A Simple Recipe for Cooking your AI-assisted Dish to Serve it in the International Digital Specimen Architecture

Wouter Addink[‡,§], Sam Leeflang[§,‡], Sharif Islam[§,‡]

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

Corresponding author: Wouter Addink (wouter.addink@naturalis.nl)

Abstract

With the rise of Artificial Intelligence (AI), a large set of new tools and services is emerging that supports specimen data mapping, standards alignment, quality enhancement and enrichment of the data. These tools currently operate in isolation, targeted to individual collections, collection management systems and institutional datasets. To address this challenge, DiSSCo, the Distributed System of Scientific Collections, is developing a new infrastructure for digital specimens, transforming them into actionable information objects. This infrastructure incorporates a framework for annotation and curation that allows the objects to be enriched or enhanced by both experts and machines. This creates the unique possibility to plug-in AI-assisted services that can then leverage digital specimens through this infrastructure, which serves as a harmonised Findable, Accessible, Interoperable and Reusable (FAIR) abstraction layer on top of individual institutional systems or datasets. An early example of such services are the ones developed in the Specimen Data Refinery workflow (Hardisty et al. 2022).

The new architecture, DS Arch or Digital Specimen Architecture, is built on the concept of FAIR Digital Objects (FDO) (Islam et al. 2020). All digital specimens and related objects are served with persistent identifiers and machine-readable FDO records with information for machines about the object together with a pointer to its machine-readable type description. The type describes the structure of the object, its attributes and describes allowed operations. The digital specimen type and specimen media type are based on existing Biodiversity Information Standards (TDWG) such as Darwin Core, Access to Biological Collection Data (ABCD) Schema and Audiovisual Core Multimedia Resources Metadata Schema, and include support for annotation operations based on the World Wide Web Consortium (W3C) Annotations Data Model. This enables AI-assisted services registered with DS Arch to autonomously discover digital specimen objects and determine the actions they are authorised to perform. AI-assisted services can facilitate various tasks such as digitisation, extract new information from specimen images, create relations with other

objects or standardise data. These operations can be done autonomously, upon user request, or in tandem with expert validation.

AI-assisted services registered with DS Arch, can interact in the same way with all digital specimens worldwide when served through DS Arch with their uniform FDO representation, even if the content richness, level of standardisation and scope of the specimen is different. DS Arch has been designed to serve digital specimens for living and preserved specimens, and preserved environmental, earth system and astrogeology samples. With the AI-assisted services, data can be annotated with new data, alternative values, corrections, and with new entity relationships. As a result, the digital specimens become Digital Extended Specimens enabling new science and application (Webster et al. 2021). With the implementation of a sophisticated trust model in DS Arch for community acceptance, these annotations will become part of the data itself and can be made available for inclusion in source systems such as collection management systems and aggregators such as Global Biodiversity Information Facility (GBIF), Geoscience Collections Access Service (GeoCASe) and Catalogue of Life.

We aim to demonstrate in the session how AI-assisted services can be registered and used to annotate specimen data. Although the DiSSCo DS Arch is still in development and planned to become operational in 2025, we already have a sandbox environment available in which the concept can be tested and AI-assisted services can be piloted to act on digital specimen data. For testing purposes, the operations on specimens are currently limited to individual specimens and open data, however batch operations will also be possible in the future production environment.

Keywords

AI, FDO, DiSSCo, annotation

Presenting author

Wouter Addink

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Hardisty A, Brack P, Goble C, Livermore L, Scott B, Groom Q, Owen S, Soiland-Reyes S (2022) The Specimen Data Refinery: A Canonical Workflow Framework and FAIR Digital Object Approach to Speeding up Digital Mobilisation of Natural History Collections. Data Intelligence 4 (2): 320-341. https://doi.org/10.1162/dint_a_00134
- Islam S, Hardisty A, Addink W, Weiland C, Glöckler F (2020) Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. Data Science Journal 19 https://doi.org/10.5334/dsj-2020-050
- Webster M, Buschbom J, Hardisty A, Bentley A (2021) The Digital Extended Specimen will Enable New Science and Applications. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.75736