Harmonised Data is Actionable Data: DiSSCo's solution to data mapping

Sam Leeflang[‡,§], Wouter Addink[§,‡]

‡ Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
§ Naturalis Biodiversity Center, Leiden, Netherlands

Corresponding author: Sam Leeflang (sam.leeflang@naturalis.nl)

Abstract

Predictability is one of the core requirements for creating machine actionable data. The better predictable the data, the more generic the service acting on the data can be. The more generic the service, the easier we can exchange ideas, collaborate on initiatives and leverage machines to do the work. It is essential for implementing the FAIR Principles (Findable, Accessible, Interoperable, Reproducible), as it provides the "I" for Interoperability (Jacobsen et al. 2020). The FAIR principles emphasise machine actionability because the amount of data generated is far too large for humans to handle.

While Biodiversity Information Standards (TDWG) standards have massively improved the standardisation of biodiversity data, there is still room for improvement. Within the Distributed System of Scientific Collections (DiSSCo), we aim to harmonise all scientific data derived from European specimen collections, including geological specimens, into a single data specification. We call this data specification the open Digital Specimen (openDS). It is being built on top of existing and developing biodiversity information standards such as Darwin Core (DwC), Minimal Information Digital Specimen (MIDS), Latimer Core, Access to Biological Collection Data (ABCD) Schema, Extension for Geosciences (EFG) and also on the new Global Biodiversity Information Facility (GBIF) Unified Model. In openDS we leverage the existing standards within the TDWG community but combine these with stricter constraints and controlled vocabularies, with the aim to improve the FAIRness of the data. This will not only make the data easier to use, but will also increase its quality and machine actionability.

As the first step towards this the harmonisation of terms, we make sure that similar values use the same term in a standard as key. This enables the next step in which we harmonise the values. We can transform free-text values into standardised or controlled vocabularies. For example: instead of using the names J. Doe, John Doe and J. Doe sr. for a collector, we aim to standardise these to J. Doe, with a person identifier that connects this name with more information about the collector.

Biodiversity information standards such as DwC were developed to lower the bar for data sharing. The downside of including minimal restraints and flexibility is that they provide room for ambiguity, leading to multiple ways of interpretation. This limits interoperability and hampers machine actionability. In DiSSCo, data will come from different sources that use different biodiversity information standards. To cover this, we need to harmonise terms between these standards. To complicate things further, different serialisation methods are used for data exchange. Darwin Core Archives (DwC-A; GBIF 2021) use Comma-separated values (CSV) files. ABCD(EFG) exposed through Biological Collection Access Service (BioCASe) uses XML. And most custom formats use JavaScript Object Notation (JSON).

In this lightning talk, we will dive into DiSSCo's technical implementation of the harmonisation process. DiSSCo currently supports two biodiversity information standards, DwC and ABCD(EFG), and maps the data to our openDS specification on a record-by-record basis. We will highlight some of the more problematic mappings, but also show how a harmonised model massively simplifies generic actions, such as the calculation of MIDS levels, which provide information about digitisation completeness of a specimen. We will conclude by having a quick look at the next steps and hope to start a discussion about controlled vocabularies.

The development of high quality, standardised data based on a strict specification with controlled vocabularies, rooted in community accepted standards, can have a huge impact on biodiversity research and is an essential step towards scaling up research with computational support.

Keywords

data harmonisation, data standards, openDS, FAIR, data interoperability, machine actions

Presenting author

Sam Leeflang

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- GBIF (2021) Darwin Core Archives – How-to Guide, version 2.2. Copenhagen: GBIF Secretariat. URL: https://ipt.gbif.org/manual/en/ipt/2.5/dwca-guide
- Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo C, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RW, Imming M, Jeffery K, Kaliyaperumal R, Kersloot M, Kirkpatrick C, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S, da Silva Santos LOB, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson M, Willighagen E, Wittenburg P, Roos M, Mons B, Schultes E (2020) FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2: 10-29. https://doi.org/10.1162/dint_r_00024