BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e26808
doi: 10.3897/biss.2.26808

OPEN ACCESS

Conference Abstract

Options to streamline and enrich biodiversity data aggregation

Donald Hobern‡, Andrea Hahn‡, Tim Robertson‡

‡ Global Biodiversity Information Facility, Copenhagen, Denmark

Corresponding author: Donald Hobern (dhobern@gbif.org)

Received: 21 May 2018 | Published: 21 May 2018

Citation: Hobern D, Hahn A, Robertson T (2018) Options to streamline and enrich biodiversity data aggregation. Biodiversity Information Science and Standards 2: e26808. https://doi.org/10.3897/biss.2.26808

Abstract

The success of Darwin Core and ABCD Schema as flexible standards for sharing specimen data and species occurrence records has enabled GBIF to aggregate around one billion data records. At the same time, other thematic, national or regional aggregators have developed a wide range of other data indexes and portals, many of which enrich the data by interpreting and normalising elements not currently handled by GBIF or by linking other data from geospatial layers, trait databases, etc. Unfortunately, although each of these aggregators has specific strengths and supports particular audiences, this diversification produces many weaknesses and deficiencies for data publishers and for data users, including: incomplete and inconsistent inclusion of relevant datasets; proliferation of record identifiers; inconsistent and bespoke workflows to interpret and standardise data; absence of any shared basis for linked open data and annotations; divergent data formats and APIs; lack of clarity around provenance and impact; etc.

The time is ripe for the global community to review these processes. From a technical standpoint, it would be feasible to develop a shared, integrated pipeline which harvested, validated and normalised all relevant biodiversity data records on behalf of all stakeholders. Such a system could build on TDWG expertise to standardise data checks and all stages in data transformation. It could incorporate a modular structure that allowed thematic, national or regional networks to generate additional data elements appropriate to the needs of their users, but for all of these elements to remain part of a single record with a single identifier, facilitating a much more rigorous approach to linked open data. Most of the other issues we

currently face around fitness-for-use, predictability and repeatability, transparency and provenance could be supported much more readily under such a model.

The key challenges that would need to be overcome would be around social factors, particularly to deliver a flexible and appropriate governance model and to allow research networks, national agencies, etc. to embed modular components within a shared workflow. Given the urgent need to improve data management to support Essential Biodiversity Variables and to deliver an effective global virtual natural history collection, we should review these challenges and seek to establish a data management and aggregation architecture that will support us for the coming decades.

Presenting author

Donald Hobern