

Scholia

Finn Årup Nielsen

Cognitive System, DTU Compute, Technical University of Denmark

28 October 2017

**Can we make a Wikidata presenter tool
targeted at bibliographic information?**

Presenting Wikidata: Reasonator

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

- Header:** Displays the name "Finn Årup Nielsen (Q20980928)" and a list of aliases: "Finn Aarup Nielsen | Nielsen FÅ | Finn Å. Nielsen | F.A. Nielsen | F Nielsen | F A Nielsen | Finn Arup Nielsen | FA Nielsen".
- researcher:** A section with a description: "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 1996 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 with a right-pointing arrow.
- From related items:** A section with a dropdown arrow.
- cast member:** A table with one entry: "Tankens anatomi dansk dokumentarfilm" (series ordinal: 2).
- doctoral student:** A table with one entry: "Lars Kai Hansen researcher" (series ordinal: 1).
- author:** A list of publications with their titles and series ordinals:
 - Right Temporoparietal Cortex Activation during Visuo-proprioceptive Conflict (series ordinal: 2)
 - Modeling of activation data in the BrainMap? database: Detection of outliers (series ordinal: 2)
 - The Real Power of Artificial Markets (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? (series ordinal: 1)
 - On clustering fMRI time series (series ordinal: 4)
 - Persistence of Web references in scientific research (series ordinal: 7)
 - Plurality and resemblance in fMRI data analysis (series ordinal: 4)
 - Mining the posterior cingulate: Segregation between memory and pain components (series ordinal: 1)
- Timeline:** A horizontal timeline showing key events:
 - 1996: academic degree: civilingeniør
 - 1993: Danish master of science in engineering
 - 1998: educated at: Technical University of Denmark
 - educated at: Technical University of Denmark
 - educated at: Technical University of Denmark
 - academic degree: Doctor of Philosophy
 - date of birth
 - academic degree: civilingeniør
- External sites:** A dropdown menu showing "official website" and "official website".
- External sources:** A dropdown menu showing various identifiers: 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), and VIAF (316671095).
- Social media:** A dropdown menu showing "SoundCloud" (fnielsen).
- Wikimedia projects:** A dropdown menu.
- Concept cloud:** A section with a right-pointing arrow.

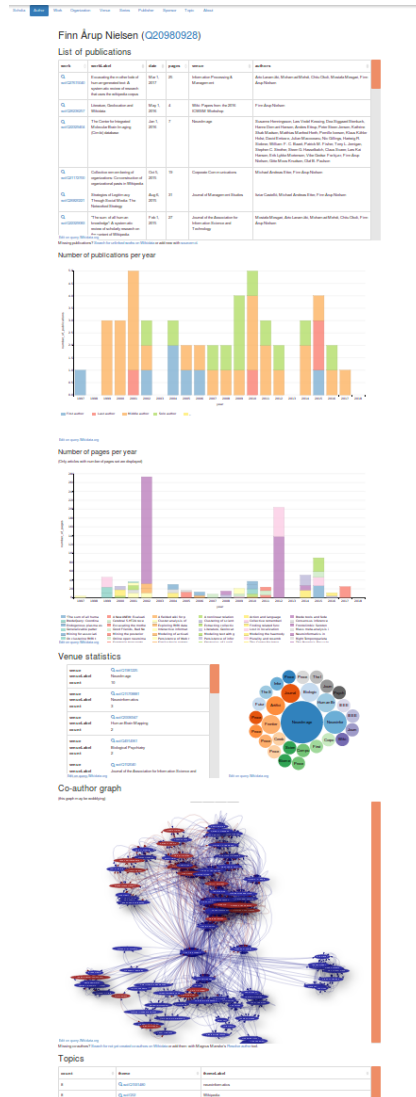
Magnus Manske's Reasonator, <https://tools.wmflabs.org/reasonator/>

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 Toolforge*.

What about citation information, aggregation of publication for an organization, etc.?

Scholia

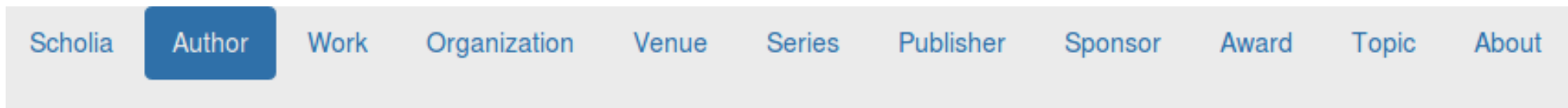


Web site with scholarly information extracted from Wikidata running from <https://tools.wmflabs.org/scholia/>.

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

Almost entirely built by using Wikidata Query Service: We generate tables, bubble charts, time lines, graphs, etc.

“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.

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 from <https://tools.wmflabs.org/scholia/author/Q20980928>.

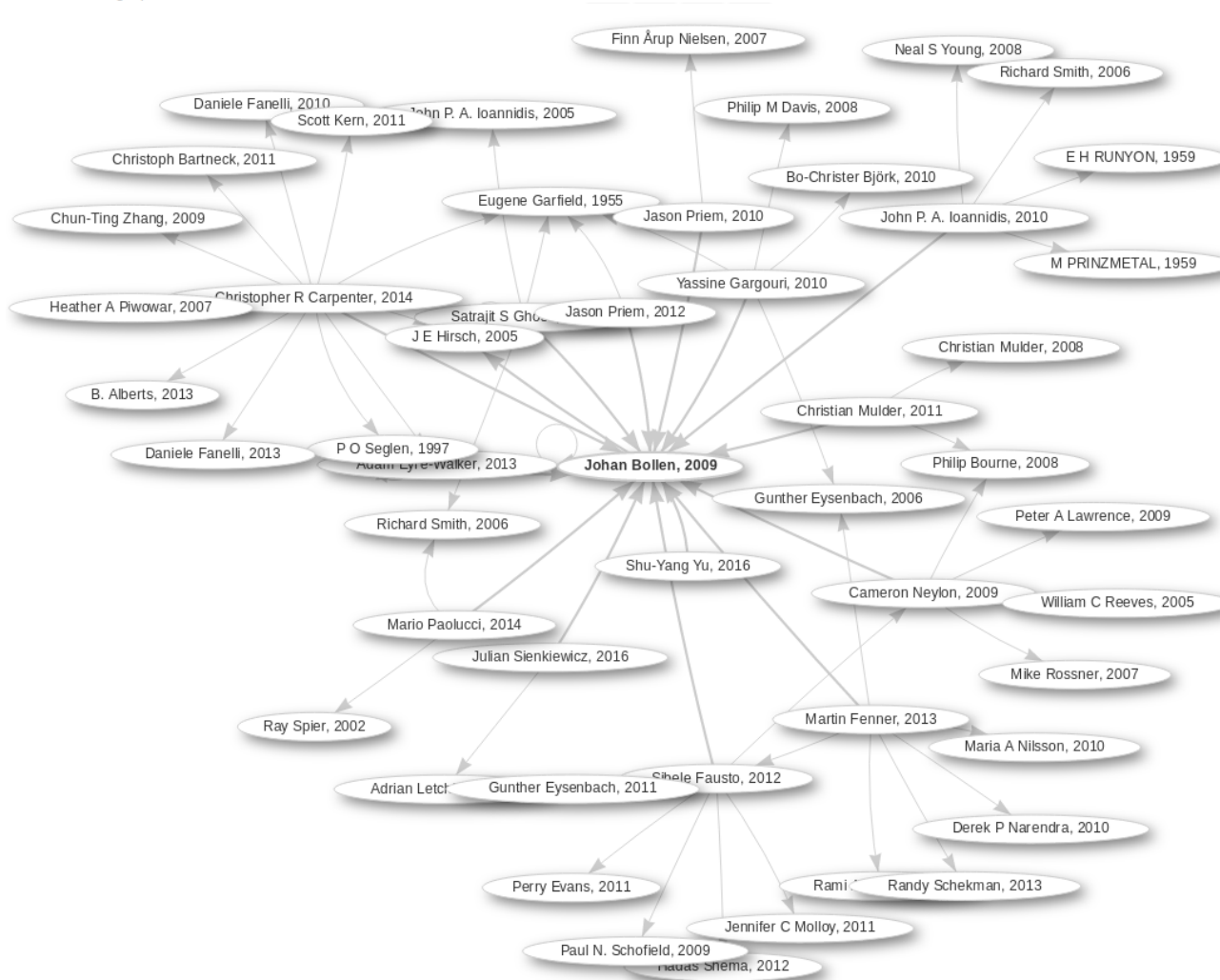
Color-coding based on author-role (first author, last author, middle author, solo author)

Using default “BarChart” of Wikidata Query Service with complex SPARQL query: <https://query.wikidata.org/#%23defaultView...>

Work aspect citation graph

Citation graph

Partial citation graph



Citation panel on *work* aspect for partial citation graph with WDQS's Graph output.

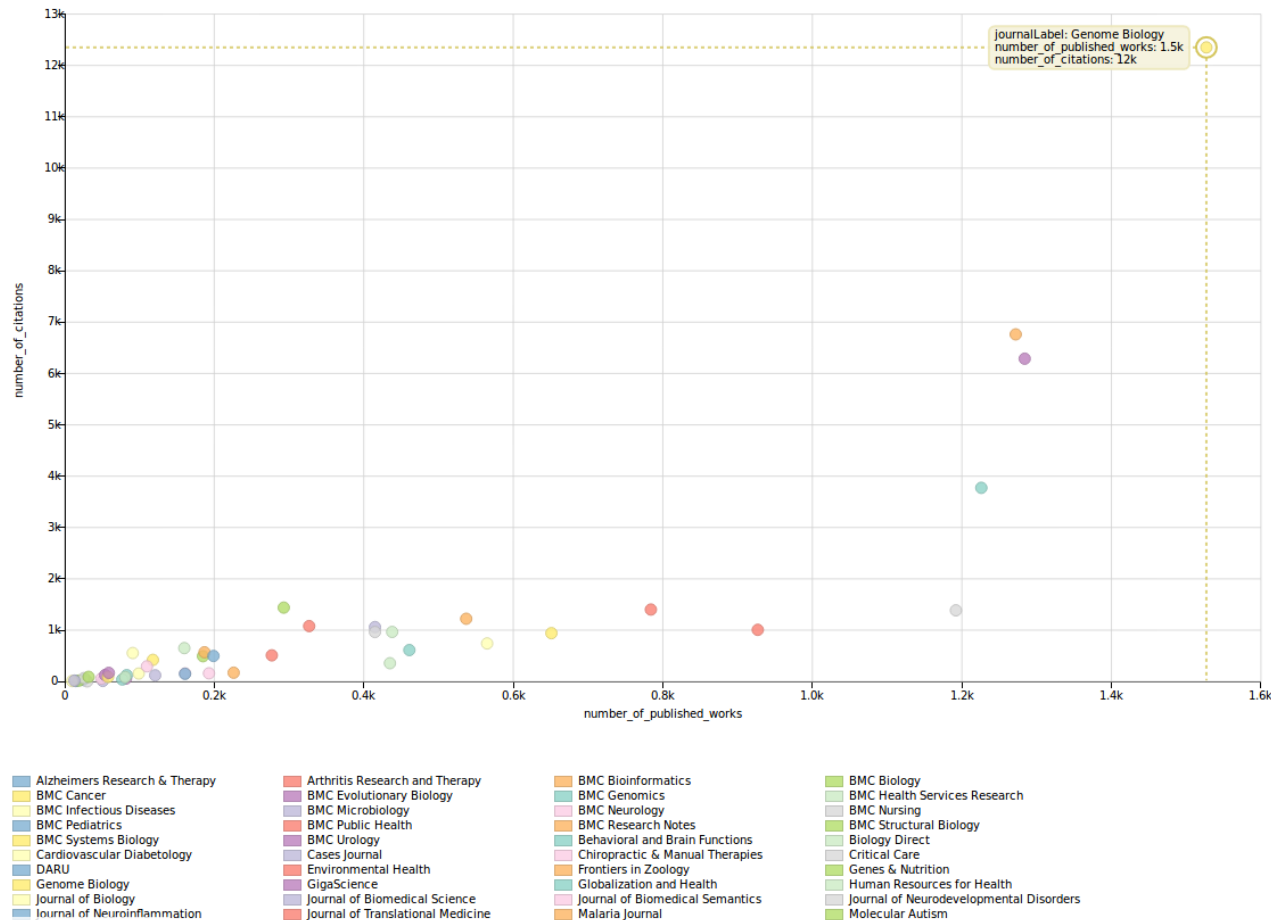
For *A principal component analysis of 39 scientific impact measures* paper.

Network constructed with the **BlazeGraph's RDF GAS API** for graph queries.

Publisher aspect

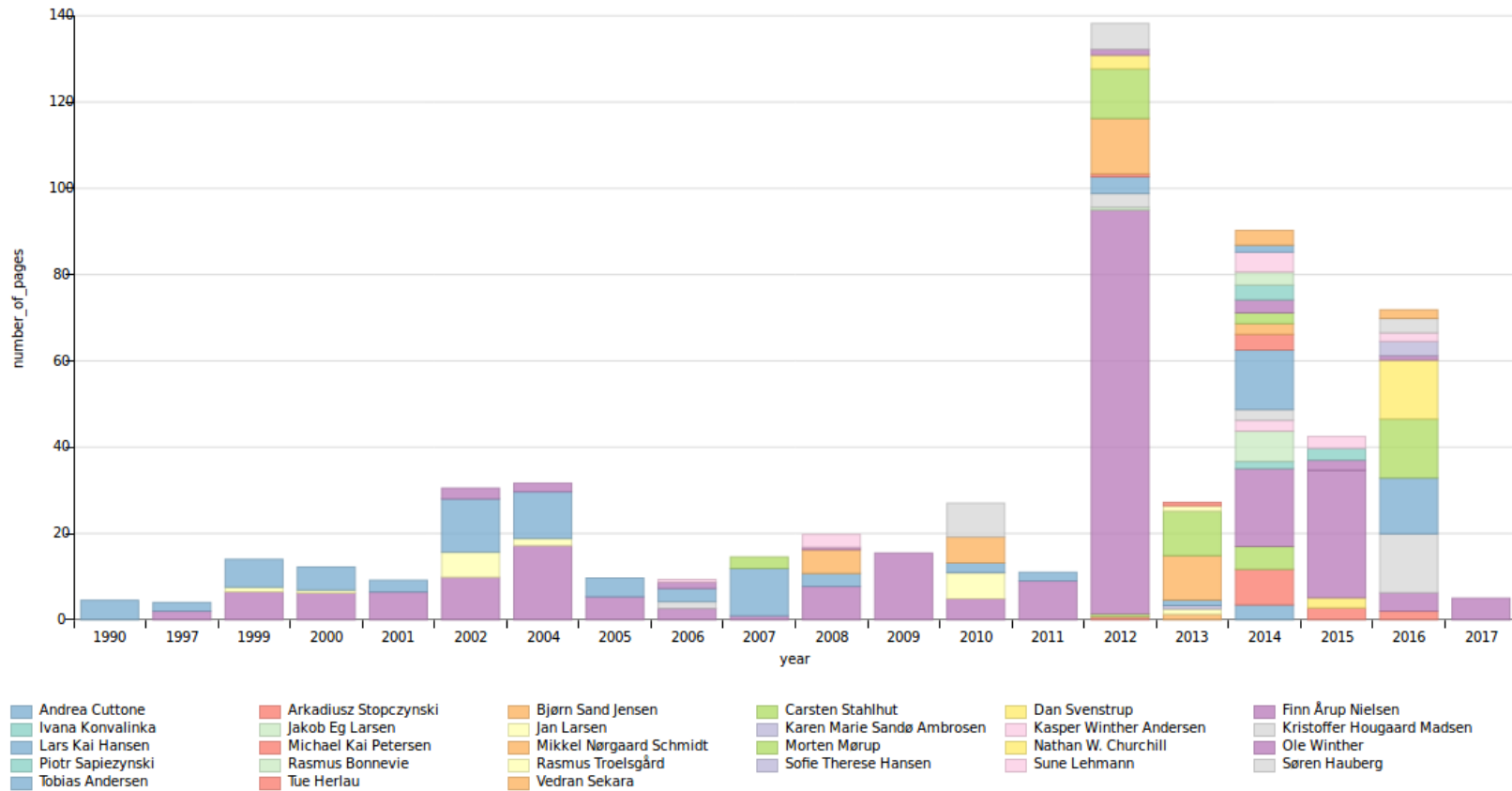
Overview of number of papers published and their citations across journals published by the publisher with Scatter output from WDQS.

Here for BioMedCentral (which may be an imprint):



<https://tools.wmflabs.org/scholia/publisher/Q463494>

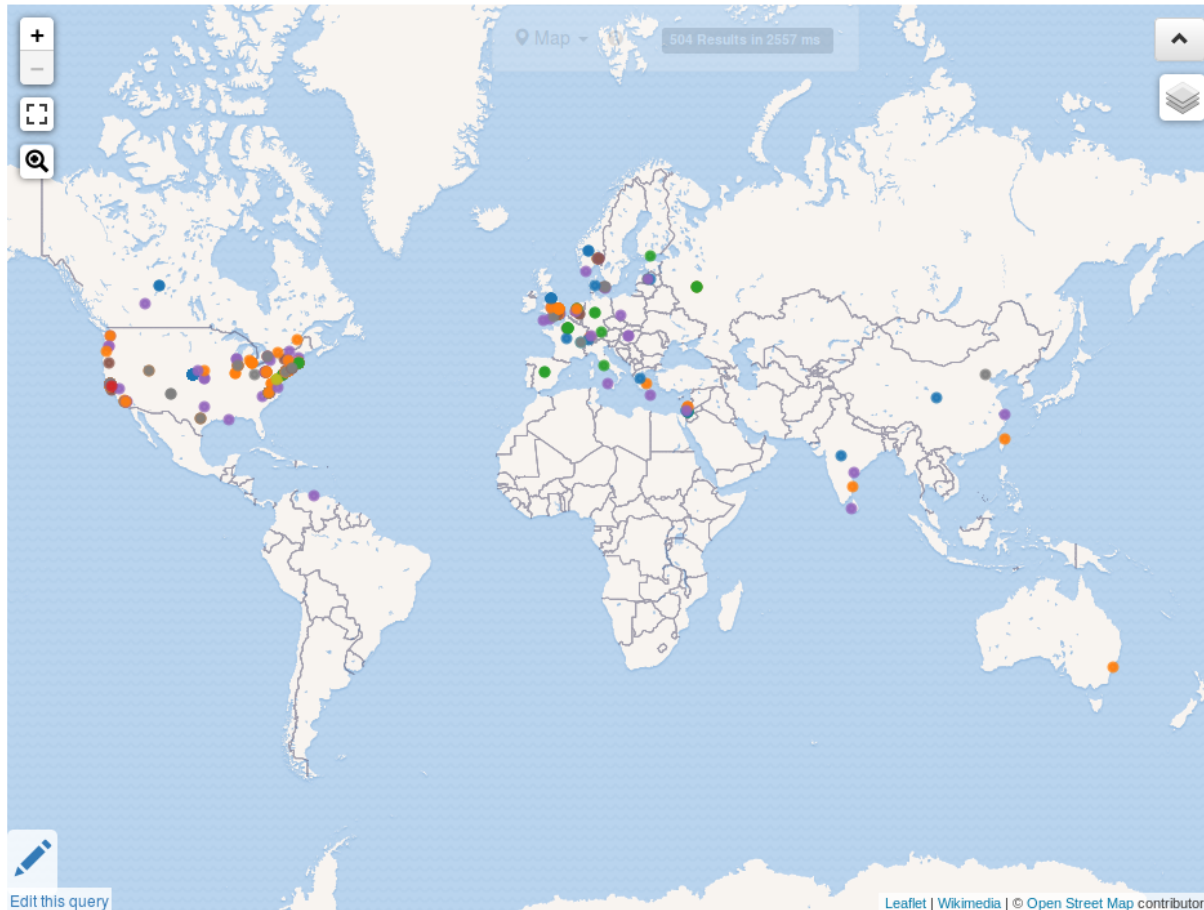
Organization aspect



Incomplete statistics on page production per year for **DTU Cognitive Systems**. Yet another color-coded WDQS Bar chart.

Award aspect

Locations of recipients



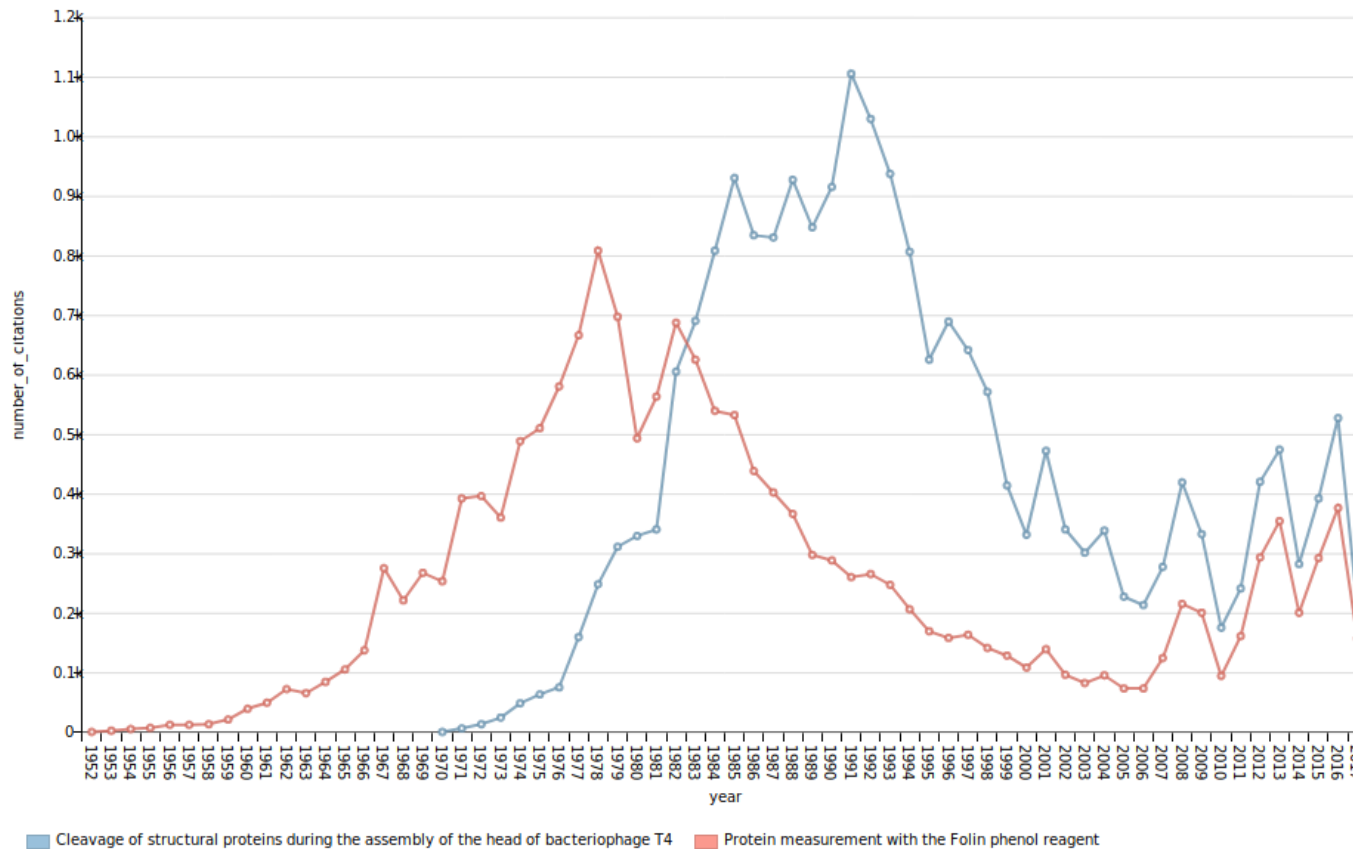
Turing award locations

The points are colored according to the “layer” column in the SPARQL result.

<https://tools.wmflabs.org/scholia/award/Q185667>

Multiple items

Citations



Aspects for multiple items: two or more works, two or more authors, two or more publishers.

Comparison of authors, universities, etc.

For instance, two works [works/Q20900776](https://www.wikidata.org/wiki/Q20900776), [Q25938983](https://www.wikidata.org/wiki/Q25938983)

Software aspect



There are several topic-related aspects: disease, protein, gene, chemical, biological pathway, software.

Software aspect: Here the panel with *Topics of works using the software* queried from P2283 with **SPM**.

Based on work by **Katherine Thornton**: Wikidata for digital preservation: A Wikidata Portal.

Scholia tool: arxiv-to-quickstatement

The screenshot shows the Scholia tool interface. On the left, the arXiv page for 'Conditional Image Generation with PixelCNN' is visible. On the right, the tool's input and result sections are shown.

Input

1606.05328

Copy and paste an ID from the [arXiv](#) preprint repository. Bare IDs (such as "1703.06103") and URIs both work.

The input ID will be queried in Wikidata and Quickstatements will not be generated if the input ID is not immediately found because of caching.

Result

```

CREATE
LAST P818 "1606.05328"
LAST P31 Q13442814
LAST Len "Conditional Image Generation with PixelCNN Decoders"
LAST P1476 en:"Conditional Image Generation with PixelCNN Decoders"
LAST P577 +2016-06-18T00:00:00Z/11
LAST P953 "https://arxiv.org/pdf/1606.05328.pdf"
LAST P820 "cs.CV"
LAST P820 "cs.LG"
LAST P2093 "Aaron van den Oord" P1545 "1"
LAST P2093 "Nal Kalchbrenner" P1545 "2"
LAST P2093 "Oriol Vinyals" P1545 "3"
LAST P2093 "Lasse Espeholt" P1545 "4"
LAST P2093 "Alex Graves" P1545 "5"
LAST P2093 "Koray Kavukcuoglu" P1545 "6"

```

[Forward to Magnus Manske's quickstatements](#)

Conversion of arXiv identifier to Magnus Manske's quickstatements which can setup up a new Wikidata item.

Implementation of Scholia

845 lines (609 sloc) | 16.7 KB

```
1  """Views for app."""
2
3  import re
4
5  from flask import (Blueprint, current_app, redirect, render_template, request,
6                    Response, url_for)
7  from werkzeug.routing import BaseConverter
8
9  from ..api import entity_to_name, wb_get_entities
10 from ..arxiv import metadata_to_quickstatements, string_to_arxiv
11 from ..arxiv import get_metadata as get_arxiv_metadata
12 from ..query import (arxiv_to_qs, cas_to_qs, doi_to_qs, github_to_qs,
13                     inchikey_to_qs, orcid_to_qs,
14                     q_to_class, random_author, twitter_to_qs)
15 from ..utils import sanitize_q
16 from ..wikipedia import q_to_bibliography_templates
17
18
19 class RegexConverter(BaseConverter):
20     """Converter for regular expression routes.
21
22     References
23     -----
24     https://stackoverflow.com/questions/5870188
25
26     """
27
28     def __init__(self, url_map, *items):
29         """Setup regular expression matcher."""
30         super(RegexConverter, self).__init__(url_map)
31         self.regex = items[0]
32
33
34     def add_app_url_map_converter(self, func, name=None):
35         """Register a custom URL map converters, available application wide.
36
```

Python (Python27 and Python35).

Uses the **Flask**, mostly selected because of its simplicity and **nice tutorial**.

JavaScript to query the Wikidata API (<https://www.wikidata.org/w/api.php>) for labels, inclusion of Wikipedia extract and setup of tables: **jQuery**, **DataTables** (elements have links to Scholia rather than Wikidata).

Test with **tox**, **flake8** and **pytest**.

Contributing

Filters Labels Milestones New issue

102 Open ✓ 52 Closed Author Labels Projects Milestones Assignee Sort

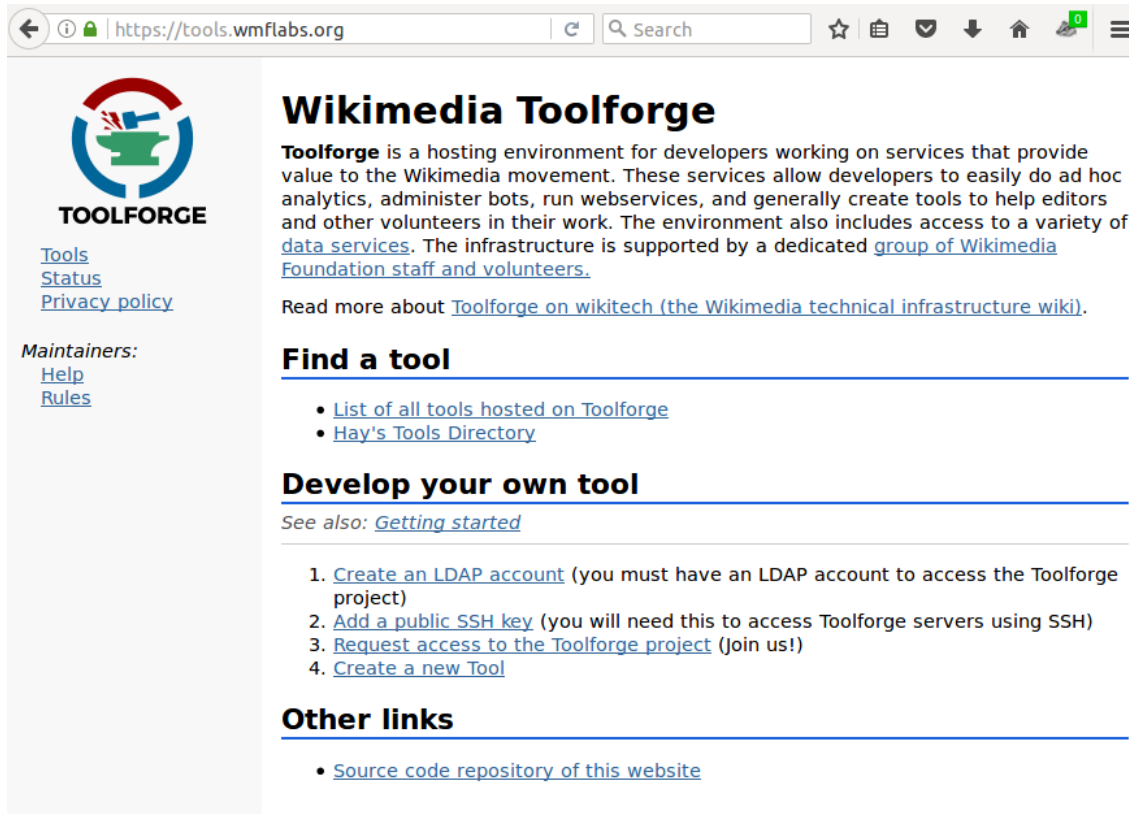
- Artifacts with the new WDQS** bug
#190 opened 2 days ago by fnielsen
- Better handling of "No data available" cases** 1
#189 opened 11 days ago by pigsonthewing
- Filter out "different from" images in "associated images" panel author aspect**
#188 opened 12 days ago by fnielsen
- Link to or display Commons media**
#187 opened 12 days ago by pigsonthewing
- Bad links on some statement values** bug 1
#186 opened 12 days ago by pigsonthewing
- Display ORCID iD on pages about individual authors** enhancement
#183 opened 20 days ago by pigsonthewing
- Co-author normalized citation for organization** enhancement 1
#179 opened 28 days ago by fnielsen
- UTF-8 character problem (perhaps zero width space)**
#177 opened on Aug 3 by fnielsen
- Add Sci-Hub link for articles that are open source** enhancement
#176 opened on Jul 31 by fnielsen
- Missing DISTINCT in award table**
#172 opened on Jul 14 by fnielsen
- what if Scholia had RSS feeds?** 1
#171 opened on Jul 14 by egonw

For contribution, GitHub users can fork **the repo at GitHub** and make pull requests under GPL.


Particularly Egon Willighagen has made a number of pull requests around proteins, biological pathways and chemicals, see, e.g., **citric acid** where he has made SPARQL queries on qualifiers and references.

We got lots of **issues**.

Deployment on Toolforge

A screenshot of the Wikimedia Toolforge website. The browser address bar shows 'https://tools.wmflabs.org'. The page features the Toolforge logo (a green and blue circular emblem with a red arc) and the text 'TOOLFORGE'. Below the logo are links for 'Tools', 'Status', and 'Privacy policy'. A 'Maintainers:' section includes links for 'Help' and 'Rules'. The main content area is titled 'Wikimedia Toolforge' and contains a paragraph describing the service as a hosting environment for developers. It includes a 'Find a tool' section with links to a list of tools and Hay's Tools Directory, a 'Develop your own tool' section with a link to 'Getting started', and an 'Other links' section with a link to the source code repository.

<https://tools.wmflabs.org>



TOOLFORGE

[Tools](#)
[Status](#)
[Privacy policy](#)

Maintainers:
[Help](#)
[Rules](#)

Wikimedia Toolforge

Toolforge is a hosting environment for developers working on services that provide value to the Wikimedia movement. These services allow developers to easily do ad hoc analytics, administer bots, run webservices, and generally create tools to help editors and other volunteers in their work. The environment also includes access to a variety of [data services](#). The infrastructure is supported by a dedicated [group of Wikimedia Foundation staff and volunteers](#).

Read more about [Toolforge on wikitech \(the Wikimedia technical infrastructure wiki\)](#).

Find a tool

- [List of all tools hosted on Toolforge](#)
- [Hay's Tools Directory](#)

Develop your own tool

See also: [Getting started](#)

1. [Create an LDAP account](#) (you must have an LDAP account to access the Toolforge project)
2. [Add a public SSH key](#) (you will need this to access Toolforge servers using SSH)
3. [Request access to the Toolforge project](#) (Join us!)
4. [Create a new Tool](#)

Other links

- [Source code repository of this website](#)

The canonical version of Scholia runs from *Toolforge*:
<https://tools.wmflabs.org/>

Toolforge (rebranded from *Wikimedia Tool Labs*), — the free cloud service provided by the Wikimedia Foundation.

Currently running gridengine. Considering Kubernetes to get better stability in response.

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{Nielsen2017ScholiaWikidataCon_slides}
```

Commands:

```
latex Nielsen2017ScholiaWikidataCon_slides.tex  
python -m scholia.tex write-bib-from-aux \  
    Nielsen2017ScholiaWikidataCon_slides.aux  
bibtex Nielsen2017ScholiaWikidataCon_slides  
latex Nielsen2017ScholiaWikidataCon_slides.tex  
latex Nielsen2017ScholiaWikidataCon_slides.tex
```

Scholia issues :(

Wikidata far from complete. Particular non-Pubmed.

Citation data lacking, but some released with I4OC.

Some queries run into WDQS time-out.

Some queries generate too large results. This is a problem from graphs such as citation networks (we put in SPARQL “LIMIT”s) and co-occurrence topics graphs (Egon Willighagen pointed to [aluminum](#)).

Scholia 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?

Building scrapers. Could focus on non-PubMed and non-DOI publications.

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, ... ?

Feeds (Egon Willighagen is working on this issue)

Thanks