Wikidata as a Tool for Mapping Investment in Open Infrastructure

An exploratory study

December 2021

IOI

Wikidata as a Tool for Mapping Investment in Open Infrastructure: An Exploratory Study

https://investinopen.org

Contact:
Anne Britton (2021 IOI Research Fellow)
Project Coordinator, Invest in Open Infrastructure
anne@investinopen.org

Kaitlin Thaney
Executive Director, Invest in Open Infrastructure
kt@investinopen.org

Report author:
Anne Britton (2021 IOI Research Fellow)

Report DOI: 10.5281/zenodo.6036284

Report dated: December 2021

Funding for this research was provided by:

Alfred P. Sloan Foundation

IOI

Executive Summary

This working paper shares the results of research conducted to explore Wikidata's current coverage of the domain of open infrastructure and investment therein. The research question investigates whether Wikidata, a collaboratively edited and multilingual knowledge graph hosted by the Wikimedia Foundation, is a viable prospect for hosting investment flow data for open infrastructure.

At present, Wikidata partially describes the domain. Coordinated efforts to collectively define relevant data categories, relationships, and values, and to align distributed editing will help to improve coverage.

This study was conducted as part of a Research Fellowship with Invest in Open Infrastructure (IOI), and is generously supported by the Alfred P. Sloan Foundation.

We invite feedback and comments directly in this document. Please feel free to add your thoughts via the commenting function. Have questions? Contact us.

Table of Contents:

IOI

Introduction

Wikidata is one of the sister projects of Wikipedia. It is a multilingual knowledge base of structured data, licensed to the public domain. Developed mostly by Wikimedia Deutschland, Wikidata launched in 2012 and has grown to include some 96 million crowdsourced data items today.

Other openly licensed knowledge bases exist. But Wikidata is remarkable in terms of its size and its open, inclusive, multilingual, and profoundly participatory design. This study explores Wikidata as a potential solution to the problem of mapping investment flows of open infrastructure. It addresses the question of whether or not Invest in Open Infrastructure (IOI)-relevant concepts, terms, and relationships are general enough to fit into Wikidata. Should their metadata representation exist separately, for instance in a subject-focused Wikibase devoted to open infrastructure?

The inspiration for the present study sprang from Wikidata work by Katherine Thornton et al. (2017) and from recent surveys of scholarly communication infrastructure and tools by Bianca Kramer and Jeroen Bosman (2015), John Maxwell et al. (2019), Katherine Skinner (2019), David Lewis (2020), SPARC Europe (2020), and COAR/Educopia (2021).

The following section details the researcher's process for surfacing IOI-relevant concepts and terms to map to properties, relationships, and items in Wikidata.

Methods

To answer the question of whether or not investment flows for open infrastructure could be mapped in Wikidata, this study looked at just ten subjects in order to limit the scope of investigation into Wikidata's 96 million data items. The study focused in particular on the subjects of interviews conducted by Yvonne Campfens in 2020 and published in partnership with IOI and SPARC Europe ("10 Key Interviews: Insights into the sustainability of open infrastructure services"): 4TU.ResearchData, arXiv, Code Ocean, Dryad, EDP Sciences, F1000 Research, Figshare, Mendeley, Our Research, and Redalyc.

A series of questions was asked about each subject: Is the subject represented in Wikidata? Is the representation accurate? Well-sourced? Does it include any financial

information? Does it describe investment-- fiscal sponsorship, in-kind support, or volunteer labor?

After these questions were answered, each subject's Wikidata representation was then expanded to deepen linkages within the Wikidata knowledge graph. Facts about investment components, if available from published sources, were added whenever possible.

Facts added to Wikidata items for organizations, products, or services included investment details – such as funder, sponsor – and contextualizing information – director, employee count, founder, inception, industry, instance, legal form, owner, use, and so on.

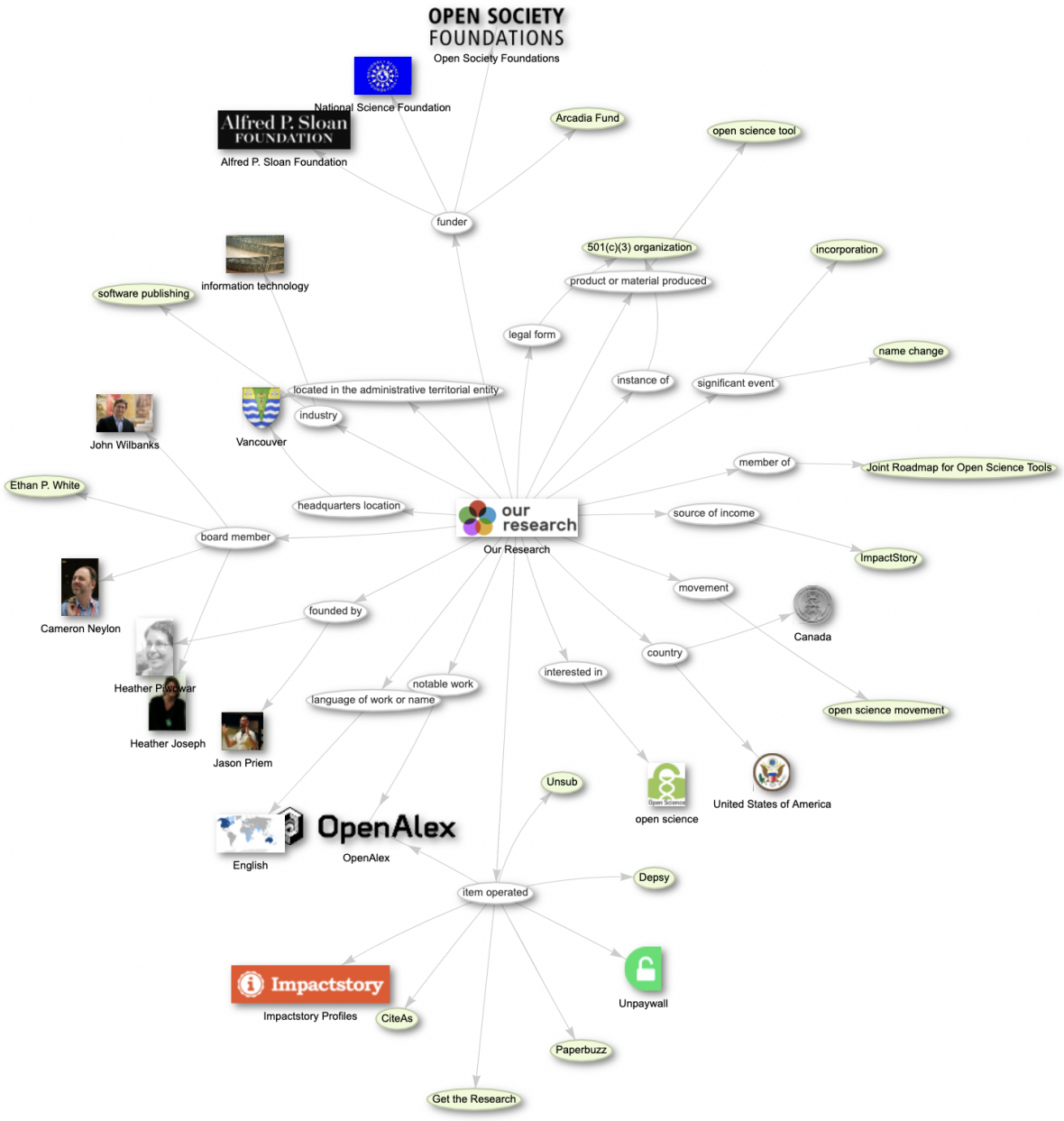
The expansion step was meant to (1) surface additional Wikidata properties (data categories), items (data values), and other internal and external resources useful in fully describing the domain at hand, as well as (2) identify IOI-relevant properties and items currently missing from Wikidata.

Limitations of the above-described method include its bias towards English-language terminology and sources of information; lack of researcher awareness of nuances in technology and investment; and researcher bias towards scholarly communication as practiced in the US.

Example item

A list of the ten example Wikidata items can be found here. For demonstration purposes, one of the ten examples, the US nonprofit Our Research (also known as ImpactStory), will be discussed briefly.

> **Our Research**
> Country: US
> Description: US nonprofit organization
> Scholia visualization: Q16996711
> Edits made to Wikidata item: September-November 2021
> New Wikidata items created: 14
> Ten Key Interviews (2020): An interview with Heather Piwowar, Co-founder, Open Research, Canada



Scholia visualization of Our Research "topic in context" (source, retrieved 22 November 2021)

The graph above visualizes the Wikidata statements (properties, relationships, and items) pulled directly from the Wikidata item for Our Research (Q16996711), using a visualization tool called Scholia. Data scientist Finn Årup Nielsen of the Technical University of Denmark developed the open-source tool in 2016 to create on-the-fly scholarly profiles. Although Our Research is not itself a scholar, Scholia provides useful visualization of the organization's founders, funders, products, and more.

OpenAlex, the latest product created by Our Research, appears in the graph as of November 2021, along with its logo. The same query, had it been run in June 2021, would not have shown OpenAlex, since its Wikidata representation had not yet been created or linked to the Wikidata representation for Our Research. The link was made in July 2021 by an editor called User:Azertus.

In this way, link by link, editors expand and enrich Wikidata items with factual detail. In turn, facts appear in tools like Scholia, or in sidebar infoboxes in Wikipedia articles.

During the present study, several edits were made to the Our Research item, and to its many related items-- its products, founders, board members, and so on.

New items were created to fill gaps in basic description-- such as Crossref Event Data, Q109266549, which is used to run Our Research's Paperbuzz, Q96473538; or Open Science Thesaurus, Q108928644, which describes Our Research's Unpaywall as an "open science tool;" or the US National Science Foundation Early-concept Grants for Exploratory Research, Q109324913, a source of funding.

Findings

In our research, we focused on Wikidata due to its size and its open, inclusive, multilingual, and participatory design. Other openly licensed knowledge bases exist, but we do not explore them in this working paper.

Currently, Wikidata partially and inconsistently describes the domain of open infrastructure. Governance, leadership, and financial and other investment details are generally not present – including board members, grants, in-kind donations, volunteer labor, revenues, assets, and expenses.

An initial overview of the subjects in September 2021 prior to the start of this study can be found here. In general, representation existed, although many statements lacked sources. Occasionally organizations, products, services, and technology appeared as one combined item, rather than distinct, separate entities. Most subjects lacked financial data.

In the course of filling in basic facts on each subject, a **working list of potentially useful IOI-relevant properties and items** was compiled, along with some examples of use.

Examples of English-language terms not found in Wikidata (September-November 2021) as distinct items: academic-led, academy-owned, catalyst grant, commercial vendor, communication infrastructure, community-led, corporate giving, data management service, fiscal host, institutional membership, library subscription, open infrastructure, publishing technology, research-sharing platform, volunteer labor, website visitor.

Discussion

Many editors have worked to represent the scholarly communication ecosystem in Wikidata. Excellent groundwork exists already. However, we face several challenges in mapping *investment flows of open infrastructure* in Wikidata.

First, components of open infrastructure can resist easy **comparison**, or clean mapping to identical data models.

As Nadia Eghbal (2016) and others have described, components of open infrastructure often begin as efforts of individuals, then become named products or services, and then develop an organizational structure to administer the product or service and its staff, funding, maintenance, and so on. Sometimes an organization forms part of a larger organization, such as a university or multinational corporation. Wikidata should express these distinctions between organizations, products, and services.

Fortunately, Wikidata editors in the past few years already have developed lists of core properties for organizations, businesses, and software, working in self-organized, thematic groups known as WikiProjects. When applicable, already-developed lists of core properties should be used to describe aspects of open infrastructure.

Often there exists a clear distinction between the name of a product or service and the organization that oversees it (such as OpenAlex, overseen by Our Research). Sometimes, however, the distinction is not clear to a general reader (as in the case of Figshare, overseen by Figshare LLP, or Dryad repository, overseen by the Dryad nonprofit body).

Second, **published sources** documenting investment flows vary in availability.

Links to trustworthy, published information sources that support statements of fact raise the overall worth of Wikidata as a public knowledge base. In the domain of open infrastructure, many published sources exist that describe technology and use. Fewer sources exist for financial information, however, an observation noted by other IOI researchers.

For example, US tax law requires registered nonprofit organizations to publicly file annual tax returns. It is relatively easy to find financial information for US organizations in this category (such as Dryad or Our Research). A US project such as ArXiv, on the other hand, does not exist as a separate body. It is overseen by Cornell University, which publishes financial information but not necessarily at the granular level itemizing hundreds of projects within the university in a given year. ArXiv itself issues some financial data, but it does not necessarily map cleanly to that of US nonprofit tax forms.

Outside of the US, financial information becomes more difficult to find, depending on one's familiarity with the context within a particular country. Mexico, for instance, requires nonprofit civic organizations to register. In the US, several non-governmental efforts focus on parsing nonprofit tax data for wider consumption by human and machine readers (such as ProPublica's NonProfit Explorer). The present study was not able to find Mexican counterparts.

Many granting agencies provide lists of grant recipients, but specific figures and dates can be hard to find. Privately held commercial enterprises rarely publish financial data. We are hopeful that more data will be made available in the coming years as IOI's work and that of the community grows in advocating for public disclosure.

Third, **investment** in general appears to be undermodelled in Wikidata. More specific properties and items could better represent grants, in-kind donations, volunteer labor, and fiscal sponsorship.

Recommendations

Based on the above-described observations of quality and quantity of linked data in Wikidata, as well as in-depth discussion with IOI staff, brief engagement with conference session attendees (see Appendix 9), and responses to a survey (see Appendix 8), the following actions would benefit IOI's mission:

Alignment

1. Lead or facilitate a community-wide effort to create and use schemas that will generate query results with high utility to IOI stakeholders.

2. Encourage the establishment of a Wikidata WikiProject to coordinate the scope and pace of editing by the IOI community. Models: Informatics/FLOSS; Economics; Companies (and its country-specific guide)

IOI

Training

3. Organize and facilitate at least one training session on Wikidata editing.

4. Organize and facilitate at least one training session on using SPARQL and the Wikidata Query Service.

5. Propose regular virtual meetings. Model: LD4 Wikidata Affinity Group's themed bi-weekly calls with guest presenters. LD4 (Linked Data for Production) is a series of funded exploratory projects among US research libraries.

Definitions

6. Issue and maintain a list of data categories of high interest to IOI, along with detailed definitions, in multiple languages, linked to corresponding Wikidata properties whenever possible.

7. Consider facilitating a community-developed glossary. Model: process that produced the Framework for Open and Reproducible Research Training glossary.

8. Consider tracking emerging consensus on vocabulary that describes the domain of open infrastructure and its funding. Model: folksonomy accumulated via TagTeam software for the Open Access Tracking Project.

9. Consider facilitating a community discussion on ontology. Resources: Open Science Thesaurus; Wikidata:WikiProject Ontology; Noy and McGuinness's 2001 Ontology Development 101.

10. Create schematic diagrams of organizational hierarchies and dependencies. Example: Holtzbrinck.

11. Create schematic diagrams for products, services, technologies. Example: Redalyc platform layers.

12. Create schematic diagrams for funding. Example: Open Collective interactive graph.

13. Create or collect visuals to help disambiguate entities with identical or near-identical names. Example: Jupyter notebook.

Gaps

14. Organize and facilitate discussion on IOI-relevant data values and categories missing from Wikidata; strategize ways to fill gaps.

15. Coordinate collective effort to fill Wikidata with IOI-relevant content on open infrastructure.

16. Establish relationship with Wikidata as an institutional data donor, so that IOI can add its own research data as a trusted neutral party, without conflict of interest.

17. Expand coverage of open infrastructure and its funding in main subject statements in bibliographic items. Example: article about CZI and bioRxiv.

18. Create property proposal for number of downloads. See recent proposal. Example: arXiv statistics.

19. Create property proposal for number of website visitors. Example: Redalyc "service to 72.000 unique users per day."

20. Consider property proposal for acceptable use policy. Example: Mind the Gap.

Other

21. Perform "nudge editing." Example: for a funder that does not provide an API to its funding data, input "unknown value" in API endpoint statement. (Phrase inspired by Cass Sunstein (2014)).

22. Monitor status of English-language Wikipedia community discussion on using Wikidata for infoboxes. Already in use in some non-English Wikipedias. Example: Redalyc in Spanish Wikipedia, EDP Sciences in Russian Wikipedia.

23. In the past few years, Wikimedia project contributors have developed automated tools and gadgets for curation and visualization, among other functions. Tools useful to the present study included Scholia, Recoin ("Relative Completeness Indicator"), SQID, Flickr2Commons, and Thomas Steiner's Wikidata tool for Google Sheets. As representation of the domain of open infrastructure deepens in Wikidata, curation tools such as inteGraality and Listeria may prove useful for aligning editing activity by diverse contributors.

IOI

Acknowledgements

Thanks to IOI for the opportunity to conduct this study, and to Asura Enkhbayar, Kaitlin Thaney, Richard Dunks, and Teri Wanderi for their helpful comments and questions. Special thanks to David Lewis for providing a text version of his 2020 Bibliographic Scan, and to Daniel Mietchen for advice on WikiProjects. Thanks also to the producers of information sources supporting statements of fact: ProPublica's Nonprofit Explorer, Open Corporates, Internet Archive's Wayback Machine; and to Wikidata developers and editors.

Glossary

Data category. "Class of data items that are closely related from a formal or semantic point of view." (source)

Folksonomy. "The emergent labeling of lots of things by people in a social context." (source)

Item. "In Wikidata, items are used to represent all the things in human knowledge, including topics, concepts, and objects. For example, the '1988 Summer Olympics', 'love', 'Elvis Presley', and 'gorilla' are all items in Wikidata." (source)

Ontology. According to Tom Gruber, "in the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application." (source)

Open infrastructure. "By 'infrastructure' we mean the sets of services, protocols, standards and software that the academic ecosystem needs in order to perform its functions throughout the research lifecycle — from the earliest phases of research, collaboration and experimentation through data collection and storage, data organization, data analysis and computation, authorship, submission, review and annotation, copyediting, publishing, archiving, citation, discovery and more. 'Open infrastructure' is the narrower set of services, protocols, standards and software that can empower communities to collectively build the systems and infrastructures that deliver new improved collective benefits without restrictions, and for a healthy global interrelated infrastructure system." (source)

Property. In Wikidata, "a property describes the data value of a statement and can be thought of as a category of data, for example 'color' for the data value 'blue'." (source)

Scholia. In Wikidata, "Scholia is a service that creates visual scholarly profiles for topics, people, organizations, species, chemicals, etc using bibliographic and other information." (source)

SPARQL. "A semantic query language for databases." (source)

Statement. In Wikidata, "a statement consists of a property-value pair, for example, 'location: Germany.'" (source)

Value. "In computer science, a value is the representation of some entity that can be manipulated by a program." (source)

Wikibase. "Wikibase is a set of MediaWiki extensions for working with versioned semi-structured data in a central repository.... Wikibase was developed for and is used by Wikidata." (source)

WikiProject. "A group of contributors who want to work together as a team to improve Wikidata. These groups often focus on a specific topic area (for example, astronomy) or a specific kind of task." (source)

Appendix

1. List of assorted properties, items, and resources relevant to open infrastructure and its funding.
2. List of assorted sample queries relevant to open infrastructure and its funding.
3. List of people interested in attending Wikidata training and/or coordinated IOI-relevant editing. (IOI internal use only)
4. Draft introductory guide to IOI-relevant Wikidata editing.
5. Example Wikidata items, before and after IOI study. Based on 2020 published interviews produced by IOI. (Alternate format here).
6. List of new Wikidata items created during the IOI study.

7. Image files uploaded to Wikimedia Commons, October-November 2021.

8. Survey to gauge IOI community engagement with Wikidata, and introductory blog post, November 2021.

9. Presentation slides, Open Source Publishing Tools session, Open Publishing Fest, 16 November 2021.