

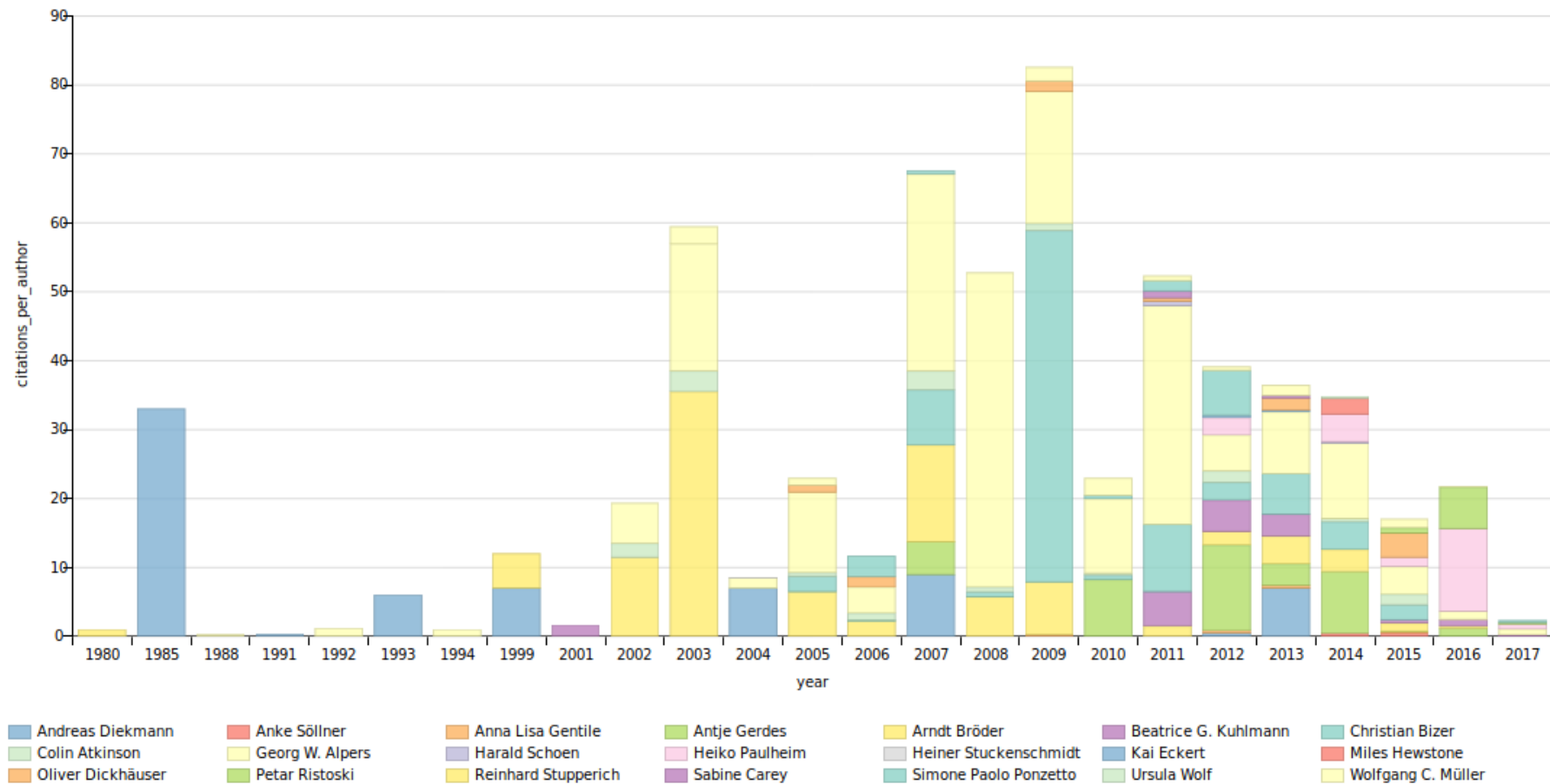
Scholia, WikiCite, Wikidata et al.

Finn Årup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

November 2018

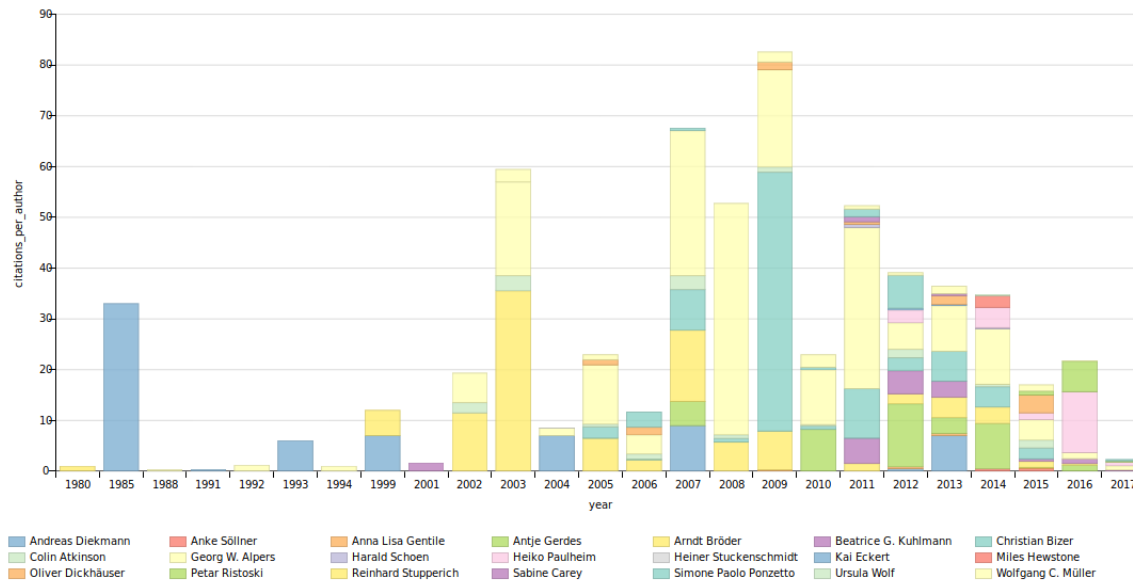
Co-author-normalized citations per year



<https://tools.wmflabs.org/scholia/organization/Q317070>

Scholia

Co-author-normalized citations per year

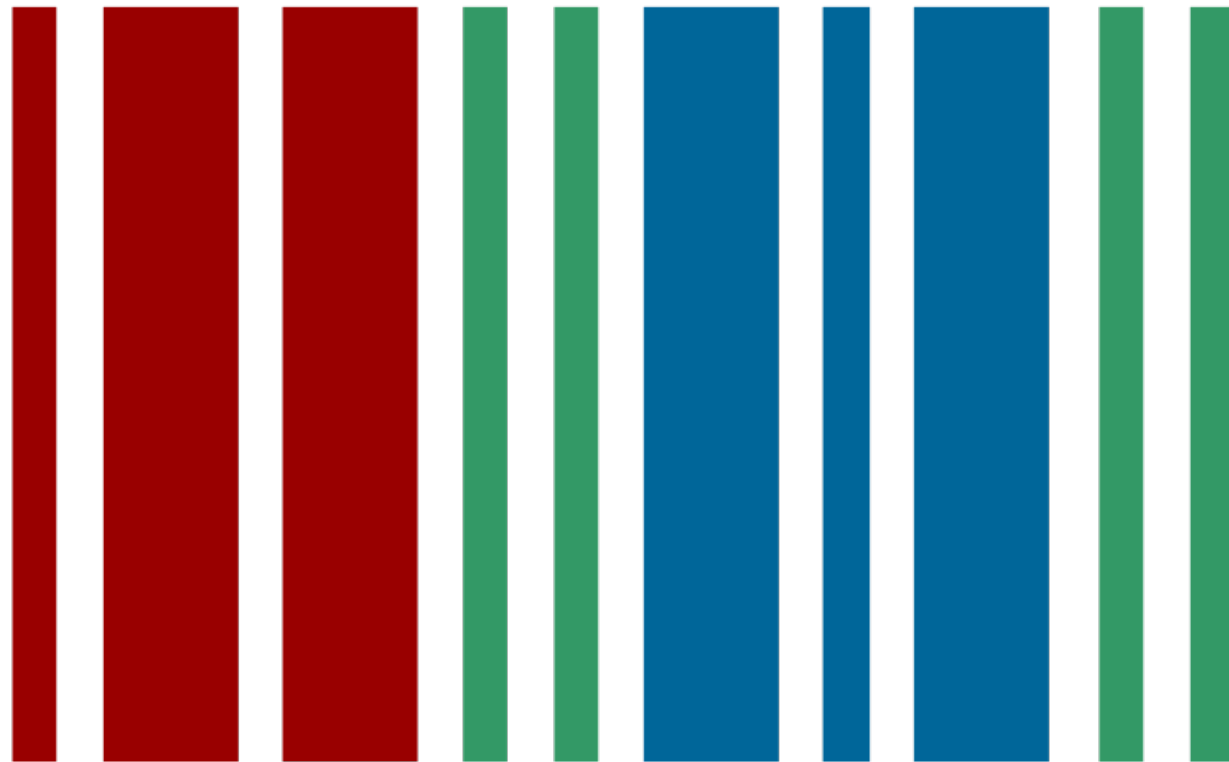


Scholia is a webservice from <https://tools.wmflabs.org/scholia/> and a Python package from <https://github.com/fnielsen/scholia>.

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

<https://www.wikidata.org>

Wikidata

Article [Discussion](#) [Read](#) [Edit](#) [View history](#) [WikiTrust](#)

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 WJ	[0 sources]
Vehicle registration	Klaus Wowereit <i>German politician</i>	[1 source]
	Klaus Wunderlich <i>German musician</i>	
Area	Klaus Waldeck <i>Austrian musician and former lawyer</i>	[2 sources]
Twin city	Klaus Wagner <i>German mathematician</i>	[3 sources]
	Klaus Wagner <i>Stalker of the British Royal Family</i>	

[\[new fact\]](#)

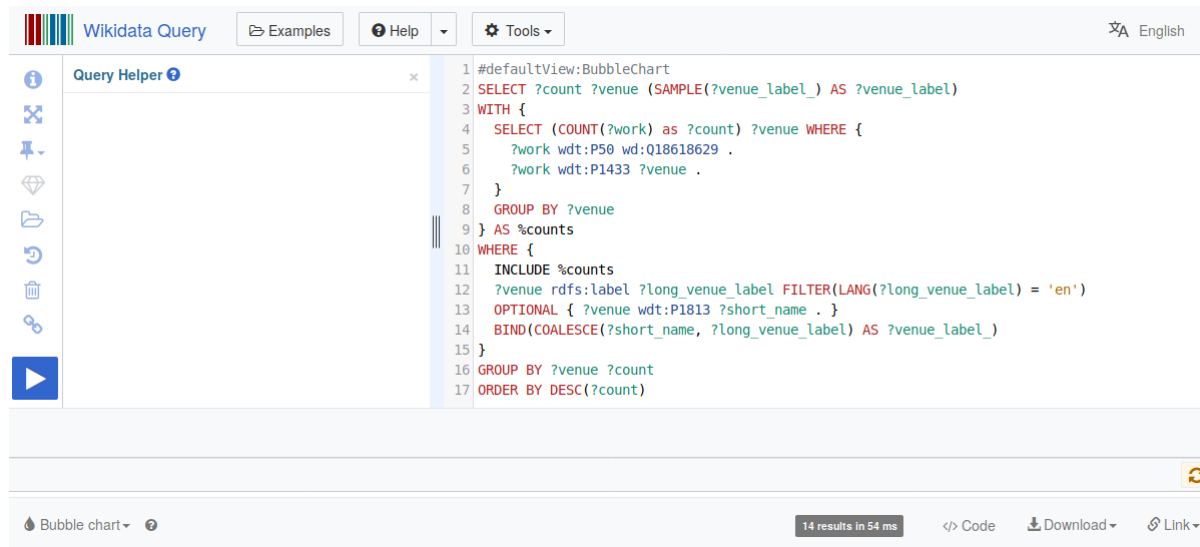
“Wikidata: Verifiable, Linked Open Knowledge That Anyone Can edit” (Dario Taraborelli)

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



The screenshot shows the Wikidata Query Service interface. At the top, there are navigation links for 'Examples', 'Help', and 'Tools', along with a language selector set to 'English'. A 'Query Helper' sidebar is visible on the left. The main area displays a SPARQL query:

```

1 #defaultView:BubbleChart
2 SELECT ?count ?venue (SAMPLE(?venue_label_) AS ?venue_label)
3 WITH {
4   SELECT (COUNT(?work) as ?count) ?venue WHERE {
5     ?work wdt:P50 wd:Q18618629 .
6     ?work wdt:P1433 ?venue .
7   }
8   GROUP BY ?venue
9 } AS %counts
10 WHERE {
11   INCLUDE %counts
12   ?venue rdfs:label ?long_venue_label FILTER(LANG(?long_venue_label) = 'en')
13   OPTIONAL { ?venue wdt:P1813 ?short_name . }
14   BIND(COALESCE(?short_name, ?long_venue_label) AS ?venue_label_)
15 }
16 GROUP BY ?venue ?count
17 ORDER BY DESC(?count)

```

Below the query, the results are displayed as a bubble chart. The chart shows seven bubbles of varying sizes and colors, representing different venues and their counts. The bubbles are labeled: 'Hypertext' (orange), 'SSS' (light blue), 'WWW'06' (dark purple), 'WWW'12' (medium blue), 'ISWC 2006' (light purple), 'ISWC 2014' (light green), and 'WWW'16' (medium green). The chart is titled 'Bubble chart' and shows '14 results in 54 ms'. There are also links for 'Code', 'Download', and 'Link'.

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.

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 <i>Aedes albopictus</i> (Diptera: Culicidae) in Southern France Q22330695	<div>Asian tiger mosquito</div> <div>Culicidae</div> <div>Chikungunya Virus</div>
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 <i>Diopatra marocensis</i> (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](#).

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.

WikiCite Statistics

Count	Description
6535113277	Total number of triples
163920228	Citations
85892308	Author name strings on items about works
17422401	Items with a PubMed ID
14571941	Items with a DOI
7045717	Items with a geolocation
4431216	Items with a PubMed Central ID
3791001	Links from items about works to items about their main subjects
3556963	Links from items about works to items about their authors
2521360	Items with a taxon name
284285	Items about authors with an ORCID profile that has public content

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

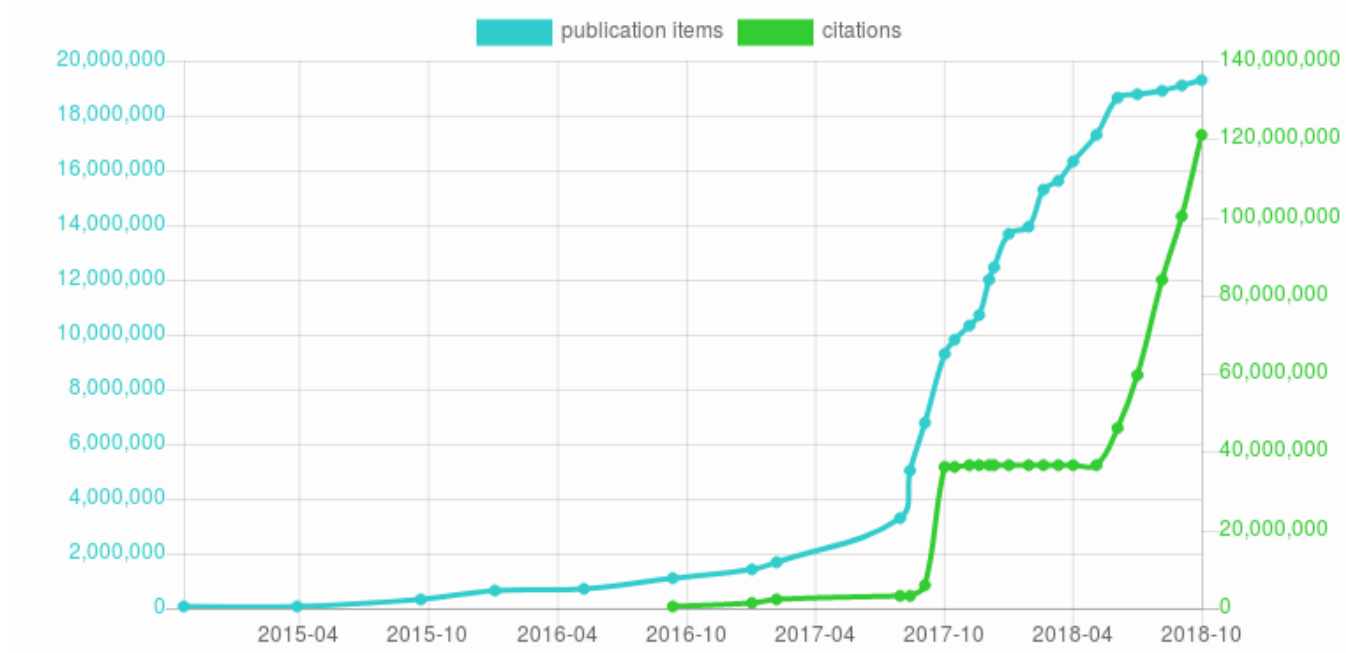
163 million citations.

17 million PubMed links.

14 million DOI links.

284 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

Scholia Author Work Organization Location Event Award Topic Tools Help

Work

Scientific articles, conference articles, books, ...

Examples

Works	Comparisons
<ul style="list-style-type: none"> The Alzheimer's disease-associated amyloid beta-protein is an antimicrobial peptide 	<ul style="list-style-type: none"> "Protein measurement using the Folin phenol reagent" and "Novel method for the detection of beta-amyloid protein"

General
Disease
Taxon
Gene
Protein
Pathway
Chemical
Chemical class
Use

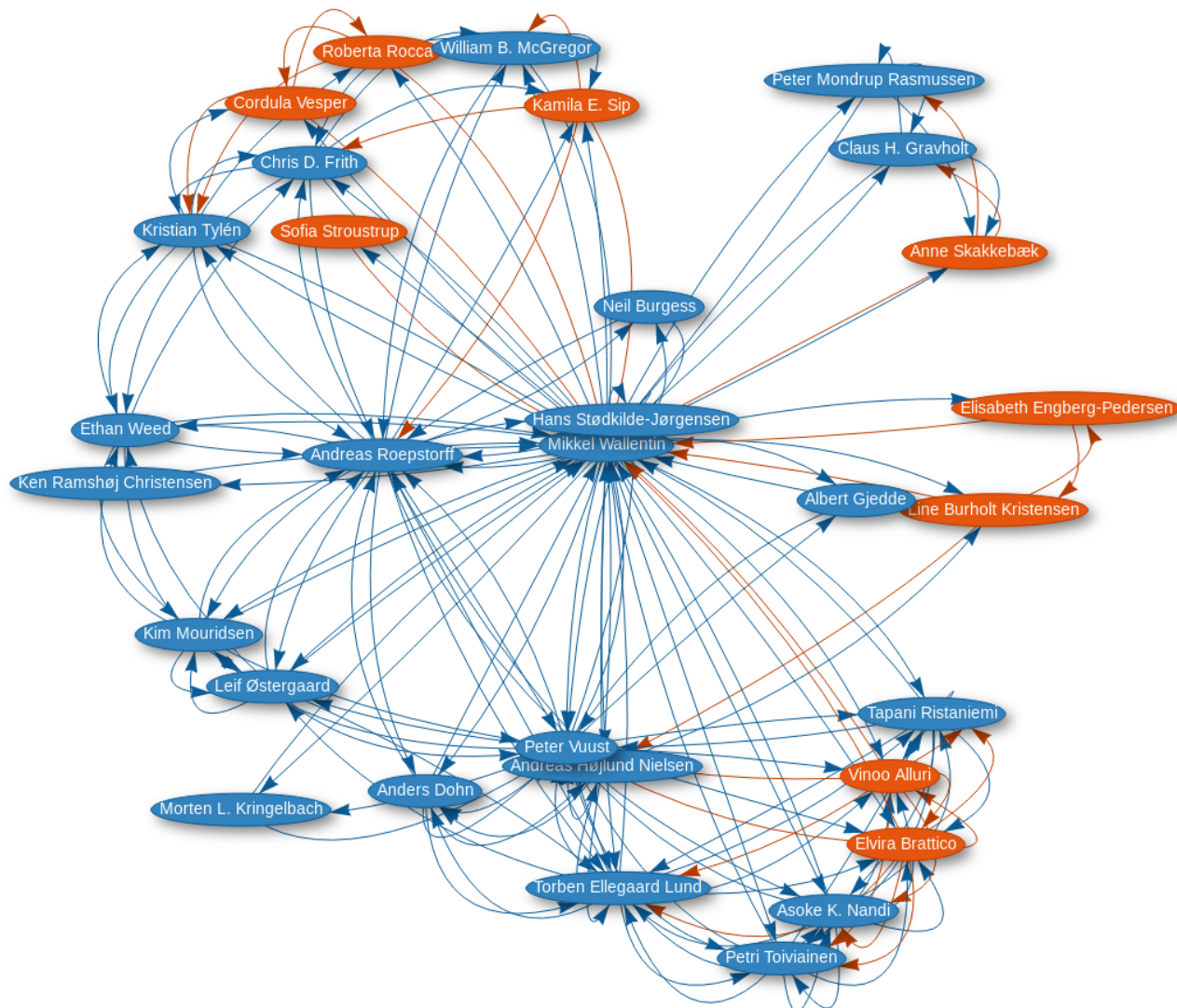
Scholia shows Wikidata data in *aspects*, author, work, organization (e.g., university, research group), venue (journal or conference), series, publisher, sponsor, location, event, award, topic, chemical, disease, etc.

For instance, the *Technical University of Denmark* may be viewed as a publisher, topic, organization, sponsor and location.

Author aspect: Co-author graph

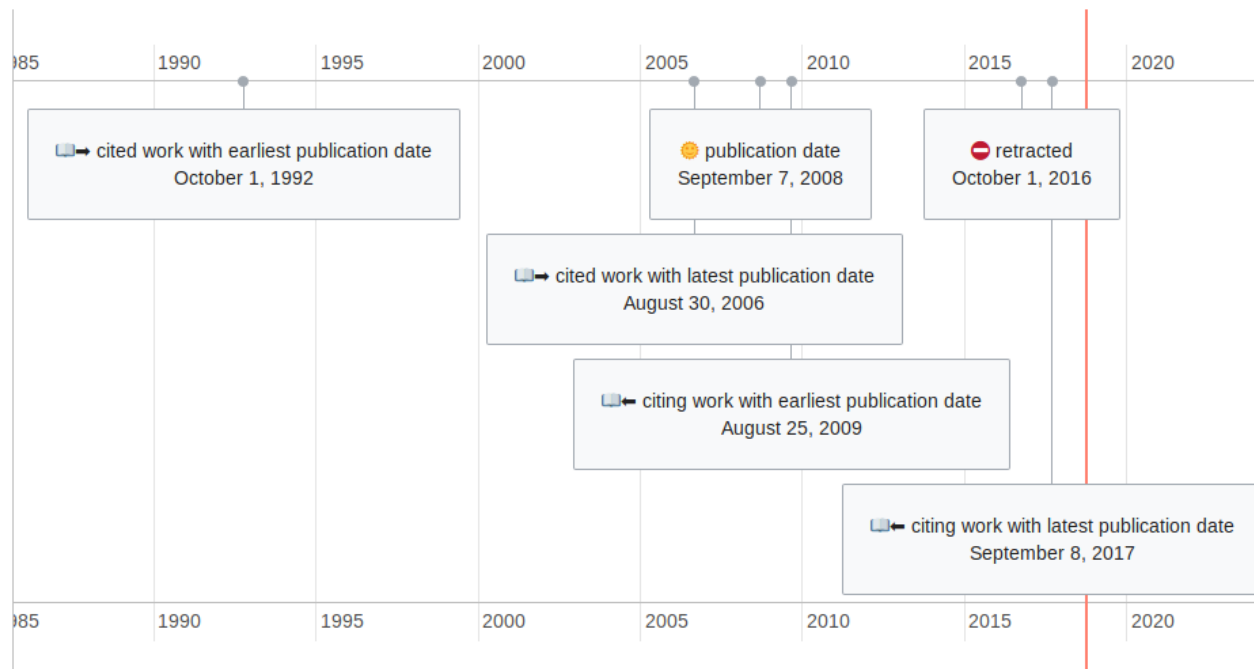
The egocentric co-author graph in Scholia's author aspect for the researcher **Mikkel Wallentin**, Aarhus University.

Colored according to gender.



Work aspect: Retractions

Timeline

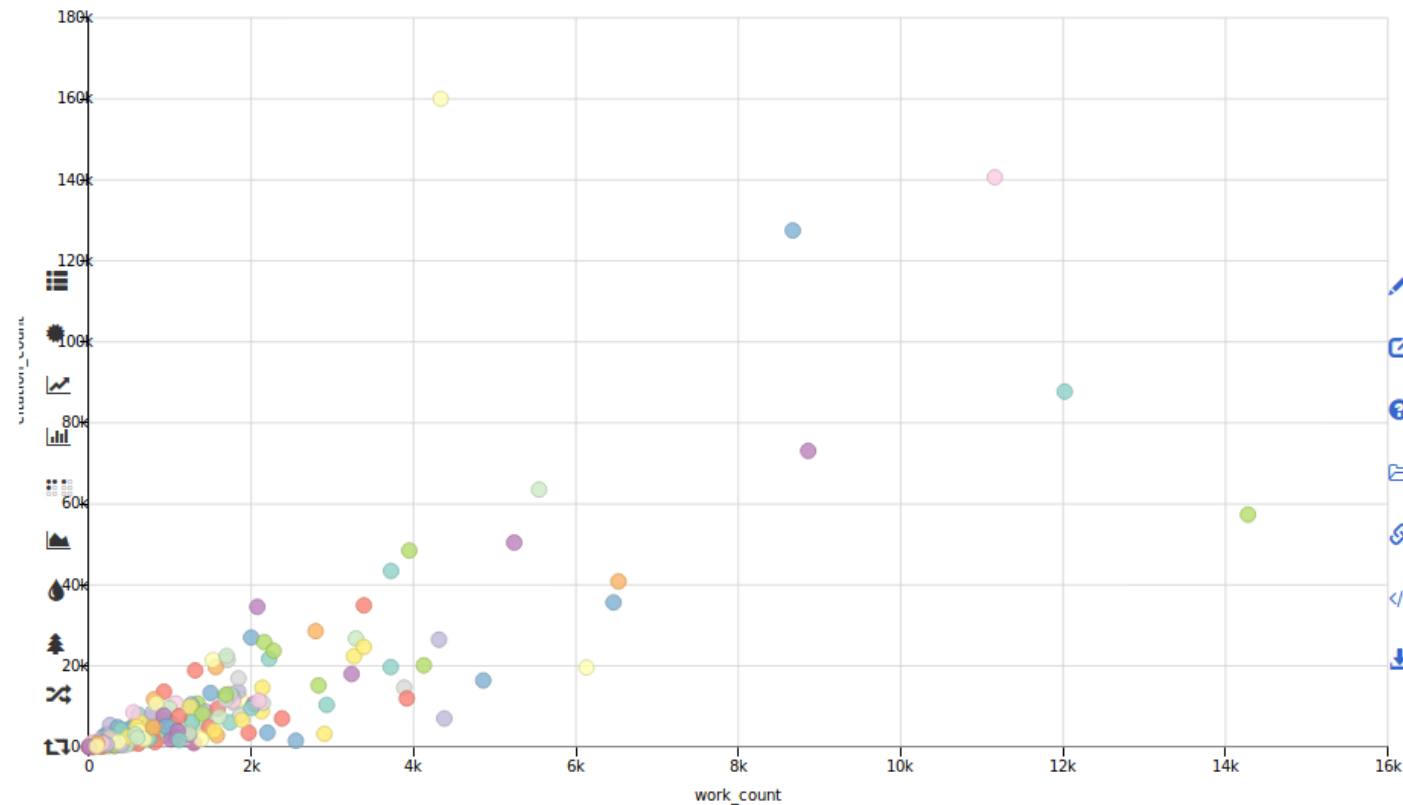


Wikidata can specify retracted papers, retraction notices and their connection.

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)

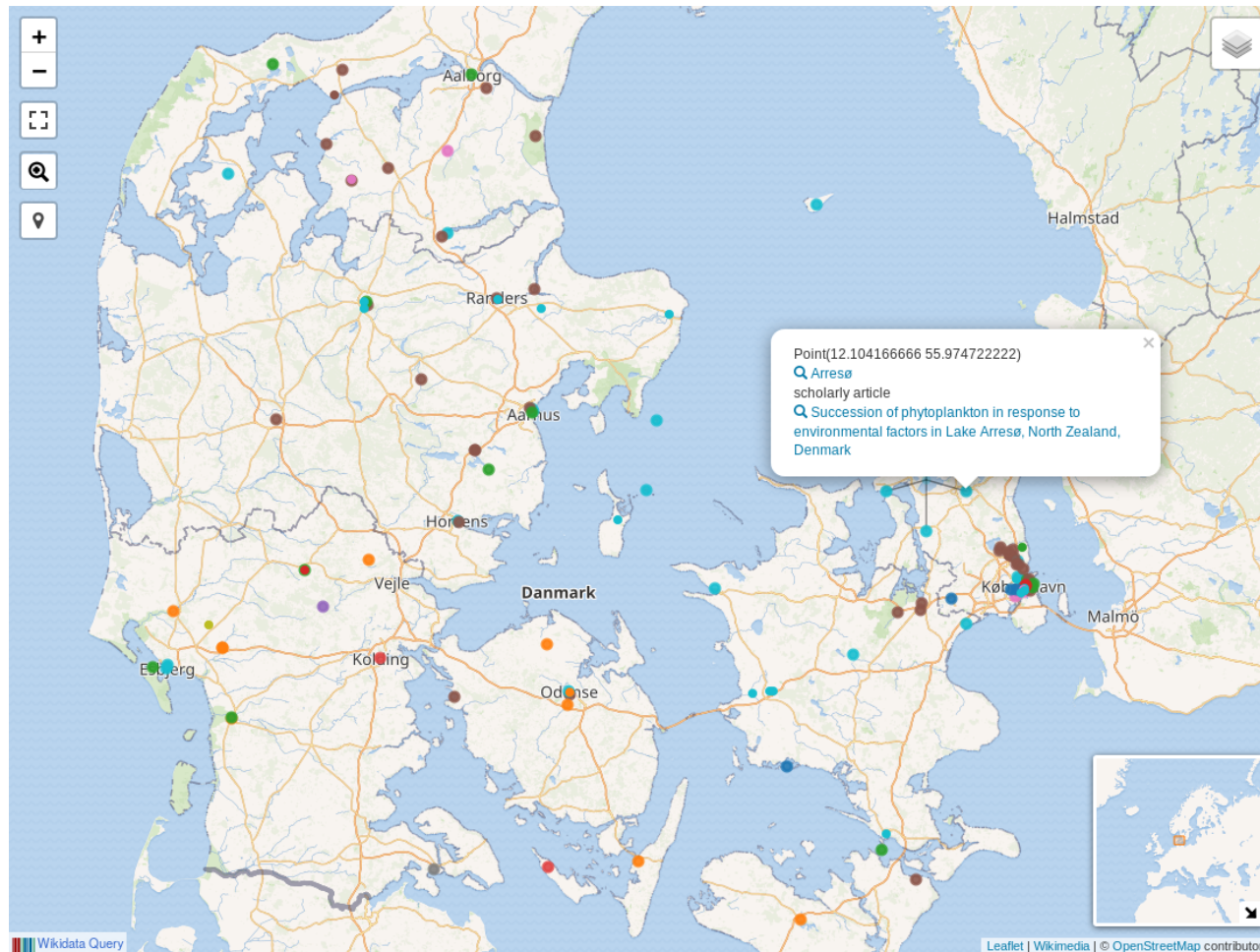
Publisher aspects



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.

Project aspect: Research projects in Scholia

Citations per budget

 Show 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., 2018).

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

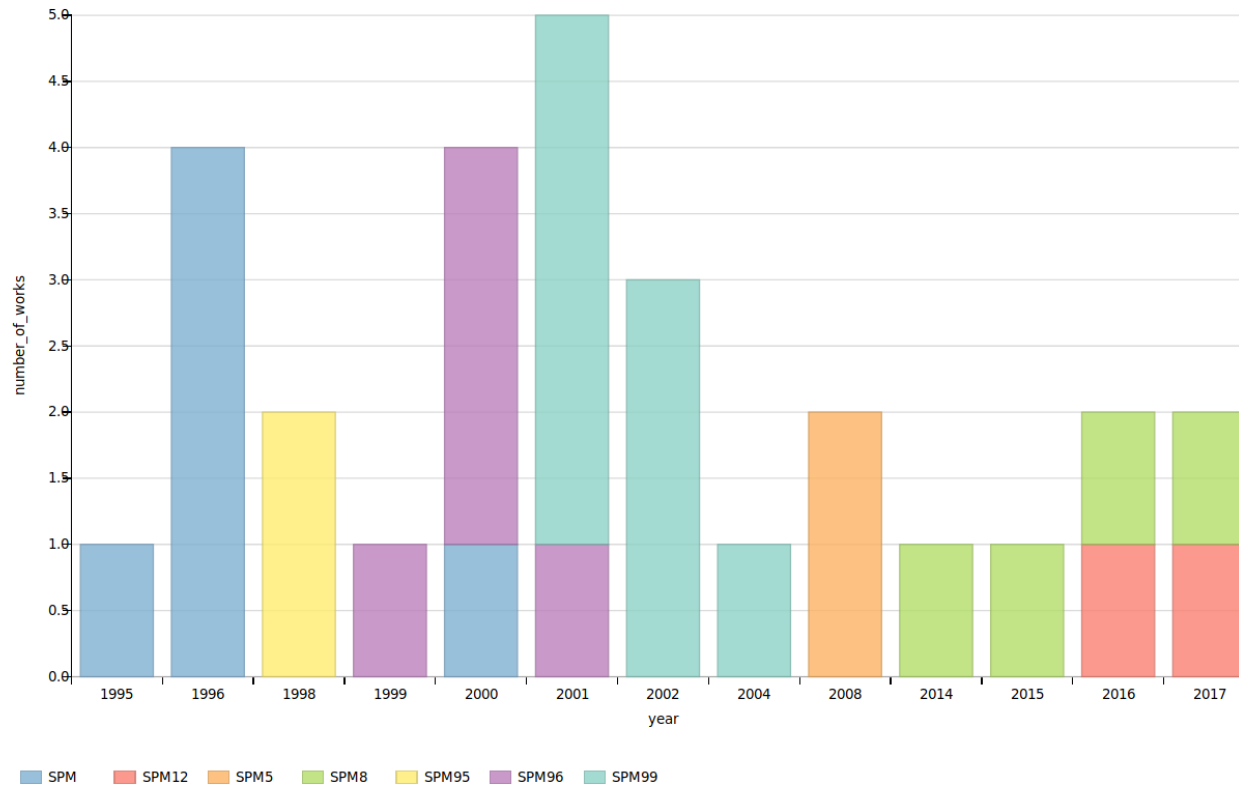
Here *citations per million budget*.

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

Use aspect

Usage over time

Works using the resource over time.

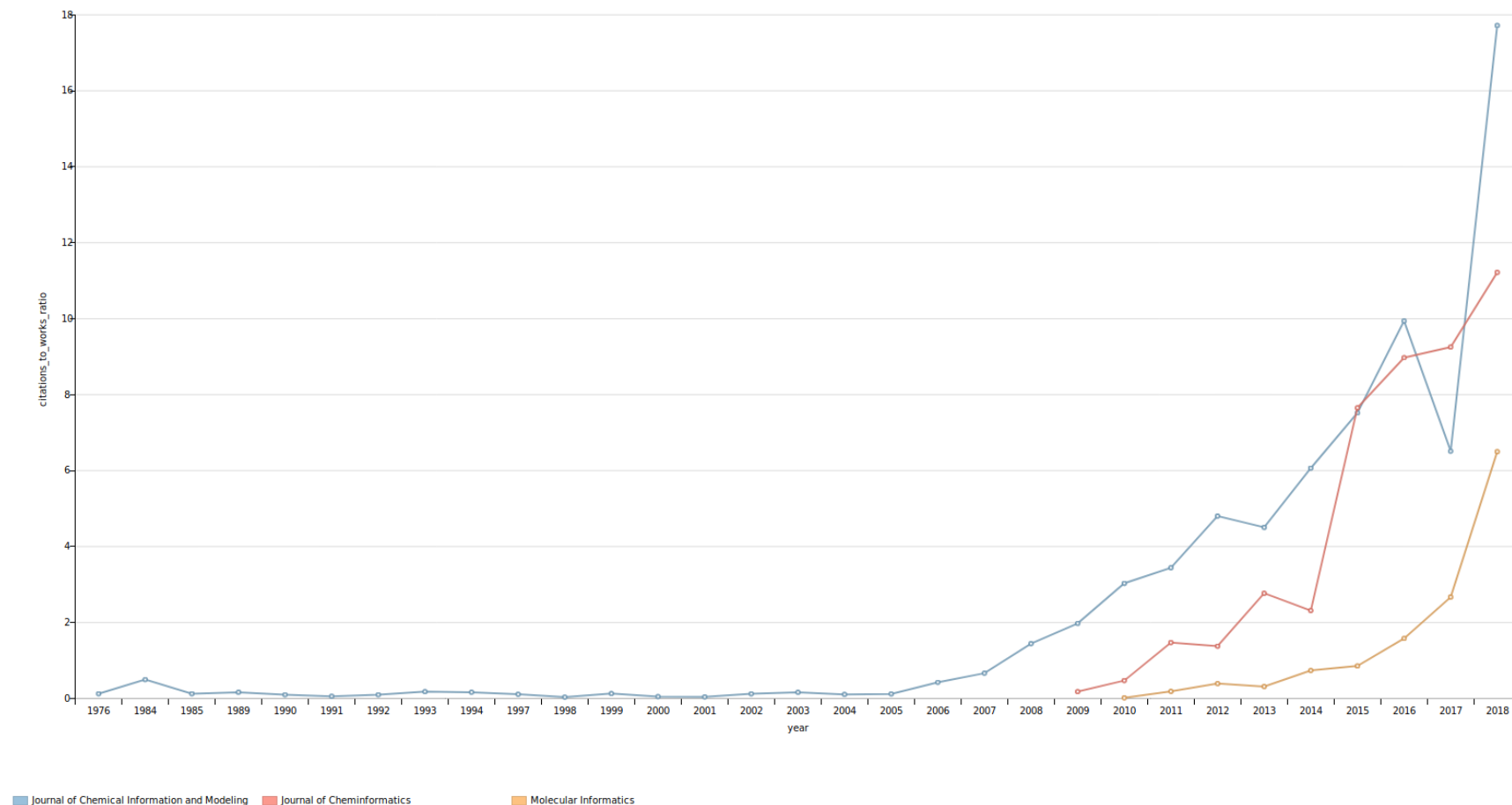


Bar chart for **usage of SPM software** (functional neuroimaging software) over time with different software versions indicated by color.

Uses the *describes a project that uses* property.

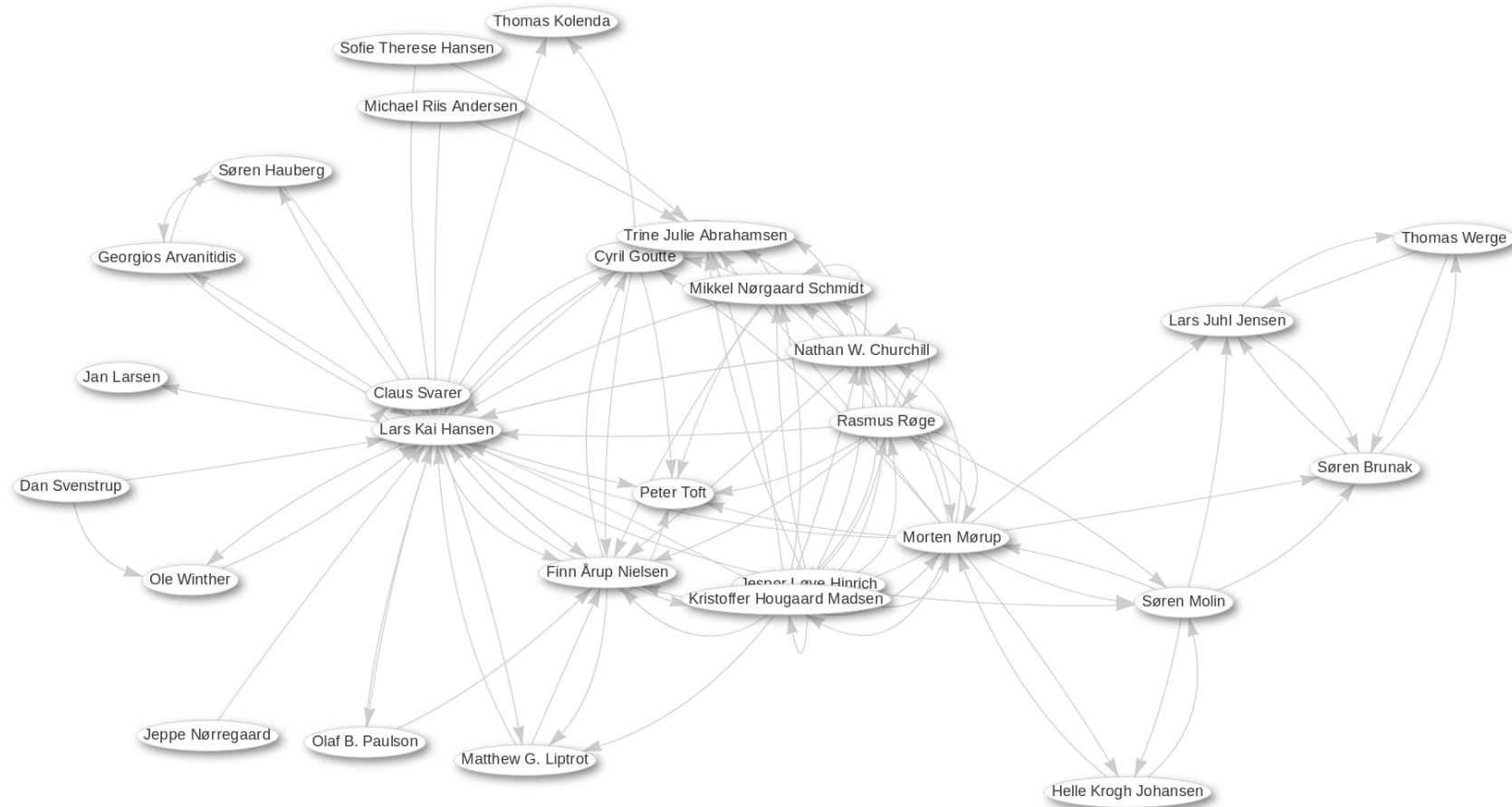
Such data is likely not available in directly machine readable format.

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/Q35/topic/Q2539>.

Geodata and Scholia

Nearby researchers

Show entriesSearch:

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	lassen 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.

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 .
```

Summary

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

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

Wikidata input is somewhat cumbersome. We rely heavily on Magnus Manske's bespoke tools.

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

WikiCite part of Wikidata continues to grow.

<https://tools.wmflabs.org/scholia>

References

- 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. (2018). The EU NanoSafety Cluster as Linked Data visualized with Scholia. DOI: [10.6084/M9.FIGSHARE.6727931](#).

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.