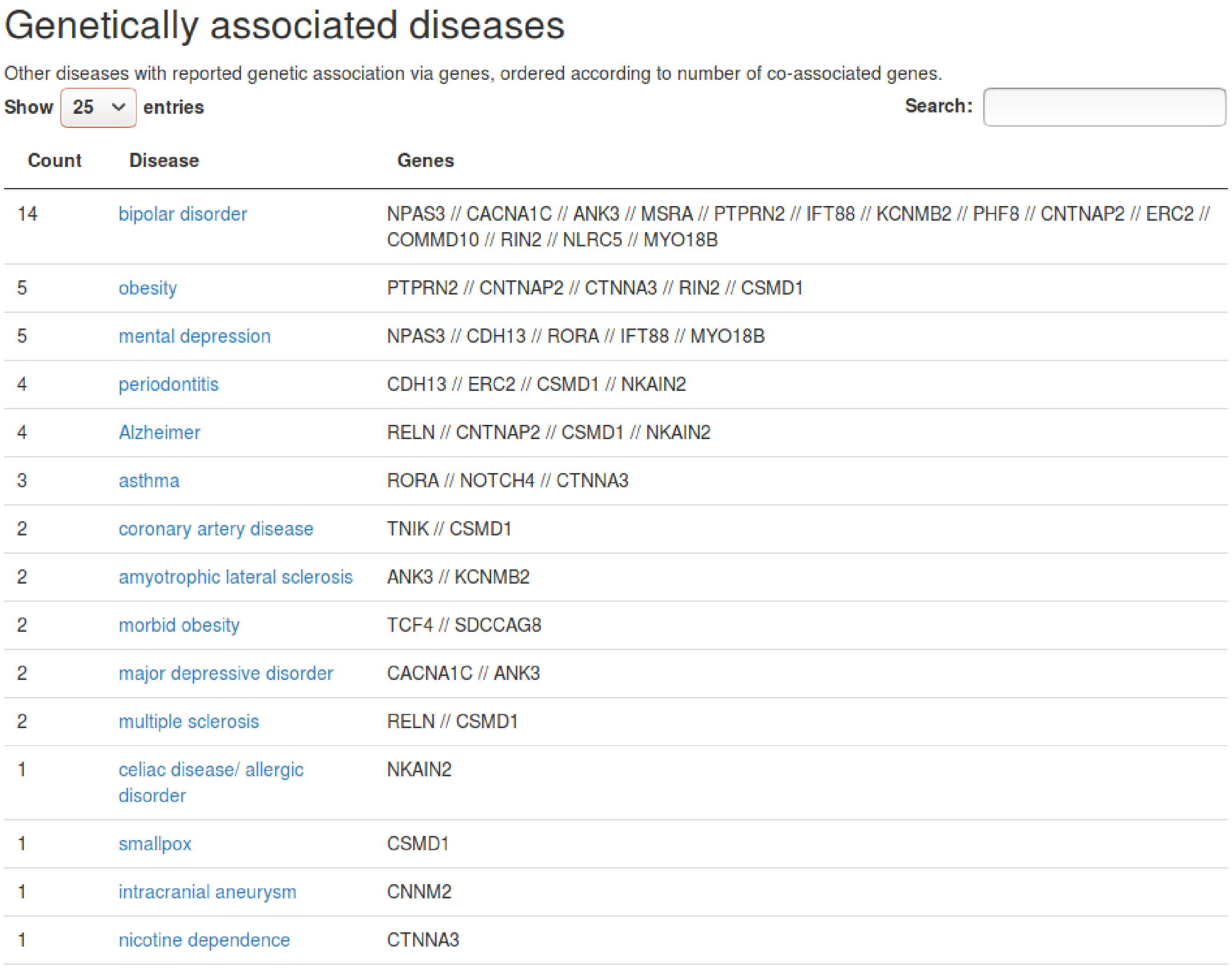
Wikipedia researchers near Tübingen: Weight information in Wikidata by the geographical distance and topic of authored works (Nielsen et al., 2018).

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

Nearby (in space and time) events also possible.

Finding related items

Related diseases with Wikidata Query Service



Genetically associated diseases

Other diseases with reported genetic association via genes, ordered according to number of co-associated genes.

Show 25 entries Search:

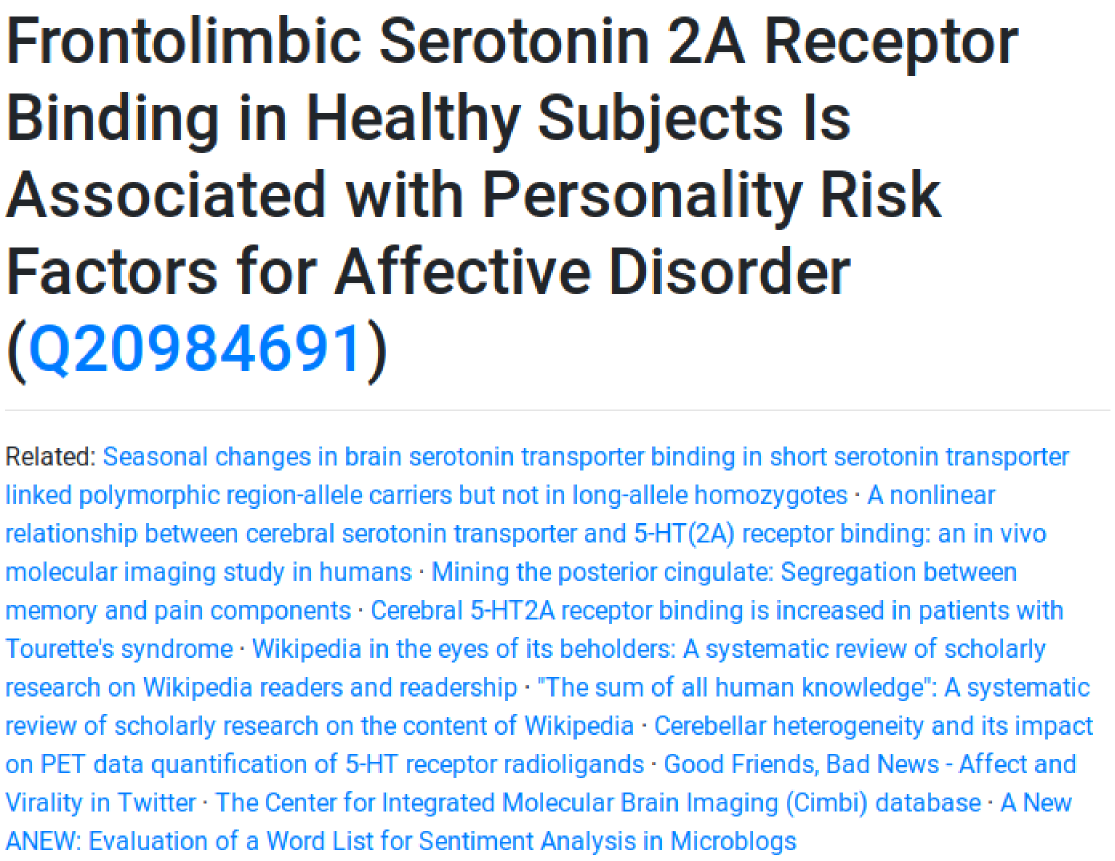
Count	Disease	Genes
14	bipolar disorder	NPAS3 // CACNA1C // ANK3 // MSRA // PTPRN2 // IFT88 // KCNMB2 // PHF8 // CNTNAP2 // ERC2 // COMMD10 // RIN2 // NLRC5 // MYO18B
5	obesity	PTPRN2 // CNTNAP2 // CTNNA3 // RIN2 // CSMD1
5	mental depression	NPAS3 // CDH13 // RORA // IFT88 // MYO18B
4	periodontitis	CDH13 // ERC2 // CSMD1 // NKAIN2
4	Alzheimer	RELN // CNTNAP2 // CSMD1 // NKAIN2
3	asthma	RORA // NOTCH4 // CTNNA3
2	coronary artery disease	TNIK // CSMD1
2	amyotrophic lateral sclerosis	ANK3 // KCNMB2
2	morbid obesity	TCF4 // SDCCAG8
2	major depressive disorder	CACNA1C // ANK3
2	multiple sclerosis	RELN // CSMD1
1	celiac disease/ allergic disorder	NKAIN2
1	smallpox	CSMD1
1	intracranial aneurysm	CNNM2
1	nicotine dependence	CTNNA3

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.

Wembedder



Frontolimbic Serotonin 2A Receptor Binding in Healthy Subjects Is Associated with Personality Risk Factors for Affective Disorder (Q20984691)

Related: Seasonal changes in brain serotonin transporter binding in short serotonin transporter linked polymorphic region-allele carriers but not in long-allele homozygotes · A nonlinear relationship between cerebral serotonin transporter and 5-HT(2A) receptor binding: an in vivo molecular imaging study in humans · Mining the posterior cingulate: Segregation between memory and pain components · Cerebral 5-HT2A receptor binding is increased in patients with Tourette's syndrome · Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership · "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia · Cerebellar heterogeneity and its impact on PET data quantification of 5-HT receptor radioligands · Good Friends, Bad News - Affect and Virality in Twitter · The Center for Integrated Molecular Brain Imaging (Cimbi) database · A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs

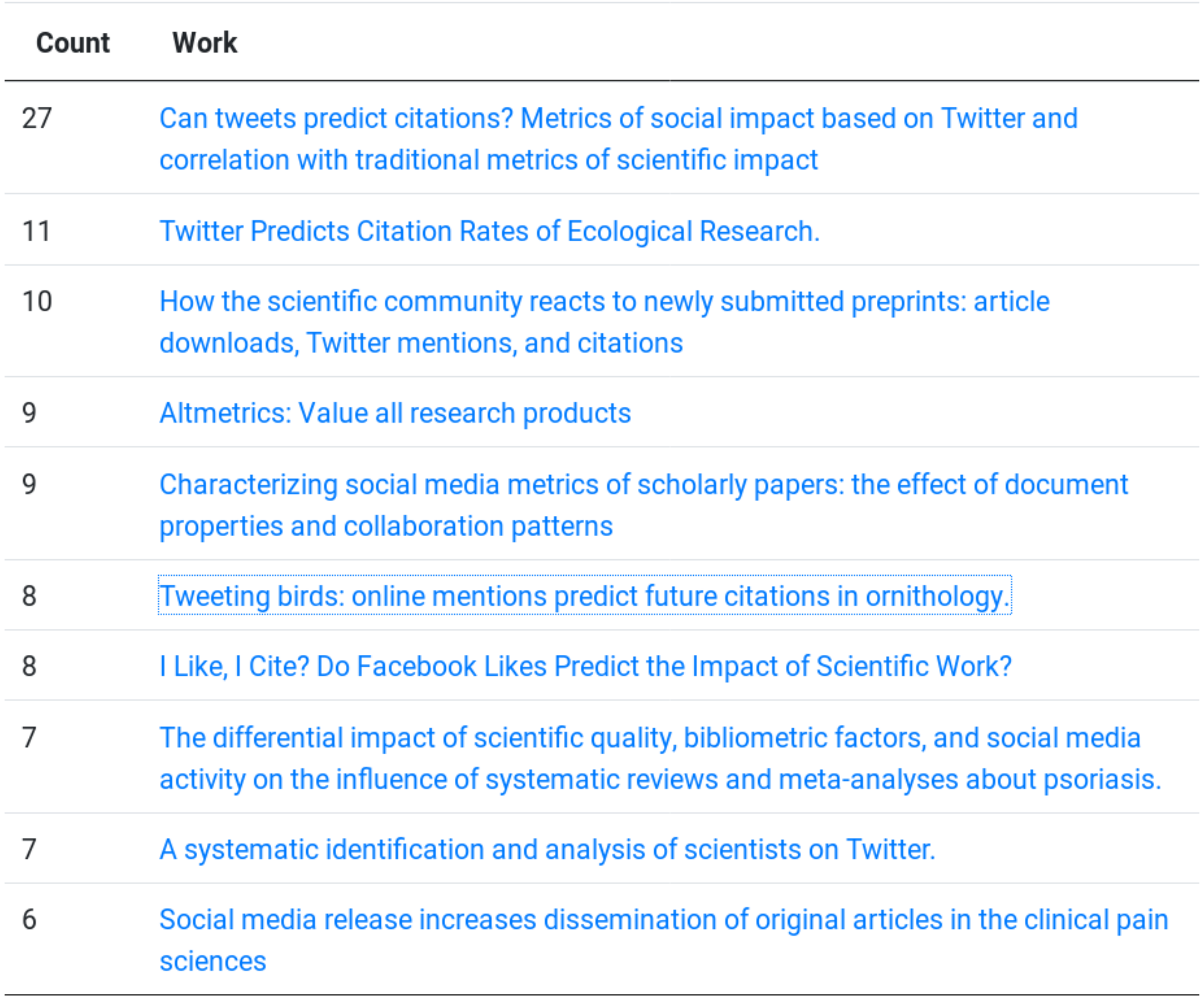
Finding related items based on word2vec-based knowledge graph embedding (Nielsen, 2017).

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



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

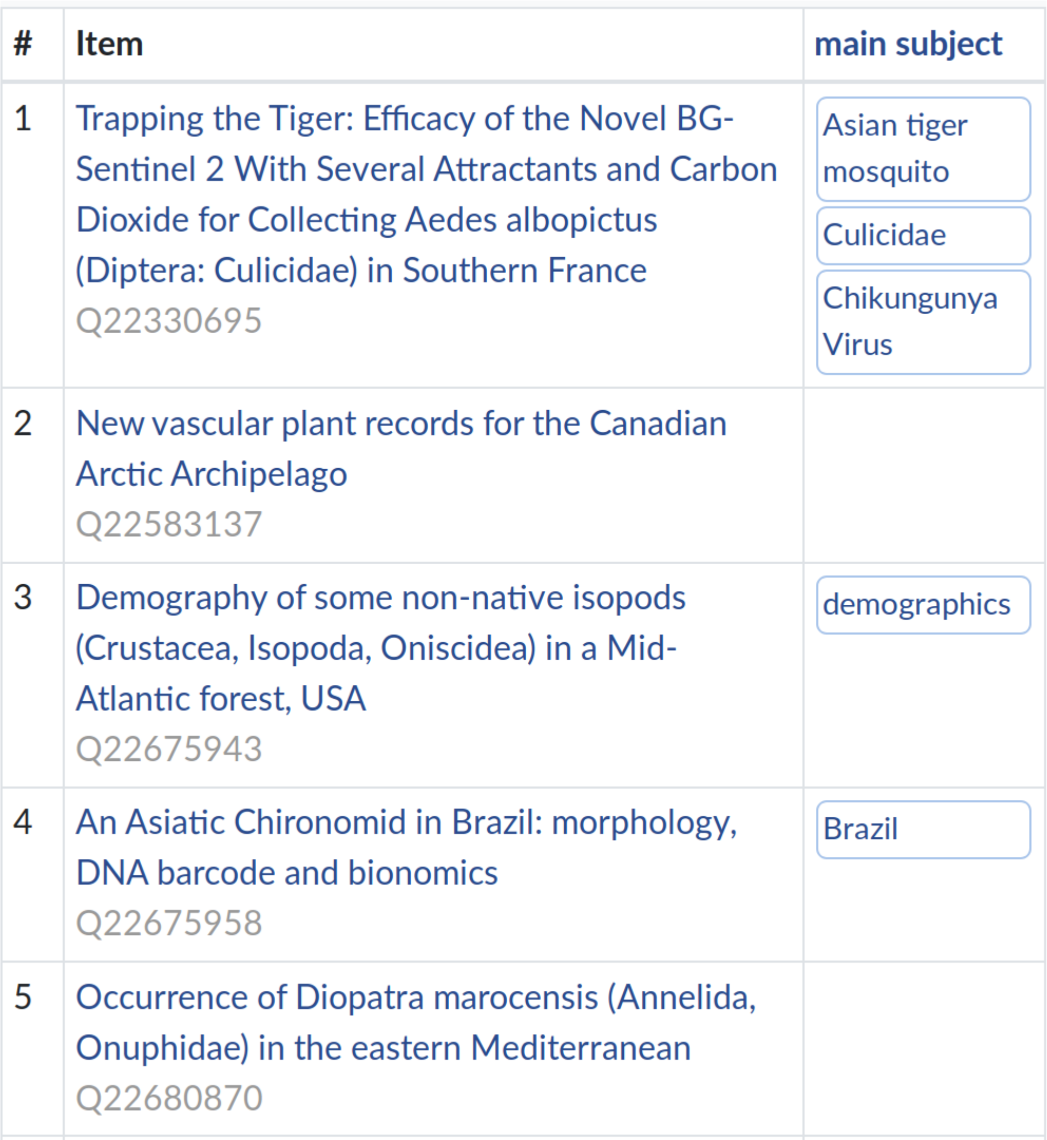
Example with *Do altmetrics work? Twitter and ten other social web services*.

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



#	Item	main subject
1	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	Asian tiger mosquito Culicidae Chikungunya Virus
2	New vascular plant records for the Canadian Arctic Archipelago Q22583137	
3	Demography of some non-native isopods (Crustacea, Isopoda, Oniscidea) in a Mid-Atlantic forest, USA Q22675943	demographics
4	An Asiatic Chironomid in Brazil: morphology, DNA barcode and bionomics Q22675958	Brazil
5	Occurrence of Diopatra marocensis (Annelida, Onuphidae) in the eastern Mediterranean Q22680870	

Manual input on the `https://www.wikidata.org` website.

Magnus Manske's tools: SourceMD 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.

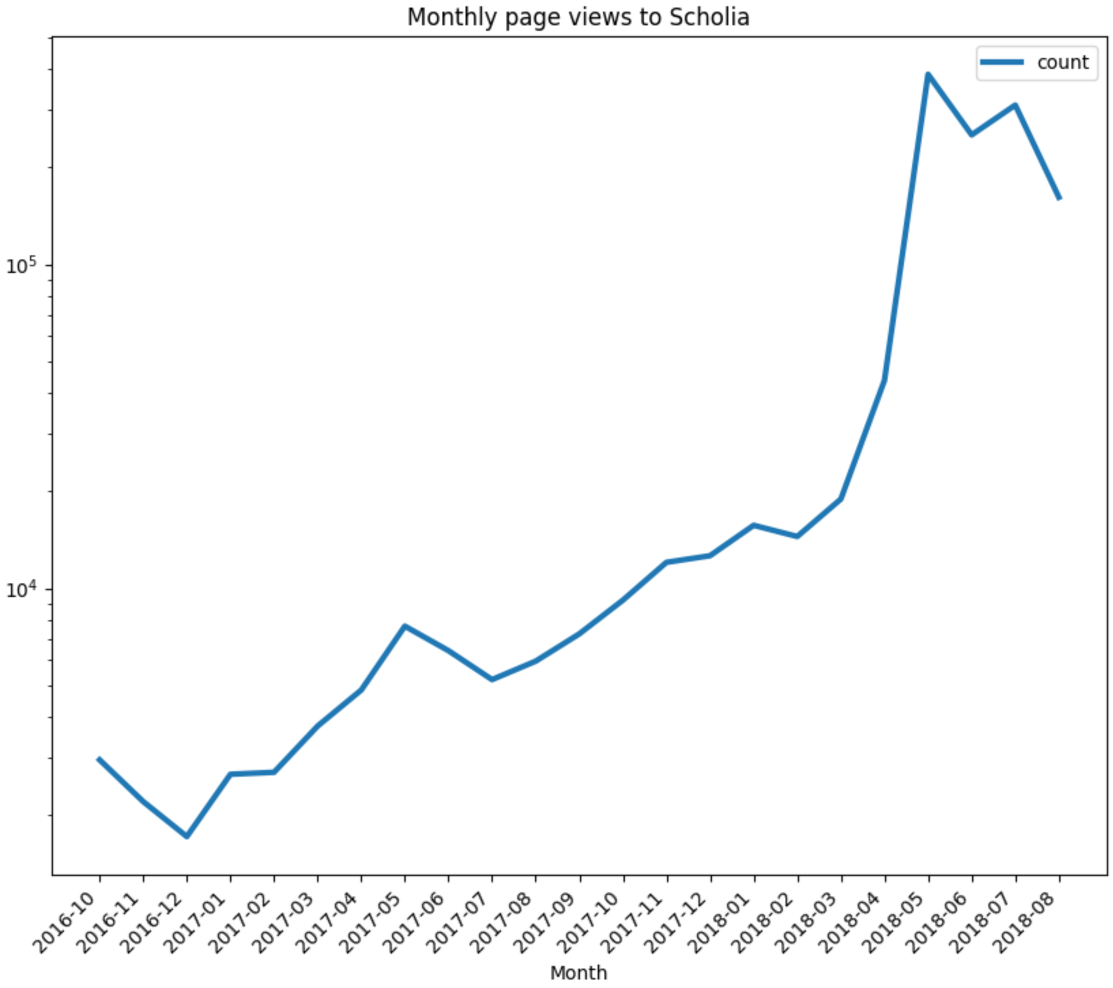
DOI bias: Documens with DOIs are easier to setup than documents without.

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.

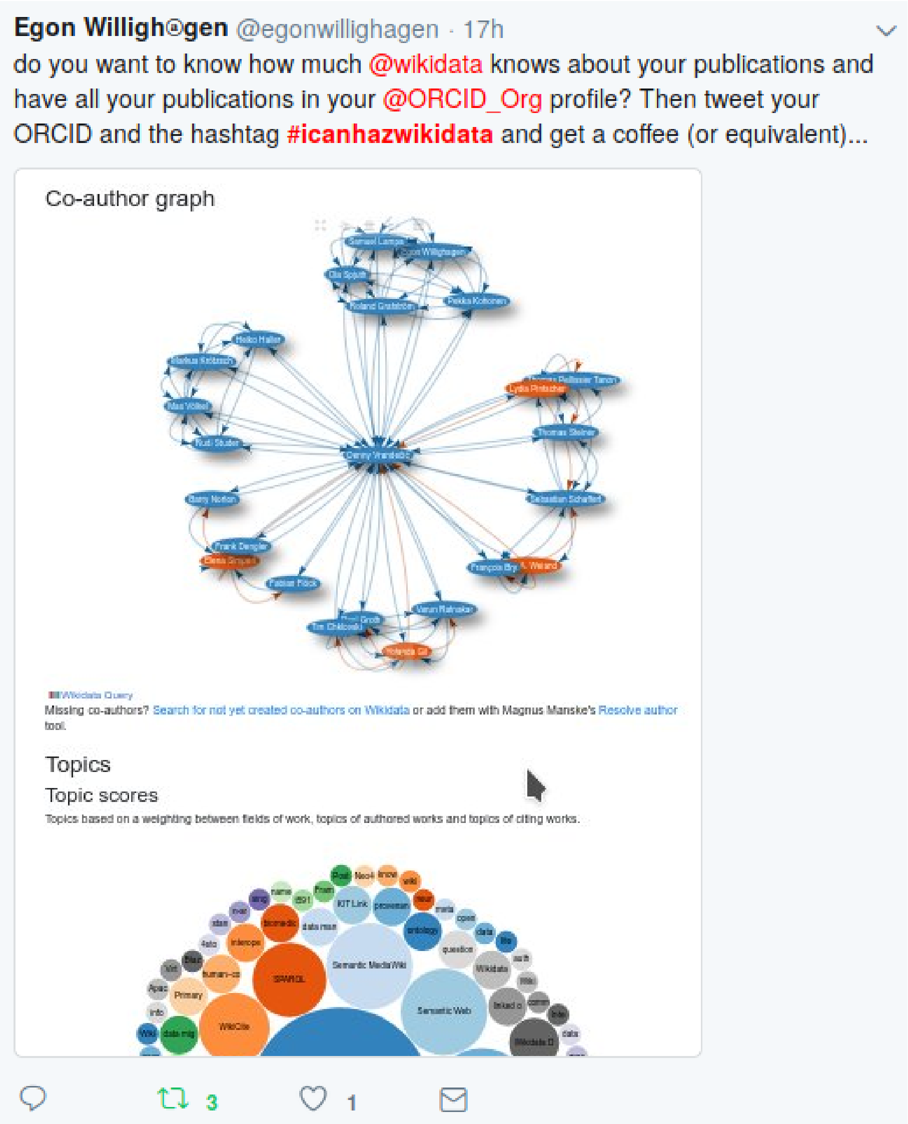
Scholia usage statistics



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 comming this way are bots or users are not known.

Scholia/Wikidata promotions



Egon Willigh@gen @egonwillighagen · 17h

do you want to know how much @wikidata knows about your publications and have all your publications in your @ORCID_Org profile? Then tweet your ORCID and the hashtag #icanhazwikidata and get a coffee (or equivalent)...

Co-author graph

Wikidata Query

Missing co-authors? Search for not yet created co-authors on Wikidata or add them with Magnus Manske's Resolve author tool.

Topics

Topic scores

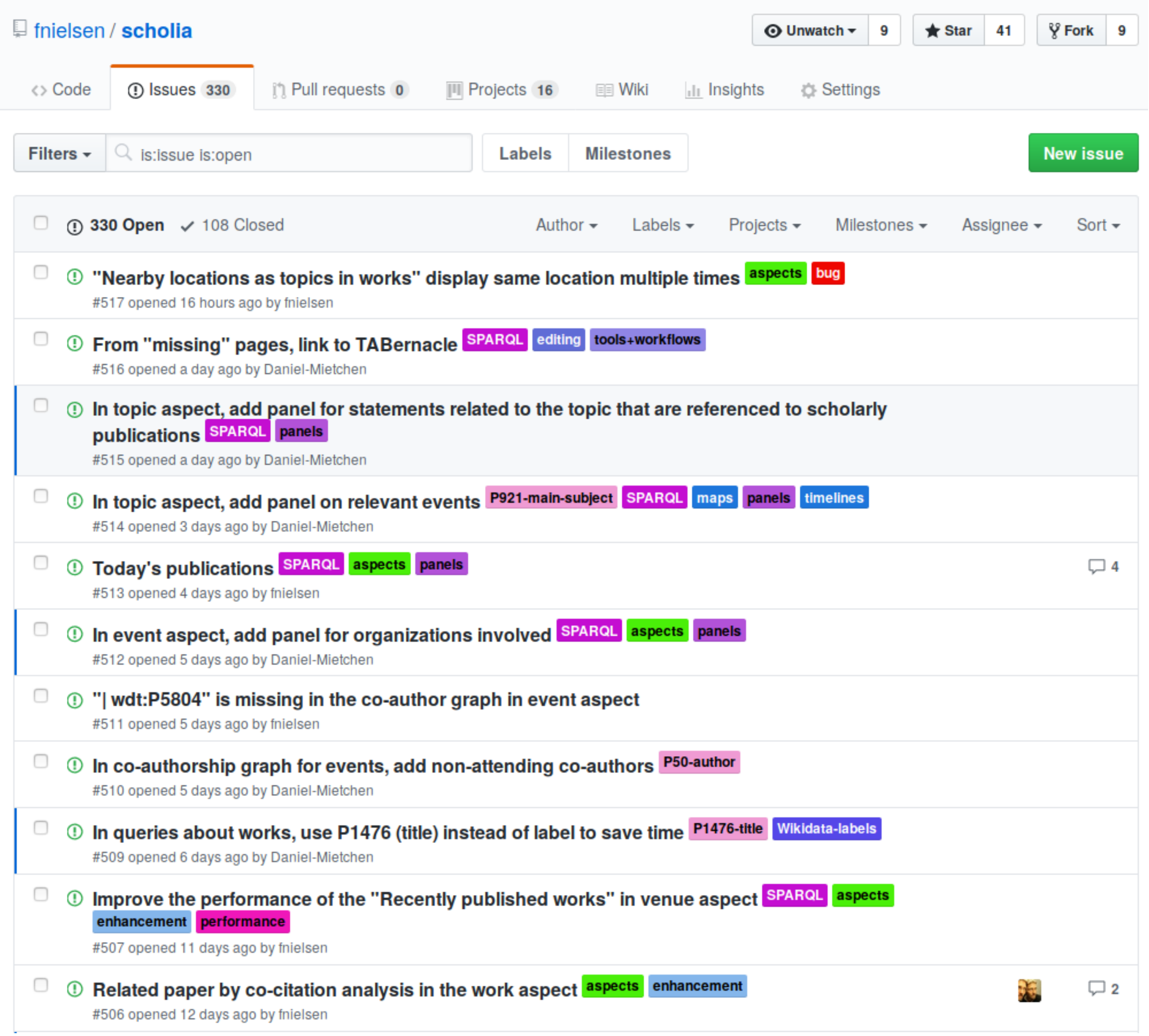
Topics based on a weighting between fields of work, topics of authored works and topics of citing works.

How do we spread the word of Scholia and Wikidata?

Here Egon Willighagen uses the hash tag *#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



fnielsen / scholia

Unwatch 9 | Star 41 | Fork 9

Code | Issues 330 | Pull requests 0 | Projects 16 | Wiki | Insights | Settings

Filters | is:issue is:open | Labels | Milestones | New issue

330 Open ✓ 108 Closed | Author | Labels | Projects | Milestones | Assignee | Sort

□ "Nearby locations as topics in works" display same location multiple times aspects bug
#517 opened 16 hours ago by fnielsen

□ From "missing" pages, link to TABernacle SPARQL editing tools+workflows
#516 opened a day ago by Daniel-Mietchen

□ In topic aspect, add panel for statements related to the topic that are referenced to scholarly publications SPARQL panels
#515 opened a day ago by Daniel-Mietchen

□ In topic aspect, add panel on relevant events P921-main-subject SPARQL maps panels timelines
#514 opened 3 days ago by Daniel-Mietchen

□ Today's publications SPARQL aspects panels 4
#513 opened 4 days ago by fnielsen

□ In event aspect, add panel for organizations involved SPARQL aspects panels
#512 opened 5 days ago by Daniel-Mietchen

□ "| wdt:P5804" is missing in the co-author graph in event aspect
#511 opened 5 days ago by fnielsen

□ In co-authorship graph for events, add non-attending co-authors P50-author
#510 opened 5 days ago by Daniel-Mietchen

□ In queries about works, use P1476 (title) instead of label to save time P1476-title Wikidata-labels
#509 opened 6 days ago by Daniel-Mietchen

□ Improve the performance of the "Recently published works" in venue aspect SPARQL aspects enhancement performance
#507 opened 11 days ago by fnielsen

□ Related paper by co-citation analysis in the work aspect aspects enhancement 2
#506 opened 12 days ago by fnielsen

Development takes place on GitHub under GPL at https://github.com/fnielsen/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 a 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

Ioannidis, J. P. A., Klavans, R., and Boyack, K. W. (2018). Thousands of scientists publish a paper every five days. *Nature*, 561:167–169. DOI: 10.1038/D41586-018-06185-8.

Nielsen, F. Å. (2017). Wembedder: Wikidata entity embedding web service. DOI: 10.5281/ZENODO.1009127.

Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2018). Geospatial data and Scholia. *Proceedings of the 3rd International Workshop on Geospatial Linked Data and the 2nd Workshop on Querying the Web of Data*. DOI: 10.5281/ZENODO.1202256.

Willighagen, E., Jahn, N., and Nielsen, F. Å. (2018a). The EU NanoSafety Cluster as Linked Data visualized with Scholia. DOI: 10.6084/M9.FIGSHARE.6727931.

Willighagen, E., Slenter, D., Mietchen, D., Evelo, C. T., and Nielsen, F. Å. (2018b). Wikidata and Scholia as a hub linking chemical knowledge. *11th International Conference on Chemical Structures. Program & Abstracts*, page 146. DOI: 10.6084/m9.figshare.6356027.v1.

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.