Geospatial Data and Scholia[⋆]

Finn Årup Nielsen, Daniel Mietchen and Egon Willighagen

[1] Cognitive Systems, DTU Compute, Technical University of Denmark, Denmark
[2] Data Science Institute, University of Virginia, Charlottesville, Virginia, USA
[3] Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, The Netherlands

Abstract. *Scholia* is a website that displays information about scientific works represented in Wikidata. It extracts the information via queries to the extended SPARQL endpoint *Wikidata Query Service* (WDQS). Scholia handles geospatial data that may be rendered on maps with the default map output format available in WDQS. We describe the use of geospatial data in Scholia, how we combine it with other data—e.g., on topics, events and affiliations—and present a set of user stories to illustrate its current capabilities and its potential in light of the ongoing expansion of Wikidata and its integration with research-related resources.

1 Introduction

The collaboratively edited database Wikidata [8] at `https://www.wikidata.org/` has become a fast growing resource of Linked Open Data with close ties to Wikipedia and its sister sites. In terms of geographic information, Wikidata items can be described using properties with the *geographic coordinate* and *geographic shape* datatypes. These datatypes enable Wikidata to represent points and polygons. The perhaps most important geographic coordinate property is *coordinate location* (P625), which is used more than 5.8 million times, but several other properties use the geographic coordinate datatype too [4]. The recently introduced geographic shape datatype that is used by the *geoshape* property (P3896) describes a polygon by referencing a map data file in the Wikimedia Commons media archive.

To be loaded into the Wikidata Query Service (WDQS)—an extended Blazegraph-based SPARQL endpoint exposed at `https://query.wikidata.org/`—the data in Wikidata is first converted into a triple representation. When translated to RDF, WDQS represents a geographic coordinate with GeoSPARQL's [1] "wktLiteral" and "Point()". WDQS extends the standard SPARQL with a function called `geof:distance` that computes the distance between two geographic coordinates. It returns the result in kilometers, whereas in GeoSPARQL, the `geof:distance` function with the same name takes a third argument for the unit [1].

[⋆] This work is licensed under Creative Commons Attribution-ShareAlike. The maps are copyrighted by OpenStreetMap contributors and licensed under Creative Commons Attribution-ShareAlike 2.0.

Apart from acting as a SPARQL endpoint, WDQS can format the output in various ways: as tables, charts, graphs and —of particular relevance in geospatial applications— as a geographic map, where geographic coordinates returned by a SPARQL query are rendered on a zoomable OpenStreetMap-derived world map. WDQS can also render points and polygons with different colors on the map. The annotated map is automatically generated by WDQS if the SPARQL query contains a column with geospatial values and the query specifies map output with the line "`#defaultView:Map`". All HTML-based output from WDQS includes a link in the lower left corner to the WDQS editor where the SPARQL query can be edited, directly in the raw SPARQL code and in a form-based editor.

Spawned by the WikiCite effort [6,7] that integrates bibliographic information with Wikidata, we have built Scholia [5], a website that exposes the WikiCite-associated data from Wikidata with a public website, primarily using calls to WDQS: `http://tools.wmflabs.org/scholia/`. Scholia presents bibliographic and scientific information in so-called aspects, e.g., for a work, an author, an organization, a sponsor, a publisher, or a topic.

The aspect is part of the URL. For example, `http://tools.wmflabs.org/scholia/topic/` is the prefix for the *topic* aspect, which includes sub*topics*. A particular Wikidata item may be viewed in one or more of these aspects: by default a university will be displayed as an organization, but can also be viewed as a sponsor, as a topic, or as a publisher. Each aspect displays multiple information panels, each constructed by specific SPARQL queries to WDQS and showing the results in a table, a plot, a network graph, or a map. And like visualizations of WDQS itself, each panel links back to the WDQS query editor. The SPARQL queries are automatically generated from predefined templates, where the variables in the templates are user-provided Wikidata item identifiers. The Scholia user can provide the identifiers either by text search or by clicking on links in the tables in the Scholia interface or —through a JavaScript gadget— from a Wikidata item page.

We have previously described Scholia [5] and our contribution here focuses on the geospatial part of Scholia and is a description of new ways that we present scientific information. We furthermore set up a few user stories to illustrate Scholia's ability to answer realistic geospatial questions. Our motivation for extending Scholia with geospatial information is to provide alternative ways to discover scientific information.

2 Related work

Magnus Manske has built several tools utilizing the geospatial data in Wikidata. His Reasonator (`https://tools.wmflabs.org/reasonator/`) shows OpenStreetMap maps if a queried item has geographic coordinates. The wikidata-todo tool (`https://tools.wmflabs.org/wikidata-todo`) can show Wikidata items on a map around a queried geographic coordinate. His mobile app WikiShootMe! (`https://tools.wmflabs.org/wikishootme/`) identifies nearby places having a Wikidata item but no associated image.

The Wikidata website itself has a feature to identify nearby Wikidata items: https://www.wikidata.org/wiki/Special:Nearby. It takes a geolocated item as an optional parameter, e.g., #/page/Q3150 for the German city of Jena. We have previously considered geospatial data in Wikidata, describing an application displaying narrative locations of literary works on a Danish map [4].

DBpedia Mobile is an application for the smartphone presenting geographically nearby information from DBpedia on a map [2,3]. DBpedia has also been used in other systems, enhancing it with a locally-stored reference geo-dataset, enabling proximity queries to line and polygonal representations of named features [9].

3 Geospatial information in Scholia

We have introduced two new Scholia aspects of particular geospatial relevance: *country* and *location*. The *country* aspect shows information with respect to a country. For instance, /scholia/country/Q35 is the *country* aspect for Denmark. Currently, this aspect shows a list of academic organizations in the country and a list of authors associated with these organizations.

The *country* aspect also shows two maps: One map displays organizations across the world for which Wikidata is aware of affiliated authors of publications co-authored with people affiliated with organizations of the queried country, as can be seen in Fig. 1. A second map displays the locations with geographic coordinates in the queried country that are mentioned as the main subject of a work, see Fig. 2, where Lake Esrum in the North-Eastern part of Denmark is the subject of a scientific article.

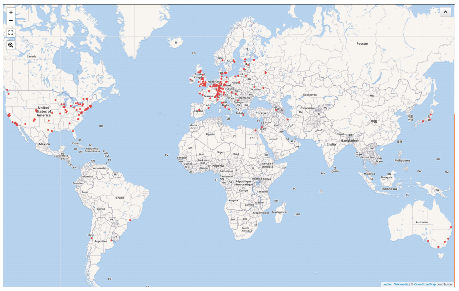
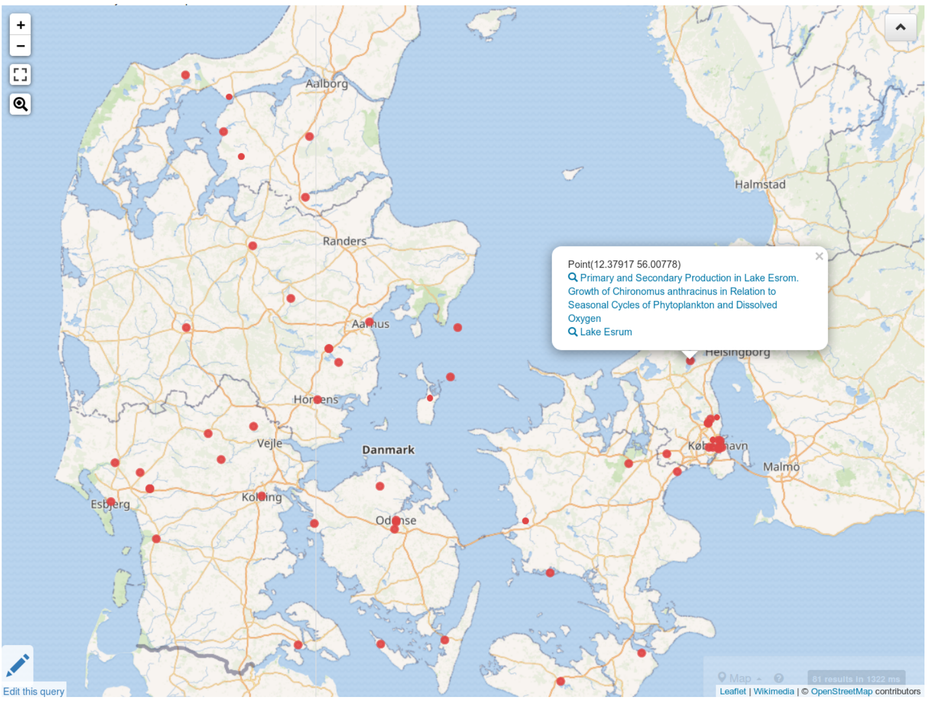


Fig. 1. Map in Scholia with international collaborators of authors based in the Netherlands.

Map © OpenStreetMap contributors. CC BY-SA 2.0.

The *location* aspect shows organizations (universities, departments, research groups, etc.) located nearby the queried location. For instance, https://tools.wmflabs.org/scholia/location/Q3806 shows the page for Tübingen and that *Knowledge Media Research Center* and *Forschungsstelle für Planungs-, Verkehrs-, Technik- und Datenschutzrecht* are amongst the nearby organizations. For this list, the SPARQL query identifies the geographic coordinates associated with Tübingen and all organizations that are a subclass of the university item or a part of such a subclass. To order the returned values according to distance, the `geof:distance` function is used. A related geospatial SPARQL query on the same aspect page lists nearby locations that are topics of works, e.g., https://tools.wmflabs.org/scholia/location/Q1142544 for the Wellcome



Point(12.37917 56.00778)
Primary and Secondary Production in Lake Esrom. Growth of Chironomus anthracinus in Relation to Seasonal Cycles of Phytoplankton and Dissolved Oxygen
Lake Esrum

Fig. 2. Denmark in the country aspect in Scholia with a panel showing the locations in Denmark that are the main subject of a work.

Map © OpenStreetMap contributors. CC BY-SA 2.0.

Trust Sanger Institute lists an article titled "Three Drawings by John James at Audley End" about Audley End House (Q758949), a nearby country house.

Another recent addition to Scholia are so-called subaspects that take two Wikidata items of different classes as arguments and facilitate a kind of faceted search. Two subaspects of geospatial relevance have been implemented so far: A country–topic subaspect and a location–topic subaspect.

The country–topic subaspect allows the Scholia user to "zoom in" on a specific topic with respect to a country. For instance, the subaspect at https://tools.wmflabs.org/scholia/country/Q30/topic/Q202864 shows information about researchers associated with the United States (Q30) who have written works about the Zika virus (Q202864). Current plots are two graphs with the co-author network and the co-citation network of US-based Zika virus researchers.

The other subaspect combines location–topic information by using the geographic coordinates associated with a location to find nearby researchers involved with the topic using a point query. For instance, https://tools.wmflabs.org/scholia/location/Q727/topic/Q910164 lists people that are close to Amsterdam (Q727) and authors of research articles about cheminformatics (Q910164).

There are a couple of other aspects in Scholia where we show geospatial information on maps. For an author, we show associated locations on a map, e.g., educational institution, employers and place of birth. For the award aspect, we show locations associated with the award recipients. We expect to implement more of such maps and more features to help Scholia users contribute content.

4 User stories

Below, we present five realistic user stories and show how they can be handled with Scholia. Their usefulness should increase as Wikidata coverage improves.

1. *You are to review research applications from Finland about machine learning and related research fields. You are based outside Finland and would like to get an overview of Finnish researchers and research organizations in that research area, their works as well as their collaboration and citation patterns.*
   Scholia can present an overview of Finnish machine learning with the country–topic subaspect at `https://tools.wmflabs.org/scholia/country/Q33/topic/Q2539`. The current page shows a list with 19 authors, with the most prolific author, Erkki Oja, listed with 5 works. The co-author graph shows one connected component, while the co-citation graph is empty. More dense graphs can be observed for Denmark.
2. *As a machine learning researcher or student, you are visiting Copenhagen and wish to identify machine learning researchers or research groups for possible collaboration.*
   Scholia presents a list of machine learning (Q2539) researchers near the center of Copenhagen (Q1748) in the location–topic subaspect at `https://tools.wmflabs.org/scholia/location/Q1748/topic/Q2539`. At the top are natural language processing researchers from the University of Copenhagen. Further down the list, one finds researchers at the Technical University of Denmark (located in a suburb of Copenhagen) and researchers at universities in the Swedish province of Scania, i.e., in a neighboring country.
3. *You are a researcher interested in Wikipedia research and planning a visit to Tübingen where you would like to meet other Wikipedia researchers.*
   This user story is basically the same as the above, but with Tübingen (Q3806) instead of Copenhagen and Wikipedia (Q52) in place of machine learning: `https://tools.wmflabs.org/scholia/location/Q3806/topic/Q52`. The list shows researchers from the University of Tübingen and the Max Planck Institute for Biological Cybernetics, but also other researchers further away at the Karlsruhe Institute of Technology.
4. *You are going to The Web Conference in April 2018 in Lyon. You want to know if there is any other relevant scientific meeting in the local area at that time, preferably just before or just after the conference.*
   The event aspect for The Web Conference 2018 (Q48910401) at `https://tools.wmflabs.org/scholia/event/Q48910401` shows related conferences, e.g., the current ranking shows the *Wiki Workshop 2018* as the most related event. This is a workshop colocated with the conference.

5. *As a university administrator, you would like to get an overview of your university's collaborations or possible collaborations in another country, e.g., an administrator at the Technical University of Denmark wants to know collaboration patterns to South Korea.*
   The country aspect for South Korea (Q884) at `https://tools.wmflabs.org/scholia/country/Q884` shows a map of organizations across the world that collaborate with South Korean organizations. Currently, the map shows only a few organizations. These organizations are located in Japan, Denmark, Germany, England and the United States. For Denmark, only a single researcher is listed as having a collaboration link to South Korea, and he is affiliated with the Technical University of Denmark.

All SPARQL queries are available from links in the lower left corner in each panel in the Scholia interface. We also note that all our user stories involve solutions with aspects (location, country and event) not available in the previously described version of Scholia [5].

5 Discussion

Scholia can address user stories like those presented in this paper, albeit with the caveat that the results returned are often still considerably limited by the amount of relevant data currently available in Wikidata. Conversely, Scholia queries provide one way to identify gaps in Wikidata coverage: The machine learning literature, scientific events and South Korean researchers, for instance, are not well covered at the moment.

Scholia queries can also help identify starting points for curating information related to those gaps. For instance, the number of different locations mentioned as topics in works vary considerably between countries and regions. For Sweden and Wales, many such mentions are available through Wikidata, whereas they number very few for South Korea. For Sweden, many of the locations stem from archeological articles published in the journal *Fornvännen* (Q4162197).

There is room for enriching the set of queries Scholia uses. For related events, for instance, time and location are not the only relevant features: The topic and type of the events should also be incorporated in the scoring of related events. An event in Wikidata may be characterized by main subject, organizer, speaker, participant and sponsor as well as the event series it is typically a part of. These features would be relevant to add to the query. The ordering for related events is currently based on a somewhat *ad hoc* scoring as a multiplication of an inverse distance and the inverse difference between the time of the events.

We note that the query associated with the first user story about Finnish machine learning researchers works without geospatial data and uses hierarchical relationships between administrative territorial entities instead. Both approaches could be combined in further queries, since Wikidata also links those administrative entities with geospatial data.

Acknowledgment: This research is funded by the Innovation Foundation Denmark through the DABAI project.

References

1. OGC GeoSPARQL - A Geographic Query Language for RDF Data (September 2012), https://portal.opengeospatial.org/files/?artifact_id=47664
2. Becker, C., Bizer, C.: DBpedia Mobile: A Location-Enabled Linked Data Browser. Proceedings of the Linked Data on the Web Workshop, Beijing, China, April 22, 2008 (July 2008), http://ceur-ws.org/Vol-369/paper13.pdf
3. Becker, C., Bizer, C.: Exploring the Geospatial Semantic Web with DBpedia Mobile. Web Semantics: Science, Services and Agents on the World Wide Web 7, 278–286 (December 2009)
4. Nielsen, F.Å.: Literature, Geolocation and Wikidata. Wiki (May 2016), http://www2.compute.dtu.dk/pubdb/views/edoc_download.php/6934/pdf/imm6934.pdf
5. Nielsen, F.Å., Mietchen, D., Willighagen, E.: Scholia, Scientometrics and Wikidata. The Semantic Web: ESWC 2017 Satellite Events (October 2017), http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/7010/pdf/imm7010.pdf
6. Taraborelli, D., Dugan, J.M., Pintscher, L., Mietchen, D., Neylon, C.: WikiCite 2016 Report (November 2016), https://upload.wikimedia.org/wikipedia/commons/2/2b/WikiCite_2016_report.pdf
7. Taraborelli, D., Pintscher, L., Mietchen, D., Rodlund, S.R.: WikiCite 2017 report (December 2017), https://ndownloader.figshare.com/files/9928825
8. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57, 78–85 (October 2014), http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext
9. Younis, E.M.G., Jones, C.B., Tanasescu, V., Abdelmoty, A.I.: Hybrid Geo-spatial Query Methods on the Semantic Web with a Spatially-Enhanced Index of DBpedia. Geographic Information Science pp. 340–353 (December 2012)