Biodiversity Information Science and Standards 8: e135228
doi: 10.3897/biss.8.135228

Conference Abstract

Simple Models, Complex Vocabularies: Developing Controlled Vocabularies for an Interdisciplinary Collection Management System in RECODE

Jutta Buschbom[‡,§], Ben Collier[‡], Matt Woodburn[‡], Sarah Vincent[‡], Elaine Tsai[‡], Kirstie Toth[‡], Marla Spencer[‡], David Smith[‡], Mike Sadka[‡], Tzy-Ting Hsu[‡], Brad Hunn[‡], Josh Humphries[‡], Itan Grinberg[‡], Lucy Ellis[‡], Steen Dupont[‡]

‡ Natural History Museum, London, United Kingdom
§ Statistical Genetics, Ahrensburg, Germany

Corresponding author: Jutta Buschbom (jutta.buschbom@nhm.ac.uk)

Received: 21 Aug 2024 | Published: 22 Aug 2024

Citation: Buschbom J, Collier B, Woodburn M, Vincent S, Tsai E, Toth K, Spencer M, Smith D, Sadka M, Hsu T-T, Hunn B, Humphries J, Grinberg I, Ellis L, Dupont S (2024) Simple Models, Complex Vocabularies: Developing Controlled Vocabularies for an Interdisciplinary Collection Management System in RECODE. Biodiversity Information Science and Standards 8: e135228. https://doi.org/10.3897/biss.8.135228

Abstract

Situated at the intersection of distinct stakeholder communities and their objectives, collection management systems (CMS) need to integrate and mediate a wide range of demands to provide functionality, user experience, and data fit for purpose. While metadata standards, (e.g., Biodiversity Information Standards (TDWG) Darwin Core (Darwin Core Task Group 2009) and its Latimer Core (Grant et al. 2024), and Pinian Core (Plinian Core Task Group 2021) extensions) and ontologies, (e.g., the World Wide Web Consortium (W3C) Provenance Ontology (Lebo et al. 2013) or the W3C Open Digital Rights Language (Iannella et al. 2018)) provide guidance for structuring data resources and workflows, controlled vocabularies standardize and harmonize the data content in those structures.

Controlled vocabularies contribute to differentiating dimensions of information present in metadata concepts and allow comprehensive, information-rich descriptions of reality by aiming to provide well-defined terms that can be clearly understood. Harmonized across scientific and applied disciplines as well as distributed data infrastructures, they

contribute to data interoperability, findability, and reusability, and thus to the basis for data sharing and the automation of work processes.

Instead of introducing challenges for users, the presentation of context-specific subsets of terms for manual selection as well as automation of context-deducible entries can improve user experiences, work environment efficiency, and (meta)data comprehensiveness. This shifts infrastructure development to an additional layer of rules and constraints (policies) that determine interface dynamics and data validation.

Setting these theoretical considerations to the test of practice, we are sharing our experiences and insights gained during the development and implementation of the new collection management system by the RECODE (Rethinking Collections Data Ecosystems) program at the Natural History Museum, London. Controlled vocabularies and their terms constitute a major component in the CMS data model. They present challenges due to their context-specificity and hierarchical nature, for which solutions need to be found.

Daily work with controlled vocabularies requires extensive documentation with functionality for creating and tracking provenance, relationships, and mappings, as well as for versioning. There is a need for open, shared repositories and work environments that foster the versatile, user-driven development of terminologies, ontologies, mappings, and digital policies.

Keywords

data quality, applicability, data harmonization, data sharing, rules and constraints layer

Presenting author

Jutta Buschbom

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Darwin Core Task Group (2009) Darwin Core. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/450

- Grant S, Jones J, Webbink K, Woodburn M, Vincent S, Trekels M, Buschbom J, Groom Q, Norton B, Sanderson R, Engelbrecht I, Baskauf SJ, Addink W, Bloom D, Breugelmans L, Chapman C, Dröge G, Grosjean M, Hahn A, Krimmel E, Paul D, Raes N, Robertson T, Ulate W, (TDWG) BIS (2024) TDWG Latimer Core (LtC) Standard. Biodiversity Information Standards (TDWG). URL: https://tdwg.github.io/ltc/
- Iannella R, Steidl M, Myles S, Rodríguez-Doncel V (2018) ODRL Vocabulary & Expression 2.2, 15 February 2018, W3C Recommendation. World Wide Web Consortium (W3C). URL: https://www.w3.org/TR/odrl-vocab/
- Lebo T, Sahoo S, McGuinness D (Eds) (2013) PROV-O: The PROV Ontology, 30 April 2013, W3C Recommendation. World Wide Web Consortium (W3C). URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/
- Plinian Core Task Group (2021) Plinian Core, a Species-level Data Specification. Biodiversity Information Standards (TDWG). URL: https://github.com/tdwg/PlinianCore