Biodiversity Information Standards

(TDWG)

http://www.tdwg.org

Minimum Information about a Digital Specimen (MIDS) Charter

A Task Group of the Collections Descriptions (CD) Interest Group

Convenors

Alex Hardisty (Cardiff University, UK)
Elspeth Haston (Royal Botanic Garden Edinburgh, UK)

Core Members

Alex R Hardisty, Cardiff University, UK
Elspeth Haston, Royal Botanic Garden Edinburgh, UK
Wouter Addink, Naturalis, NL
Mathias Dillen, Meise Botanic Garden, BE
Quentin Groom, Meise Botanic Garden, BE
Falko Glöckler, Museum für Naturkunde Berlin, DE
Deborah Paul, Florida State University / iDigBio, US
Mareike Petersen, Museum für Naturkunde Berlin, DE
Hannu Saarenmaa, Bioshare Digitization, FI
Anton Güntsch, Botanic Garden and Botanical Museum Berlin, DE

Motivation

In its most general sense, digitization in natural sciences is the process of converting analog information about physical specimens to digital form, which includes electronic text, images and other representations.

However, the term ‘digitization’ is understood diversely. It can mean, for example: creating database records (of various extents); making images of collections containers, specimens and/or their label(s); a level of data capture (transcription, excluding or including interpretation of data); and more recently, semantic enrichment of data, and notions of ‘born digital’/’digital by default’. From one digitization initiative to another, the outputs can vary widely because aims, practices and procedures vary across different collection types and institutions. Thus, when a curator, collections manager or scientist talks of something being digitized it is not apparent in an objective way what is meant. Nor is it apparent what ‘sufficient digitization’ means and when (if at all) digitization is

complete. Furthermore, most collections need to report on the progress of digitization to the management and/or funding agencies and therefore, agreed measures are needed.

A harmonizing framework captured as a TDWG standard can help clarify levels (depth) of digitization and the minimum information captured and published at each level. This would help to ensure that enough data are captured, curated and published against specific requirements so they are useful for the widest range of possible purposes; as well as making it easier to consistently measure the extent of digitization achieved over time and to set priorities for remaining work. Such a framework would also be beneficial for 'born digital' specimens where digital data is captured from the outset, beginning with the gathering event.

Inspired by the idea of ‘minimum information standards’ adopted in other areas of biology we name this proposed TDWG standard as ‘Minimum Information about a Digital Specimen’ (MIDS) - the topic of the present task group. This harmonizing framework *includes making the data publicly available* because open access policies in countries around the world require that digital data should be findable and accessible, even at the lowest level of available digitized information.

Goals, outputs and outcomes

- Beginning with the existing draft MIDS specification (July 2020)[1] developed with input from the DiSSCo and ADBC programmes, the goal now is to broaden applicability and to achieve international consensus that leads to widespread adoption and implementation.
- Outputs will include the draft standard itself, a summary of the use cases served, reports of pilot implementations and evaluations of the standard’s content, and appropriate proposals for MIDS support by other TDWG standards, such as Darwin Core, ABCD, CD, etc.
- The complete draft minimum information standard for Minimum Information about a Digital Specimen (MIDS) will be submitted for consideration as a TDWG Standard (Autumn 2021).

Strategy

- A working session at TDWG 2020 will introduce the topic to the TDWG community.
- Through a series of monthly virtual working sessions, the task group will review, discuss and improve the existing draft specification to accommodate the variability of digitization processes, procedures and workflows across different collection types and institutions.
- GitHub will be used to co-ordinate the work and documentation. A task-driven work plan will be developed here which will be openly available. This platform will support the focussed working sessions and will help identify and engage user participation from across the bio- and geodiversity community, from all parts of the world.
- Use cases from across the community will be documented and summarised. The MIDS standard will then be tested against these use cases as part of the evaluation process, and adjusted as needed.
- Implementation of the MIDS standard will also be evaluated throughout the development process to determine any critical difficulties or barriers.

[1] Draft MIDS definition, version 0.11, 21July2020. https://bit.ly/MIDSv011.

Stakeholders

A range of key stakeholders from bio- and geodiversity domains can be identified as beneficiaries of MIDS. The task group will engage their participation, especially for the implementation of MIDS. Stakeholders include:

- Developers of collection management systems for automating MIDS calculation and management of missing data; of crowdsourcing platforms in relation to field inclusion and management of missing data; and of other software tools;
- Digitization and administration staff, for example to identify and manage missing data and to calculate costs of and plan for further digitization;
- Management for developing and managing digitization strategies, and receiving reports;
- Public relations staff for public communication of the progress of digitization; and,
- Domain experts/researchers, for example for assessing and developing usability in research and teaching, and for data mining.

Becoming involved

- This Task Group would welcome anyone who has a practical interest in minimum information standards, Digital Specimen information, experience with digitization processes and workflows and the subsequent management (including making public) of the outputs of digitization, including reporting requirements for management and funding agencies.
- Contact the Conveners.

Context/history

Minimum information standards have been an initiative in biosciences to provide sets of guidelines for reporting data derived by relevant scientific methods[2]. As a general principle, however there is no reason confining them to bioscientific disciplines. Minimum information standards can be applied wherever else it is necessary to capture and present (publish) data for interoperability and re-use by others. When followed, minimum information standards should ensure that such data can be easily verified, analysed and clearly interpreted by the wider scientific community. Minimum information standards also facilitate structured databases, public repositories, and development of processes, procedures and software tools.

The Minimum Information Standards for Scientific Collections (MISC)/Authority Files Working Group[3] was established in 2012 by iDigBio, the National Resource for Advancing Digitization of Biodiversity Collections (ADBC) funded by the National Science Foundation. It was not an attempt to establish a standard for minimum information for scientific collections, but rather an attempt to suggest to data providers what data should be provided for ingestion to the iDigBio infrastructure. The guidance (MISC 2012) denoted three categories of elements - i) required, ii) highly desired, or iii) complementary - that were felt necessary to support better practices for discoverability, research use, and cross-linking (through the use of globally unique identifiers (GUID), for example). This work helped the USA community move toward understanding what was needed to enhance discovery, research use, and linking.

[2] Wikipedia: Minimum Information Standards. https://en.wikipedia.org/wiki/Minimum_information_standard.

[3] https://www.idigbio.org/wiki/index.php/MISC/Authority-File-Working-Group.

Design study work funded by the European Union Horizon 2020 ICEDIG (https://icedig.eu/) project (2018 - 2020) for the future European Distributed System of Scientific Collections (DiSSCo) research infrastructure identified that when discussing digitization, many people have different understandings of the term. This leads to confusion and uncertainty when something is described as having been digitized. Thus, the idea for a 'minimum information standard' was born to serve a range of aims that include:

- Offering clarity to collection owners about the minimum information they should be publishing out of digitization initiatives to make digital specimen information useful for multiple purposes of teaching and learning, research, etc.;
- Assisting the global effort to digitize natural science collections, estimated to be 3 billion specimens worldwide by providing a structured framework that clarifies the outcomes of digitization and the level of digitization achieved; to assist prioritization of the remaining work;
- Supporting and contributing towards assessments of fitness for purpose of data (suitability) for feeding specific types of data processing pipelines; and,
- Assisting researchers to know what information to include in their journal articles and data deposits about specimens they have used in their research.

Drawing on prior work, drafting commenced on a specification for Minimum Information about a Digital Specimen (MIDS), (Hardisty et al. draft) and this is continuing in the present DiSSCo Prepare (https://www.dissco.eu/prepare/) project. This work is offered as the starting point for a TDWG task group.

Considering the global nature of the aims outlined, it is appropriate now to propose a TDWG Task Group to prepare a draft global standard on the topic.

Relation to other TDWG interest/task groups

Audubon Core (AC IG): Image/media types and characteristics are outside the scope of MIDS. MIDS confines itself to indicating the availability (or not) of images and other media types. Thus, the expectation is that AC might be used alongside MIDS.

Biodiversity Data Quality: (BDQ IG): MIDS recognises the importance of both quality tests and assertions and vocabularies of value although precise mechanisms of inclusion/alignment have not yet been studied. MIDS can include statements of compliance in terms of what is expected, for example in terms of completeness. The BDQ IG might want to consider implementing the MIDS levels for high-level data quality checks (at least for the quality parameter "completeness").

Collection Descriptions (CD IG): The proposed task group should come under the supervision of the CD IG as the subject matter, MIDS is relevant to the digital description of objects in physical collections of natural science materials. The convenors of that IG have been consulted and give their support.

Darwin Core (DW IG) & ABCD IG: MIDS maps the information elements expected to be present against the appropriate Darwin Core and Access to Biological Collections Data (ABCD) terms. The expectation is that MIDS will further align as convergence in this area proceeds.

Earth Sciences and Paleobiology (ESP IG): The work of the MIDS TG complements the outputs of the ESP IG and is oriented to ensure that minimum information about both biological and non-biological (i.e., fossil, rock, mineral) specimens can be presented.

Resources

- Hardisty, A., Addink, W., Dillen, M., Groom, Q., Haston, E., et al. (*Draft*) Minimum Information about a Digital Specimen (MIDS) v0.11, July 2020. http://bit.ly/MIDSv011 - *draft text of a specification*.
- Borsch, T., Stevens, A.-D., Häffner, E., Güntsch, A., Berendsohn, W.G., et al. (2020): A complete digitization of German herbaria is possible, sensible and should be started now. Research Ideas and Outcomes 6: e50675. https://doi.org/10.3897/rio.6.e50675.
- MISC 2012. iDigBio MISC Data Element Catalog (Phase 1, V0, rev. 15 December 2012). https://www.idigbio.org/wiki/images/c/c9/Phase_I_Report.pdf.

Alex Hardisty, Elspeth Haston

July 2020.