Describing datasets in Wikidata

Denny Vrandečić
orcid.org/0000-0002-9593-2294
Google
San Francisco, CA, USA
vrandecic@google.com

***Abstract*—We propose to use Wikidata to provide metadata for datasets when the traditional approach via Schema.org is not feasible. We describe and discuss the proposal, and believe that the process described in this paper can help with increasing findability and accessibility of certain datasets.**

***Index Terms*—dataset findability, Schema.org, Wikidata**

I. Introduction

Describing datasets with Schema.org [1] provides the means to provide crucial metadata about datasets which makes the datasets findable and accessible. A dataset publisher or a dataset catalog provider can create a Website describing a dataset and annotate it with the Schema.org metadata necessary to allow for the effective indexing of the dataset so it can be surfaced in a dataset search engine such as Google Dataset Search [2], [3].

One hurdle with that process is that the publisher of the dataset needs to markup their site with Schema.org annotations. This might not always be feasible or easy enough, since the publisher may not have the technical or organizational support to change the HTML markup of their sites accordingly, or because the dataset was published before Schema.org annotations became widely used, and no one is working on the dataset anymore, or because the publishers simply don't think that it is a good use of their time.

In this paper we propose to use Wikidata [4] in such cases to provide metadata about datasets. We discuss how this approach would work—both from the side of the metadata provider (Section II) and the metadata ingester (Section III)—and the advantages (Section IV) and disadvantages (Section V) of using Wikidata when providing Schema.org annotations is unfeasible.

We believe that the process proposed in this paper can help with making more datasets findable and accessible, and solve current border cases of metadata about important datasets not being available.

II. Metadata provider

We propose to use Wikidata to provide metadata about a dataset. Currently, Wikidata already describes more than 25,000 datasets (of which about a third have articles on the English Wikipedia).

Wikidata also already has properties that match to a number of the properties defined for Schema.org Datasets. Table I shows a few of those (note that the mapping to Schema.org is already part of Wikidata). All required properties for `schema:Dataset` are already mapped (bolded in the table). Properties that are useful and are currently still missing can be proposed through the Wikidata property proposal process.[1]

The table also lists how many of the datasets have at least one statement with one of the given mapped properties.

One issue we found is that Google Dataset Search asks for description to be comprehensive, comprising several sentences if not paragraphs. In Wikidata, descriptions are meant to be short and used for disambiguation. One way to solve this is to pull in either the articles in Wikipedia, or the official Website, and use those as the description when indexing.

The proposal is to create a Wikidata item for the dataset we want to provide metadata for, and then add the relevant statements to it in Wikidata.

III. Metadata ingester

In addition to crawling the relevant Schema.org annotations from the Web, we suggest that a metadata ingester also crawls and updates the relevant dataset descriptions from Wikidata.

An interesting question will be to compare the metadata coming from the different sources, but since datasets are already reconciled this is not a new problem.

Since Wikidata is large and updates very frequently, it would be rather costly to mirror all of Wikidata just for the sake of keeping the dataset metadata up to date. We suggest to use SPARQL queries [5] in order to find all items of interest and their respective data, and then run these SPARQL queries periodically to update the data. This allows to use this additional metadata for a rather low cost for all participants.

[1] `https://www.wikidata.org/wiki/Wikidata:Property_proposal`

TABLE I
Some relevant properties and the number of datasets using the property in Wikidata

Wikidata	Schema.org	count
label	**name**	25,533
description	**description**	21,959
P921	about	1,044
P50	author	799
P2860	citation	19*
P767	contributor	8
P580	dateCreated	50
P577	datePublished	9,703

* Note that for citations, we counted how many citations the datasets have in Wikidata, and not how often the datasets cite something.

IV. Advantages

There are several advantages to using this process:

- It might be much easier, and sometimes the only feasible way, to provide the required metadata to describe a dataset. As said, a dataset creator might not have the capability anymore to update the Website describing their dataset (maybe they changed affiliations, maybe the original affiliation closed down and only an archival version of the dataset is available, etc.), and in this case they can use Wikidata to provide the necessary metadata to improve findability.
- It might be that the original dataset providers are not available anymore, or do not have the time to provide the metadata, or do not see the benefit in doing so. Wikidata allows for metadata provision to be decentralized to the community, i.e. anyone could provide the necessary metadata, not just the dataset publisher.
- By relying on a widely accepted, actively community edited knowledge base the data is expected to fulfill certain minimal standards, and also to be able to widely avoid spam entries.
- By having a single repository as an additional source for metadata about datasets we reduce the cost for crawling and integrating the metadata considerably. In theory we could imagine a more generic linked data approach towards solving the issue at hand, following the AAA approach (Anyone can say Anything about Anything [6]). This is suitable for many other domains, but seems to incur excessive costs on the metadata ingester to collect the data and bring it to a usable state. By using Wikidata as a single and community editable platform to provide the metadata we provide a cost-efficient alternative.

V. Disadvantages

There are also clear disadvantages to the proposed approach:

- The metadata is provided on a third party Website, where it can be edited by anyone, and must thus receive additional scrutiny. Publishing the metadata on the Website of the dataset provider will confer the authority of the dataset provider to the metadata, whereas the same is not true for Wikidata.[2]
- Wikidata is a community maintained resource, and it is important to not overload the community with too many datasets that might be deemed non-notable. If someone is planning to publish thousands or millions of datasets, the time would be far better invested in setting up a process to also publish the metadata describing these datasets instead of setting up a bridge and publish the data in Wikidata.
- Any ingester, such as Google Dataset Search, has to implement an entire second pipeline to gather, validate, and keep metadata up to date. This will incur a continuous cost on the whole product.

[2]Wikidata's Signed Statements project is going to alleviate these issues partially, but they are not implemented yet and will require additional work to use them. See `https://phabricator.wikimedia.org/T138708`

VI. Summary

We think that using Wikidata as an additional approach towards publishing metadata for datasets can improve the findability and accessibility of certain datasets. We think that it is preferable for dataset publishers to provide their metadata themselves, using Schema.org, but Wikidata can provide an option in case this is not feasible.

If this proposal is accepted by the relevant communities, both the relevant scientific communities and the Wikidata community, the next steps would be to either identify or create ways to express all properties that are needed for describing datasets in Wikidata, to extend metadata ingesters with enriching their data with Wikidata data, and to encourage tools that are addressing the findability of datasets to use this as an additional source of metadata. Then all the pieces will be in place to provide a secondary process to describe datasets using Wikidata, in order to make them more findable and accessible.

Acknowledgments

Thanks to Krzysztof J. Gorgolewski, Natasha Noy, Dan Brickley, and the anonymous reviewer for their valuable comments on the idea and the paper.

References

[1] R. V. Guha, D. Brickley, and S. Macbeth, "Schema. org: evolution of structured data on the web," *Communications of the ACM*, vol. 59, no. 2, pp. 44–51, 2016.

[2] D. Brickley, M. Burgess, and N. Noy, "Google Dataset Search: Building a search engine for datasets in an open web ecosystem," in *The World Wide Web Conference*. ACM, 2019, pp. 1365–1375.

[3] A. Canino, "Deconstructing Google Dataset Search," *Public Services Quarterly*, vol. 15, no. 3, pp. 248–255, 2019. [Online]. Available: https://doi.org/10.1080/15228959.2019.1621793

[4] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014. [Online]. Available: http://doi.acm.org/10.1145/2629489

[5] C. B. Aranda, O. Corby, S. Das, L. Feigenbaum, P. Gearon, B. Glimm, S. Harris, S. Hawke, I. Herman, N. Humfrey, N. Michaelis, C. Ogbuji, M. Perry, A. Passant, A. Polleres, E. Prud'hommeaux, A. Seaborne, and G. T. Williams. (2013) SPARQL 1.1 overview. W3C Recommendation. W3C. [Online]. Available: https://www.w3.org/TR/sparql11-overview/

[6] G. Klyne and J. J. Carroll. (2004) Resource Description Framework (RDF): Concepts and abstract syntax. section 2.2.6. W3C Recommendation. W3C. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/