A Data Standard for Dynamic Collection Descriptions

Matt Woodburn‡, Gabriele Droege§, Sharon Grant|, Quentin Groom¶, Janeen Jones|, Maarten Trekels¶, Sarah Vincent‡, Kate Webbink|

‡ Natural History Museum, London, United Kingdom
§ Botanic Garden and Botanical Museum Berlin, Berlin, Germany
| Field Museum of Natural History, Chicago, United States of America
¶ Meise Botanic Garden, Meise, Belgium

Corresponding author: Matt Woodburn (m.woodburn@nhm.ac.uk)

Abstract

The utopian vision is of a future where a digital representation of each object in our collections is accessible through the internet and sustainably linked to other digital resources. This is a long term goal however, and in the meantime there is an urgent need to share data about our collections at a higher level with a range of stakeholders (Woodburn et al. 2020). To sustainably achieve this, and to aggregate this information across all natural science collections, the data need to be standardised (Johnston and Robinson 2002).

To this end, the Biodiversity Information Standards (TDWG) Collection Descriptions (CD) Interest Group has developed a data standard for describing collections, which is approaching formal review for ratification as a new TDWG standard. It proposes 20 classes (Suppl. material 1) and over 100 properties that can be used to describe, categorise, quantify, link and track digital representations of natural science collections, from high-level approximations to detailed breakdowns depending on the purpose of a particular implementation.

The wide range of use cases identified for representing collection description data means that a flexible approach to the standard and the underlying modelling concepts is essential. These are centered around the 'ObjectGroup' (Fig. 1), a class that may represent any group (of any size) of physical collection objects, which have one or more common characteristics. This generic definition of the 'collection' in 'collection descriptions' is an important factor in making the standard flexible enough to support the breadth of use cases.

For any use case or implementation, only a subset of classes and properties within the standard are likely to be relevant. In some cases, this subset may have little overlap with those selected for other use cases. This additional need for flexibility means that very few classes and properties, representing the core concepts, are proposed to be mandatory.

Metrics, facts and narratives are represented in a normalised structure using an extended MeasurementOrFact class, so that these can be user-defined rather than constrained to a set identified by the standard. Finally, rather than a rigid underlying data model as part of the normative standard, documentation will be developed to provide guidance on how the classes in the standard may be related and quantified according to relational, dimensional and graph-like models.

So, in summary, the standard has, by design, been made flexible enough to be used in a number of different ways. The corresponding risk is that it could be used in ways that may not deliver what is needed in terms of outputs, manageability and interoperability with other resources of collection-level or object-level data. To mitigate this, it is key for any new implementer of the standard to establish how it should be used in that particular instance, and define any necessary constraints within the wider scope of the standard and model. This is the concept of the 'collection description scheme,' a profile that defines elements such as:

- which classes and properties should be included, which should be mandatory, and which should be repeatable;
- which controlled vocabularies and hierarchies should be used to make the data interoperable;
- how the collections should be broken down into individual ObjectGroups and interlinked, and
- how the various classes should be related to each other.

Various factors might influence these decisions, including the types of information that are relevant to the use case, whether quantitative metrics need to be captured and aggregated across collection descriptions, and how many resources can be dedicated to amassing and maintaining the data.

This process has particular relevance to the Distributed System of Scientific Collections (DiSSCo) consortium, the design of which incorporates use cases for storing, interlinking and reporting on the collections of its member institutions. These include helping users of the European Loans and Visits System (ELViS) (Islam 2020) to discover specimens for physical and digital loans by providing descriptions and breakdowns of the collections of holding institutions, and monitoring digitisation progress across European collections through a dynamic Collections Digitisation Dashboard. In addition, DiSSCo will be part of a global collections data ecosystem requiring interoperation with other infrastructures such as the GBIF (Global Biodiversity Information Facility) Registry of Scientific Collections, the CETAF (Consortium of European Taxonomic Facilities) Registry of Collections and Index Herbariorum.

In this presentation, we will introduce the draft standard and discuss the process of defining new collection description schemes using the standard and data model, and focus on DiSSCo requirements as examples of real-world collection descriptions use cases.

Keywords

collection descriptions, TDWG, data standards, biodiversity, geodiversity, natural sciences, DiSSCo

Presenting author

Matt Woodburn

Presented at

TDWG 2021

Acknowledgements

Many thanks to all the interest and task group members contributing to this work.

Funding program

Support from COST (European Cooperation in Science and Technology) as part of the Mobilise Action CA17106 on Mobilising Data, Experts and Policies in Scientific Collections; and SYNTHESYS+ a Research and Innovation action funded under H2020-EU.1.4.1.2. Grant agreement ID: 823827.

Conflicts of interest

References

- Islam S (2020) European Loans and Visits System (ELViS) as a Use Case for a Collection Descriptions Standard. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59253
- Johnston P, Robinson B (2002) Collections and Collection Description. URL: http://www.ukoln.ac.uk/cd-focus/briefings/bp1/bp1.pdf
- Woodburn M, Paul DL, Addink W, Baskauf SJ, Blum S, Chapman C, Grant S, Groom Q, Jones J, Petersen M, Raes N, Smith D, Tilley L, Trekels M, Trizna M, Ulate W, Vincent S, Walls R, Webbink K, Zermoglio P (2020) Unity in Variety: Developing a collection description standard by consensus. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59233

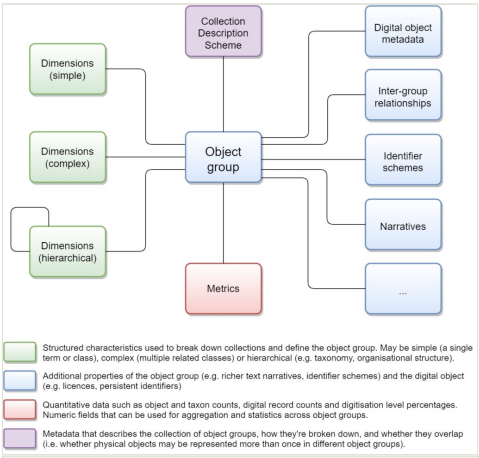

Figure 1.

A simplified representation of the data model.

Supplementary material

Suppl. material 1: Provisional list of classes.

Authors: Gabi Dröge, Sharon Grant, Quentin Groom, Janeen Jones, Maarten Trekels, Sarah Vincent, Kate Webbink, Matt Woodburn and other contributors to the TDWG Collection Descriptions Data Standard Task Group

Data type: data standard class definitions

Brief description: A list of the proposed classes, with associated definitions, in the standard for collection descriptions. A number of classes have been borrowed from Darwin Core rather than defined anew, as indicated in the BorrowedFrom field. In these cases, the definition shown here may have minor modifications to better relate it to the collection descriptions context.

Download file (2.56 kb)