BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59211
doi: 10.3897/biss.4.59211

OPEN ACCESS

Conference Abstract

How do you Develop a Data Standard? Wikibase might be the Solution…

Maarten Trekels[‡], Matt Woodburn[§], Deborah L Paul[|], Sharon Grant[¶], Kate Webbink[¶], Janeen Jones[¶], Quentin Groom[‡]

‡ Meise Botanic Garden, Meise, Belgium
§ Natural History Museum, London, United Kingdom
| Florida State University, Tallahassee, United States of America
¶ The Field Museum of Natural History, Chicago, United States of America

Corresponding author: Maarten Trekels (maarten.trekels@plantentuinmeise.be)

Received: 01 Oct 2020 | Published: 09 Oct 2020

Citation: Trekels M, Woodburn M, Paul DL, Grant S, Webbink K, Jones J, Groom Q (2020) How do you Develop a Data Standard? Wikibase might be the Solution…. Biodiversity Information Science and Standards 4: e59211. https://doi.org/10.3897/biss.4.59211

Abstract

Data standards allow us to aggregate, compare, compute and communicate data from a wide variety of origins. However, for historical reasons, data are most likely to be stored in many different formats and conform to different models. Every data set might contain a huge amount of information, but it becomes tremendously difficult to compare them without a common way to represent the data. That is when standards development jumps in.

Developing a standard is a formidable process, often involving many stakeholders. Typically the initial blueprint of a standard is created by a limited number of people who have a clear view of their use cases. However, as development continues, additional stakeholders participate in the process. As a result, conflicting opinions and interests will influence the development of the standard. Compromises need to be made and the standard might look very different from the initial concept.

In order to address the needs of the community, a high level of engagement in the development process is encouraged. However, this does not necessarily increase the usability of the standard. To mitigate this, there is a need to test the standard during the early stages of development. In order to facilitate this, we explored the use of Wikibase to create an initial implementation of the standard.

© Trekels M et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Wikibase is the underlying technology that drives Wikidata. The software is open-source and can be customized for creating collaborative knowledge bases. In addition to containing an RDF (Resource Description Framework) triple store under the hood, it provides users with an easy-to-use graphical user interface (see Fig. 1). This facilitates the use of an implementation of a standard by non-technical users. The Wikibase remains fully flexible in the way data are represented and no data model is enforced. This allows users to map their data onto the standard without any restrictions. Retrieving information from RDF data can be done through the SPARQL query language (W3C 2020). The software package has also a built-in SPARQL endpoint, allowing users to extract the relevant information:

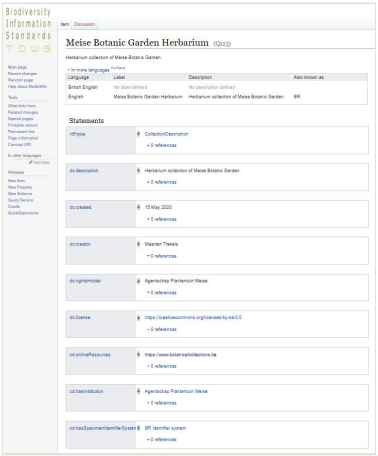

Figure 1.

A screenshot of the description of an herbarium from the test instance of the Wikibase.

- Does the standard cover all use cases envisioned?
- Are parts of the standard underdeveloped?
- Are the controlled vocabularies sufficient to describe the data?

This strategy was applied during the development of the TDWG Collection Description standard. After completing a rough version of the standard, the different terms that were defined in the first version were transferred to a Wikibase instance running on WBStack (Addshore 2020). Initially, collection data were entered manually, which revealed several issues. The Wikibase allowed us to easily define controlled vocabularies and expand them as needed. The feedback reported from users then flowed back to the further development of the standard. Currently we envisage creating automated scripts that will import data en masse from collections. Using the SPARQL query interface, it will then be straightforward to ensure that data can be extracted from the Wikibase to support the envisaged use cases.

Keywords

TDWG, collections description, data standard, standard development

Presenting author

Maarten Trekels

Presented at

TDWG 2020

Funding program

H2020 project Synthesys+ under grant agreement 823827

References

- Addshore (2020) WBStack. https://www.wbstack.com/. Accessed on: 2020-8-07.
- W3C (2020) SPARQL Query Language for RDF. https://www.w3.org/TR/rdf-sparql-query/. Accessed on: 2020-8-07.