Q392754: Criação de perfis académicos com Wikidata e Scholia – projecto-piloto

Q392754: Generating NOVA SBE scholarly profiles using Wikidata and Scholia – a pilot project

Ana Catarina Mateus Reis

Social Sciences DataLab, NOVA School of Business and Economics
Universidade NOVA de Lisboa
catarina.reis@novasbe.pt

Miguel Mimoso Correia

Teresa e Alexandre Soares dos Santos Library, NOVA School of Business and Economics
Universidade NOVA de Lisboa
miguel.correia@novasbe.pt

Resumo

A Wikidata é uma plataforma de conhecimento gratuita, aberta, multilingue e colaborativa. Funciona como um repositório de dados estruturados e ligados que podem ser usados noutros projetos, como o Scholia, para criação de perfis académicos e visualização de dados. Neste artigo apresentamos o nosso contributo para enriquecer o perfil da NOVA School of Business and Economics com metadados relevantes, permitindo a apresentação de informação acerca da produção científica da escola. Numa primeira abordagem de carácter exploratório, percebemos que a Universidade NOVA de Lisboa e a NOVA SBE tinham já alguma informação na Wikidata, mas que precisava de ser enriquecida e contextualizada. Partindo de um pequeno grupo de investigadores, extraímos dados do nosso Current Research Information System (CRIS) que depois importámos para a Wikidata. Este projecto-piloto permitiu-nos ganhar experiência e competências em

Abstract

Wikidata is a free, open, multilingual and collaboratively edited knowledge base. It works as a repository of structured and linked data that that can be used by other projects, such as Scholia, to generate scholarly profiles and data visualizations. In this paper we present our contribution to enrich NOVA School of Business and Economics (NOVA SBE) profile with meaningful metadata, enabling the display of useful information about the scientific production of the School. In an exploratory approach, we realized that Universidade NOVA de Lisboa and NOVA SBE already had scattered information on Wikidata but it needed to be enriched and contextualized. We started with a small group of researchers, extracted data from our Current Research Information System (CRIS) and then imported it to Wikidata. This pilot allowed us to gain experience and skills on Wikidata. As we continue to add more data, we

Wikidata. À medida que acrescentamos mais dados, esperamos conseguir avaliar o impacto dos dados abertos e ligados na disseminação de informação científica.

expect to be able to evaluate the impact of open linked data in the dissemination of scholarly information.

Palavras-chave: Wikidata, Scholia, comunicação científica.

Keywords: *Wikidata, Scholia, scholarly communication.*

1. Introduction

Wikidata is a free, open, multilingual and collaboratively edited knowledge base, developed by the Wikimedia Foundation. Launched in 2012 ("Wikidata," 2019), it works as a hub of structured data that can be used by other Wikimedia projects, such as Wikipedia or Wikiquotes, and tools like Scholia. Anyone can contribute to this platform, as well as copy, modify and distribute its content. All data is published under the Creative Commons Public Domain Dedication – CC 1.0 Universal ("Wikidata: Data access," 2019). Edits on Wikidata can be made manually (one item at the time) or using tools or bots (Lemus-Rojas and Pintscher, 2018, p.154).

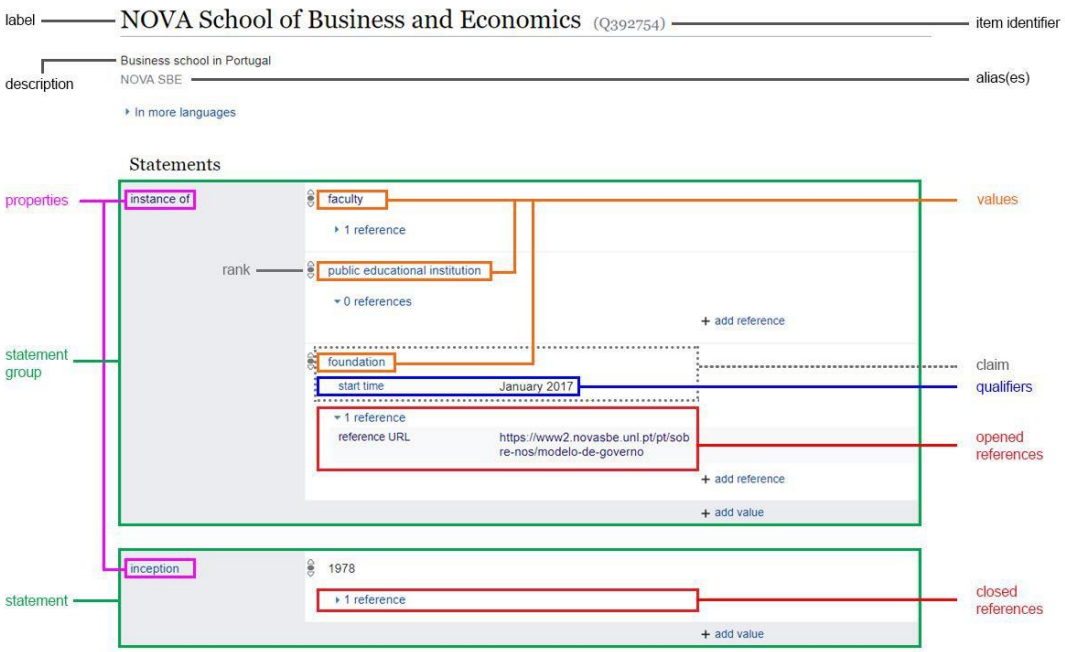
According to a recent survey by OCLC (Online Computer Library Center) Wikidata was, in 2018, the 5th most used data source platform for linked data projects, compared to a 15th place in ranking on a previous survey in 2015. It competes directly with well-known linked data sources, such as id.loc.gov (Library of Congress), VIAF (Virtual International Authority File), DBpedia and Geonames (Smith-Yoshimura, 2018).

The Wikidata data model works similarly to Resource Description Framework (RDF), as most of the data is encoded via an item (rdf: subject), a property (rdf: predicate) and a value for that property (rdf: object), enabling knowledge to be represented in a machine-readable way.

There are two types of entities on Wikidata – items, which represent topics, concepts and objects, and properties – each one with its own wiki page. Each entity has a unique identifier: an item identifier is a sequential number prefixed with Q, while a property identifier is prefixed with P. They also have a Uniform Resource Identifier (URI) consisting of the pattern: https://www.wikidata.org/entity/*identifier*. This acts as a permanent link to a group of metadata that represents the entity, like DOI numbers do. Q392754 is NOVA SBE's identifier on Wikidata.

As to the data structure, entities have a label (mandatory field), which is "the most common name that the item would be known by" ("Help:Label," 2019); a description (that enables to disambiguate items with a similar or the same label); a list of aliases (other names the entity is known by); and statements.

Figure 1. Wikidata data model



Properties are paired with, at least, one value, forming a statement. They are also connected to items, thus creating a structure of linked data. Qualifiers are used to give additional context information beyond the property values. It is possible to add references for the values provided, improving data quality and reliability. External identifiers, which can be added to items, can also help to achieve this goal. At the same time, they describe the item and provide a connection to external databases. Finally, items have a list of links to pages about the item on other Wikimedia projects.

2. The pilot project

Teresa e Alexandre Soares dos Santos Library and Social Sciences DataLab are committed to developing open knowledge and open access projects aligned with NOVA SBE research goals. By contributing data about the scientific production of their schools to open data projects like Wikidata, libraries can have a major role in the dissemination of science. Lemus-Rojas and Pintscher (2018) urge information professionals to use their expertise to contribute to Wikidata. The authors point to the experience of these professionals in the creation and implementation of standards, as well as their “deep appreciation and understanding of the value of creating structured metadata” (Lemus-Rojas and Pintscher, 2018, p.150).

This pilot project is a good opportunity to explore the potential of Scholia to disseminate the work of NOVA SBE’s researchers, under the FAIR principles (Findable, Accessible, Interoperable and Reusable). Wikidata improves data findability using unique and persistent identifiers, as well as rich metadata to describe those items. This metadata is machine-readable, allowing the automatic registration and indexation of datasets by search engines, thus improving discoverability. Being a collaborative open database, anyone with an internet connection can access and edit its contents. Although there are risks and challenges associated with collaborative platforms, collaborative stewardship and actions of data improvement by multiple users contribute to a more complete description of items and relations between them. The focus on structured and interoperable data ensures that they can be reused in

different contexts and applications. For instance, having scholarly information available on Wikidata can help generate better and more credible Wikipedia articles.

Filling Wikidata with rich, open and accessible information allows new insights and new perspectives about in-house research. Using Scholia as an aggregator for the visualization of query-based data will allow us to uncover dependencies, relationships and bibliometric information referring to our research community.

We started the project by identifying what information was available on Wikidata. Although NOVA SBE's Wikidata page had already some information, when looking at the Scholia profile we verified that the queries returned no results. This was an indication that there were no links between Universidade NOVA de Lisboa and NOVA SBE profiles. Relevant information about these two items was added, including the statements "subsidiary" e "parent organization," creating relationships between the two items.

Our pilot was based on a small group of NOVA SBE researchers. Bibliographic information was collected by using NOVA's Current Research Information System (CRIS) and exported to an Excel file. The next step was cleaning the data and select what to import to Wikidata.

Since there are many properties that can be used on Wikidata, we went through a careful selection of the properties to use for item description. We wanted to make sure that we included all the most relevant properties for building scholarly profiles and bibliographic objects. Three main types of items were identified: authors, scholarly articles and books. The tables below present the selected properties for author profiles, scholarly articles and books.

Table 1. List of Wikidata properties for authors.

Property Label	**Property ID**
instance of	P31
sex or gender	P21
country of citizenship	P27
date of birth	P569
occupation	P106
employer	P108
educated at	P69
affiliation	P1416
ORCID-id	P496
Scopus Author ID	P1153
ResearcherID	P1053
VIAF ID	P214
ISNI	P213

Table 2. List of Wikidata properties for scholarly articles.

Property Label	Property ID
instance of	P31
title	P1476
main subject	P921
author	P50
author name string	P2093
language of work or name	P407
publication date	P577
number of pages	P1104
published in	P1433
volume	P478
page(s)	P304
full work available at	P953
issue	P433
DOI	P356

Table 3. List of Wikidata properties for books.

Property Label	Property ID
instance of	P31
author	P50
author name string	P2093
language of work or name	P407
publisher	P123
title	P1476
publication date	P577
number of pages	P1104

The next step consisted in creating missing profiles from our group of researchers, as well as their publications. From a universe of around 200 professors and researchers we populated Wikidata with 70 profiles and their scientific production. Some researchers already had a profile on Wikidata but there was no link to NOVA SBE. Similarly, article items were also found but without a connection to their authors. We also added co-authors that did not have an item page. It was not possible to create statements with all the properties for all the researchers, because sometimes that information did not exist (for example, not all the researchers have an ORCID number) or we could not find the information.

The last step was automation and mass import. We started by adding information manually, but we soon found out that the process was too time consuming, and we would need to automate at least some tasks. We searched the tools that are available to easily create new items and statements, as well as adding statements to existing items. The tools used during the pilot were: OpenRefine, QuickStatements and SourceMD.

We took the data extracted from the CRIS and cleaned it. Based on the selected properties, we deleted columns that contained information that we did not want to use. Then we uploaded the Excel to OpenRefine in order to finish cleaning it. One important step is the process of reconciliation in which

OpenRefine tries to match the fields in the uploaded table with existing Wikidata items. This is essential in order to avoid duplication, and it is quicker than to check manually if an item already exists.

It is also necessary to define a Wikidata schema, which corresponds to the way the data is structured in Wikidata. The data edited in OpenRefine was then exported to QuickStatements format. With QuickStatements it is possible to add and remove statements, labels, descriptions and aliases, as well as create new items. The data is converted into a sequence of commands, which can be pasted into an editor. Running the commands creates or modifies items and statements. We also used SourceMD, that allows for the creation of items corresponding to scholarly articles, using the DOI.

3. Scholia

Scholia is a web-based application that handles bibliographic information through Wikidata. This was made possible with the release of the Wikidata Query Service (WDQS), in 2015, by the Wikimedia Foundation. This service allows users to run queries on the data contained in Wikidata, using SPARQL as the query language. SPARQL is a semantic query language able to retrieve and manipulate data stored in RDF format ("SPARQL," 2019). There are, as of March 2019, 26 different tools (including Scholia) "that allow to query the data in different ways" ("Wikidata:Tools/Query data," 2019).

Scholia is almost entirely built by using WDQS to generate tables, bubble charts, time lines, graphs and other visualizations. This service can display on-the-fly visualizations of profiles for researchers, as well as organizations, journals, publishers, research topics and even individual scientific works through SPARQL-based queries. For example, in an "organization" page we can see lists and graphs that display information on employees and affiliated researchers, co-authors, advisors, recent publications, page production, citations, awards and gender distribution. The "author" page presents information on publications, number of pages, venues, co-authors, topics, associated images, locations, and citations.

Figure 2. Screenshot of Scholia with list of researchers affiliated to NOVA SBE.

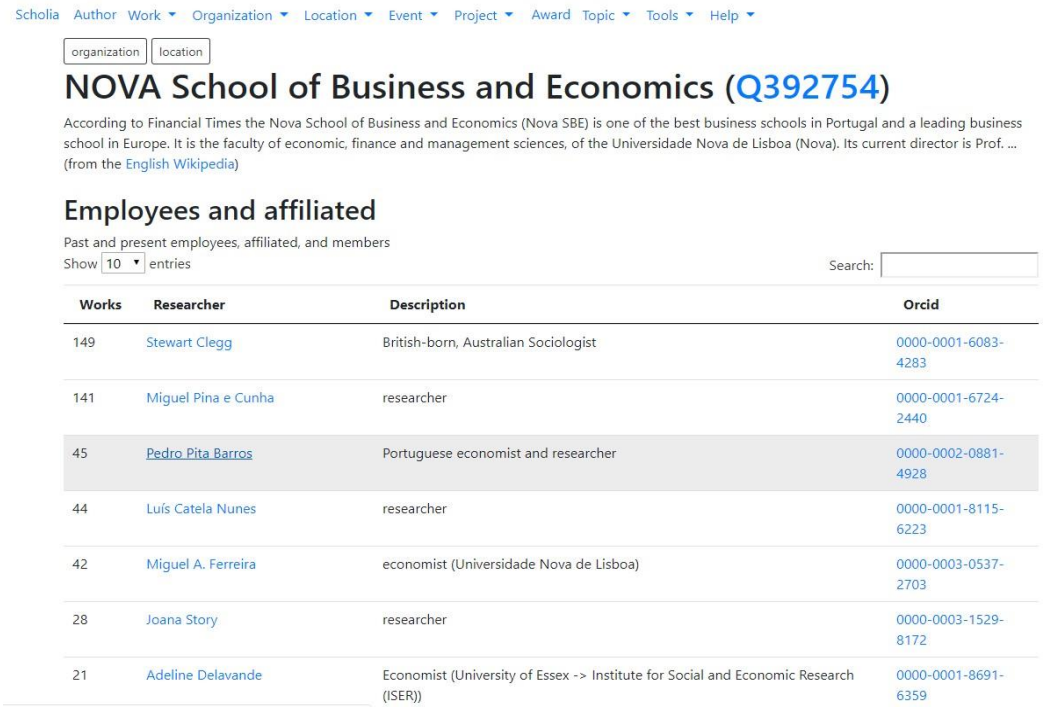


Figure 3. Screenshot of Scholia with co-author graph.

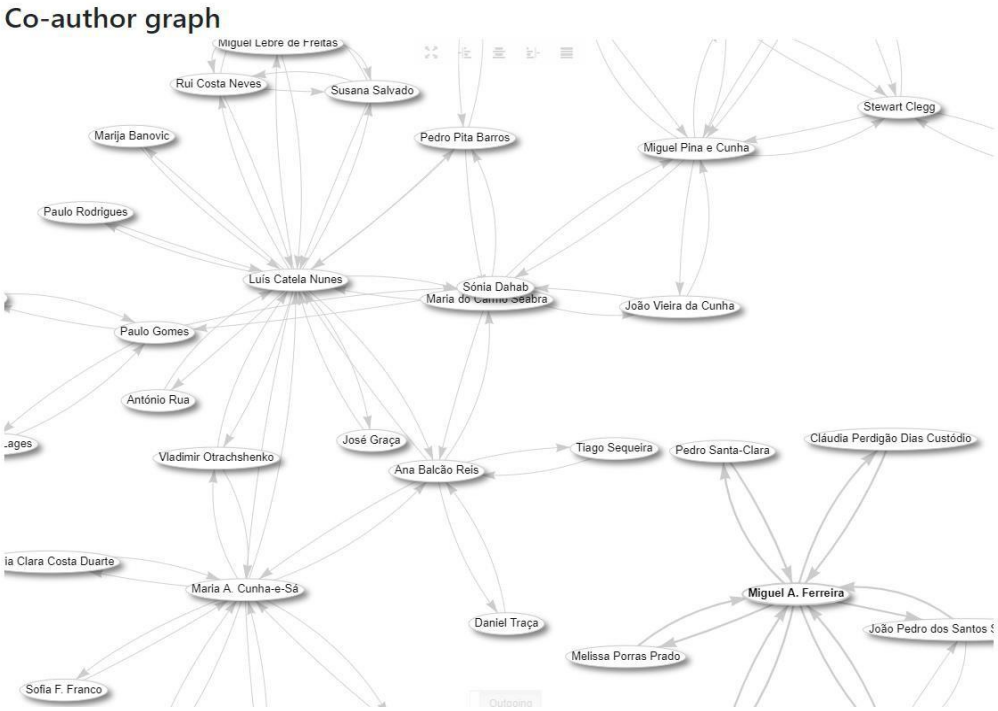
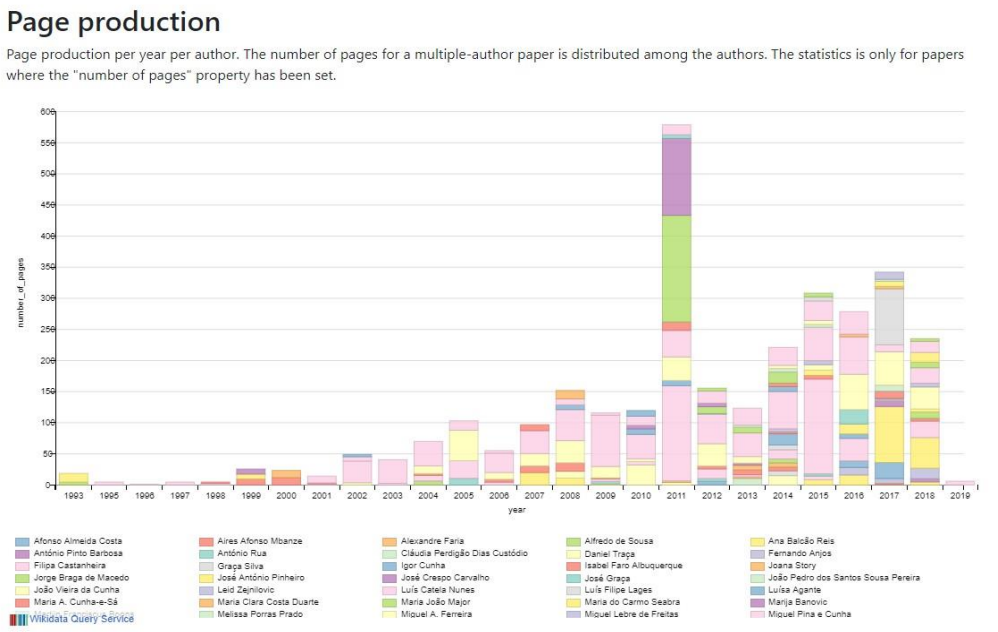


Figure 4. Screenshot of Scholia with page production graph.



4. Next Steps

Our goal is to be able to display complete information on NOVA SBE researchers and their publications in Scholia and try to regularly update it to Wikidata in a semi-automated way. In order to import data in a large scale we will continue to use tools, such as OpenRefine, QuickStatements and SourceMD, to clean metadata and upload them directly to Wikidata. The list of selected properties may be changed

if we consider there are other relevant properties. We also want to contribute data on cited works and advisors.

Another goal is to be able to create synergies with the NOVA SBE Research Office, in order to keep aligned with the School's research strategy. To ensure that the information on Wikidata remains up-to-date and as complete as possible, it will be necessary to ensure that these processes are a part the library routine integrated in the science dissemination strategy. It is also important to be aware of other projects related to scholarly information dissemination using Wikidata

We want to give more visibility to the project, communicating our results inside the NOVA community, as well as other higher education organizations. We consider relevant to engage library professionals and show how the library can play an important role in the dissemination of scholarly information by using semantic web tools.

5. Conclusions

At the end of the pilot, there were 70 NOVA SBE researchers in Wikidata (approximately 35% of the total number of researchers that are affiliated with NOVA SBE). Some profiles are still incomplete, as well as the scientific production of some researchers, while some of them already have all their scientific production on Wikidata. In total, there are already more than 600 scholarly articles and books produced by the School's researchers.

It is possible to create personalized queries using WDQS. The query results can be rendered in multiple forms, for example tables, images, timelines, maps, line charts or graphs, that can then be exported to JSON, Excel and other formats or embedded in websites using HTML iframe elements.

During the pilot we were challenged by the dispersion and lack of documentation. There are some tutorials online that are useful to understand essential concepts and tools, but we often were not able to find simple answers to our questions and necessities. This required a lot of "trial and error" to find good solutions, for example, understanding how to use the available tools to do mass imports.

Data visualization communicates complex information in a more intuitive and meaningful way, and helps to identify patterns and understand difficult concepts. Although all the information used in this pilot is already freely available online, Scholia helps to bring awareness and gain new insights and perspectives about in-house research. Scholia summarizes scholarly information in an attractive way, allowing to build a stronger relationship with data.

By publishing our scholarly data into Linked Open Data systems, like Wikidata, we allow this data to be interlinked and used in SPARQL semantic queries. We can ask complex questions to data using standard web technologies that can be understood both by humans and computers.

At the same time, we think that the mission of the Wikimedia projects is well aligned with the libraries' mission. Wikidata presents some issues, namely the fact that it is far from complete, biased, and citation data is lacking, which reflects on Scholia results. Libraries have the necessary sources to ensure information quality on Wikipedia and Wikidata, and in return they can have their data in open and freely accessible platforms with high impact and visibility on the web. With this article we hope to inspire other organizations to make their own contributions to Wikidata.

Referências Bibliográficas

ARLITSCH, K., SHANKS, J. (2018). Wikipedia and Wikidata Help Search Engines Understand Your Organization: Using Semantic Web Identity to Improve Recognition and Drive Traffic. In

M. PROFFITT (Ed.), *Leveraging Wikipedia: Connecting Communities of Knowledge* (ALA Editions, pp. 159-196). Chicago, IL.

ERXLEBEN, F., GÜNTHER, M., KRÖTZSCH, M., MENDEZ, J., & VRANDEČIĆ, D. (2014). Introducing Wikidata to the Linked Data Web. *The Semantic Web – ISWC 2014*, 50-65. doi:10.1007/978-3-319-11964-9_4.

LEMUS-ROJAS, M., & ODELL, J. D. (2018). Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia. *Journal of Librarianship and Scholarly Communication*, *6* (1), 2272. doi: 10.7710/2162-3309.2272.

LEMUS-ROJAS, M., & PINTSCHER, L. (2018). Wikidata and Libraries: facilitating open knowledge. In M. Proffitt (Ed.), *Leveraging Wikipedia: Connecting Communities of Knowledge* (ALA Editions, pp. 143–158). Chicago, IL.

NIELSEN, F. Å., MIETCHEN, D., & WILLIGHAGEN, E. (2017). Scholia, Scientometrics and Wikidata. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-70407-4_36.

SMITH-YOSHIMURA, K. (2018). Hanging Together. The OCLC Research blog: The rise of Wikidata as a linked data source. Retrieved January 30, 2019, from http://hangingtogether.org/?p=6775.

VRANDEČIĆ, Denny & KRÖTZSCH, Markus. (2014). Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*. 57. 78-85. doi: 10.1145/2629489.

“Wikidata.” 2019. Wikidata.
https://www.wikidata.org/wiki/Q2013.

“Wikidata: Data Access.” 2019. Wikidata.
https://www.wikidata.org/wiki/Wikidata:Data_access.

“Help: Label.” 2019. Wikidata.
https://www.wikidata.org/wiki/Help:Label.

“SPARQL.” 2019.Wikipedia.
https://en.wikipedia.org/wiki/SPARQL.

“Wikidata: Tools/Query data.” 2019. Wikidata.
https://www.wikidata.org/wiki/Wikidata:Tools/Query_data.