BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37201
doi: 10.3897/biss.3.37201

OPEN ACCESS

Conference Abstract

Exposing the Dark Data of Undigitized Collections: A TDWG global standard for collection descriptions

Matt Woodburn[‡], Deborah L Paul[§], William Ulate[|], Niels Raes[¶]

‡ Natural History Museum, London, United Kingdom
§ Florida State University, Tallahassee, United States of America
| Missouri Botanical Garden, Saint Louis, United States of America
¶ Naturalis Biodeiversity Center, Leiden, Netherlands

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

Received: 12 Jun 2019 | Published: 19 Jun 2019

Citation: Woodburn M, Paul D, Ulate W, Raes N (2019) Exposing the Dark Data of Undigitized Collections: A TDWG global standard for collection descriptions. Biodiversity Information Science and Standards 3: e37201. https://doi.org/10.3897/biss.3.37201

Abstract

Aggregating content of museum and scientific collections worldwide offers us the opportunity to realize a virtual museum of our planet and the life upon it through space and time. By mapping specimen-level data records to standards and publishing this information, an increasing number of collections contribute to a digitally accessible wealth of knowledge. Visualizing these digital records by parameters such as collection type and geographic origin, helps collections and institutions to better understand their digital holdings and compare them to other such collections, as well as enabling researchers to find specimens and specimen data quickly (Singer et al. 2018).

At the higher level of collections, related people and their activities, and especially the great majority of material that is yet to be digitised, we know much less. Many collections hold material not yet digitally discoverable in any form. For those that do publish collection-level data, it is commonly text-based data without the Global Unique Identifiers (GUIDs) or the controlled vocabularies that would support quantitative collection metrics and aid discovery of related expertise and publications. To best understand and plan for our world's bio- and geodiversity represented in collections, we need standardised, quantitative collections-level metadata. Various groups planet-wide are actively developing tools to capture this much-needed metadata, including information about the backlog, and more

detailed information about institutions and their activities (e.g. staffing, space, species-level inventories, geographic and taxonomic expertise, and related publications) (Smith et al. 2018).

The Biodiversity Information Standards organization (TDWG) Collection Descriptions (CD) Data Standard Task Group aims to provide a data standard for describing natural scientific collections, which enables the ability to provide:

1. automated metrics, using standardised collection descriptions and/or data derived from specimen datasets (e.g., counts of specimens) and
2. a global registry of physical collections (either digitised or non-digitised).

The group will also produce a data model to underpin the new standard, and provide guidance and reference implementations for the practical use of the standard in institutional and collaborative data infrastructures.

Our task group includes members from a myriad of groups with a stake in mobilizing such data at local, regional, domain-specific and global levels. With such a standard adopted, it will be possible to effectively share data across different community resources. So far, we have carried out landscape analyses of existing collection description frameworks, and amassed a portfolio of use cases from the group as well as from a range of other sources, including the Collection Descriptions Dashboard working group of ICEDIG ("Innovation and consolidation for large scale digitisation of natural heritage"), iDigBio (Integrated Digitized Biocollections), Smithsonian, Index Herbariorum, the Field Museum, GBIF (Global Biodiversity Information Facility), GRBio (Global Registry of Biodiversity Repositories) and fishfindR.net. These were used to develop a draft data model, and between them inform the first iteration of CD draft data standard.

A variety of challenges present themselves in developing this standard. Some relate to the standard development process itself, such as identifying (often learning) effective tools and methods for collaborative working and communication across globally distributed volunteers. Others concern the scope and gaining consensus from stakeholders, across a wide range of disciplines, while maintaining achievable goals. Further challenges arise from the requirement to develop a data model and standard that support such a variety of use cases and priorities, while retaining interoperability and manageability of the data.

We will present some of these challenges and methods for addressing them, and summarise the progress and draft outputs of the group so far. We will also discuss the vision of how the new standard may be adopted and its potential impact on collections discoverability across the natural science collections community.

Keywords

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

Presenting author

Matt Woodburn

Presented at

Biodiversity_Next 2019

Acknowledgements

We recognize here the entire list of interest and task group members contributing to this work.

References

- Singer R, Love K, Page L (2018) A survey of digitized data from U.S. fish collections in the iDigBio data aggregator. PLOS ONE 13 (12). https://doi.org/10.1371/journal.pone.0207636
- Smith V, Paul D, Woodburn M, Grant S, Singer R, Love K (2018) Shining a New Light on the World's Collections. www.idigbio.org/content/shining-new-light-world%E2%80%99s-collections. Accessed on: 2019-4-05.