BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e141097
doi: 10.3897/biss.8.141097

OPEN ACCESS

Conference Abstract

Pioneering Museum Data Transformation: Unifying Legacy Systems and Enhancing Data Integrity

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

‡ Natural History Museum, London, United Kingdom

Corresponding author: Marla Spencer (marla.spencer@nhm.ac.uk), Sarah Vincent (s.vincent@nhm.ac.uk), Steen Dupont (steen.dupont@nhm.ac.uk)

Received: 05 Nov 2024 | Published: 06 Nov 2024

Citation: Spencer M, Vincent S, Smith D, Buschbom J, Collier B, Ellis L, Grinberg I, Hsu T-T, Hunn B, Humphries J, Sadka M, Tsai E, Toth K, Woodburn M, Dupont S (2024) Pioneering Museum Data Transformation: Unifying Legacy Systems and Enhancing Data Integrity. Biodiversity Information Science and Standards 8: e141097. https://doi.org/10.3897/biss.8.141097

Abstract

In the digital era, museums confront the challenge of modernising legacy data systems to align with current standards. Part of the RECODE (Rethinking Collections Data Ecosystems) Programme (Dupont et al. 2022) examines the complex transition from disparate departmental object models to a unified system, reflecting the broader museum community's struggle with data standardisation. The case study centres on consolidating five distinct data models accumulated over two decades, each tailored to different departmental needs under various constraints. The data mapping process identifies and aligns similar fields across models, facilitating the integration of analogous data points and exposing redundancies and unique practices embedded over years. Major challenges include data complexity, diversity, and quality issues (cleaning, standardizing, and deduplicating), compounded by the massive scale of the data: currently, the Natural History Museum London (NHM) has 262,831,590 records across 8,175 fields, and a total volume of 12 TB of data, which define both the ongoing data modelling work and our future data migration.

A key innovation in our process is the immediate visibility of data issues that became apparent upon migrating to Amazon Web Services (AWS). AWS serves as the staging

environment for our transition to the new NHM Collections Management System (CMS) and offers an unprecedented platform for directly addressing these challenges. It enables us to query all our data—across all departments and fields—in just seconds. This capability provides a comprehensive view that was previously unattainable.

The core of this presentation is sharing "lessons learned" from navigating the intricacies of an ongoing CMS transition within a museum. The endeavor to untangle and unify diverse data models is a common challenge (IEEE Big Data Governance and Metadata Management Industry Connections Activity 2020,Wu et al. 2022, Wu et al. 2021,Little et al. 2022,Woodburn et al. 2022) highlighting the importance of community engagement and knowledge exchange. Our findings underscore the necessity of cross-departmental collaboration and the benefits of a data-driven data modelling approach.

By sharing our journey and the developed comprehensive object model (Collier and Woodburn 2022), we aim to contribute valuable insights to museums undergoing similar transformations. The RECODE Programme sheds light on the practical aspects of CMS modernisation through the analytics and knowledge derived from over two decades of collections data and data use.

Keywords

legacy data models, collection management system, RECODE programme, data standardization, data mapping, data complexity, Amazon AWS, data migration, collaboration, data modeling

Presenting author

Marla Spencer

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Collier B, Woodburn M, et al. (2022) Rethinking Collection Management Data Models. Biodiversity Information Science and Standards 6 (e91297). [In English]. https://doi.org/10.3897/biss.6.91297

- Dupont S, Woodburn M, Collier B, Ellis L, Hey G, Hsu T, Sadka M, Smith D, et al. (2022) RECODE: Towards a next-gen collections management system as part of the institutional and community data ecosystem. Biodiversity Information Science and Standards 6 (e90886). [In English]. https://doi.org/10.3897/biss.6.90886
- IEEE Big Data Governance and Metadata Management Industry Connections Activity (2020) Big Data Governance and Metadata Management: Standards Roadmap. IEEE Standards Association. URL: https://standards.ieee.org/wp-content/uploads/import/governance/iccom/bdgmm-standards-roadmap-2020.pdf
- Little H, Buschbom J, Best J, Norton B, Lewers K, Blum S, Leeflang S, Niño A, Koivula H, et al. (2022) Standards Mapping Task Group A Task Group of Technical Architecture Group (TAG). https://www.tdwg.org/about/committees/tag/standards-mapping/. Accessed on: 2024-10-27.
- Woodburn M, Webbink K, Jones J, Grant S, Paul D, Trekels M, Groom Q, Vincent S, Droege G, Ulate W, Trizna M, Raes N, Buschbom J, et al. (2022) Latimer Core terms. https://ltc.tdwg.org/terms/. Accessed on: 2024-10-27.
- Wu M, Juty N, Collins J, Duerr R, Ridsdale C, Shepherd A, Verhey C, Castro LJ, et al. (2021) Guidelines for publishing structured metadata on the web. RDA Research Metadata Schemas Recommendations. https://zenodo.org/records/5727048#.YaSw19BBy70. Accessed on: 2024-10-27.
- Wu M, Hagan P, Cecconi B, Richard S, Verhey C, et al. (2022) A Collection of Crosswalks from Fifteen Research Data Schemas to Schema.org RDA Research Metadata Schemas Supporting Output. https://zenodo.org/records/6341481#.Yii3GC8w3rA. Accessed on: 2024-10-27.