Calculating the Digitization Level of Specimens with the Minimum Information about a Digital Specimen (MIDS) Standard

Mathias Dillen[‡], Pieter Huybrechts[‡], Quentin Groom[‡], Lynn Delgat[§]

‡ Meise Botanic Garden, Meise, Belgium
§ Howest, University of Applied Sciences, Bruges, Belgium

Corresponding author: Mathias Dillen (mathias.dillen@plantentuinmeise.be)

Abstract

Natural history specimens constitute physical evidence for past observations of nature. They hold further value as the backbone of taxonomy and as historical samples that can be subjected to further analysis. Yet, as physical objects scattered across collections around the world, their scientific use cases are limited by an overall lack of FAIRness, i.e. not easily Findable, Accessible, Interoperable or Reusable. Digitization of these specimens through imaging and categorical metadata capture can improve this FAIRness and has been done to some extent for decades already, but only recently have technical developments in the field of imaging and information technology made it possible for the fruits of these digitization efforts to be widely distributed and utilized.

Digitization can be done in many different ways and while protocols may be well formulated during a project or within the responsible digitization team, they are often not communicated beyond to users, get lost with time and are not available for analysts to assess the state of digitization or make requests concerning material for which further information may be available. Hence, as digitization is ongoing, it is a difficult exercise to estimate how much has been digitized, and to what extent, at the collection level or on a larger scale. The Minimum Information about a Digital Specimen (MIDS) standard that is currently under development by a Working Group of Biodiversity Information Standards (TDWG) aims to address this problem by defining hierarchical levels of digitization, each associated with a set of criteria for a level to be achieved by an individual specimen.

MIDS has been in development since work in the ICEDIG project in 2019 and its earlier drafts have been used in surveys to try and determine digitization status, often through coarse estimates based on the experience of curators. As a result, these scores cannot be considered reproducible or particularly reliable. Ideally, MIDS scores can be calculated automatically based on a mapping made between the data model of the source and the MIDS criteria. These mappings should also take into account any data value that is known not to be reflective of digitization status. While in an ideal world there would be only one

accepted mapping for any data model, different practices causing interoperability conflicts and different kinds of specimens will likely continue to require slight modifications.

To make this concrete, we constructed a few JSON schemas that specify such mappings, based on the current specification of the MIDS standard and the state of biodiversity data in a few sources, including Darwin Core archives for occurrence data as produced by the Global Biodiversity Information Facility (GBIF). These schemas could be incorporated into existing data publication workflows to automatically calculate MIDS levels. We have also developed an R Shiny app with a user interface to make calculations and simple adjustments of the schemas. We welcome anyone interested to further develop the syntax and philosophy behind the schemas and their integration into other systems.

Keywords

data standard, JSON schema, Rshiny, interoperability, FAIR

Presenting author

Mathias Dillen

Presented at

TDWG 2022

Funding program

A major part of the work was done in the context of an internship from the Howest, University of Applied Sciences in Bruges at Meise Botanic Garden. This work was also facilitated by the DiSSCo Flanders project, funded by the Research Foundation – Flanders (FWO) research infrastructure under grant number I001721N.

Conflicts of interest