Minimum Information about a Digital Specimen (MIDS)

Update, April 2020 (Elspeth Haston, RGBE & Alex Hardisty, CU)

Summary - What is MIDS?

- The Minimum Information about a Digital Specimen (MIDS) standard specifies the expected information elements recommended to be present in a Digital Specimen after digitization.
  - Digital Specimens are the digital representations in cyberspace of their physical counterparts in natural science collections.
  - Digitisation is the process of converting (analog) information about physical specimens to digital format, which therefore includes electronic text, images and other representations.
- MIDS is a ‘minimum standard’. It recommends *the minimum amount of information* that is useful for various purposes.

In the context of CETAF strategic targets

Natural history collection management and collection access

- 10% of the 1.5 billion specimens in our natural history collections are databased, *digitized and digitally available*, and scientific collection visits increase by 10%
- CETAF best practice and common collections policies are implemented in the majority of member institutions
- *All CETAF institutions have self-assessed collections*

Biodiversity informatics and information technologies

- CETAF interoperable standards … adopted …

Update since CETAF DWG, Vienna 2019.

- Reformatting and structuring in style of a prospective TDWG standard – in progress, with informative and normative sections.
- Several discussions leading to consolidated definitions of each of 4 MIDS levels and information elements expected to be present at each. Added mappings to Darwin Core and ABCD terms.
- Added proposal for extension mechanism at level 3.
- Simplified the expectations of image presence/absence at any level to a single flag
  - “i”, indicating presence/absence. Image types now out of scope.
- Version 0.10 coming soon. Open for further comment.
- Gained prospective support from several institutions.

Structure of MIDS

- MIDS accommodates variability in details of the digitization process by defining several levels of ‘minimum information’.
  - Based on major research and curation requirements put forward by national museums and herbaria. 3 levels and a pre-level (zero).
- It defines the information elements expected to be present at each level.
  - It *always allows the presence of other information elements beyond the minimum*, including institutional and collection specific elements. Extensible at level 3.
- MIDS flags ‘image(s) are available’ but no more. Image types and characteristics are out of scope.*

* but could be defined elsewhere

Table 1: MIDS levels 1 – 3 plus a pre-level (0)

Note: No notion of completeness or of a full or complete record because new information is always valuable and can be added to an existing record at any time.

MIDS level	Record extent	Purpose
1	Basic	A minimal record of specimen information, largely based on current practices, enabling similar discovery capabilities on-line as researchers and curators would have by more traditional means.
2	Regular	Key information fields that have been agreed over time as essential for most scientific purposes.
3	Extended/rich	Other data present or known about the specimen, including links to third-party sources.

Table 1: MIDS levels 1 – 3 plus a pre-level (0)

Note: No notion of completeness or of a full or complete record because new information is always valuable and can be added to an existing record at any time.

MIDS level	Record extent	Purpose
1	Basic	A minimal record of specimen information, largely based on current practices, enabling similar discovery capabilities on-line as researchers and curators would have by more traditional means.
2	Regular	Key information fields that have been agreed over time as essential for most scientific pu
3	Extended/rich	Other data present o third-party sources.
0	*Catalogue (pre-level)*	*A skeletal record making the association between an identifier of a physical specimen and its digital representation, allowing for unambiguous attachment of all other information.*

MIDS levels exist in a context of publishing specimen data and making it accessible on-line, in association with a persistent identifier such as an NSId. Pre-level (0) is available to serve some specific operational needs in this context, encountered in different collections practices.

MIDS level 1 Basic Information elements	DwC term (latest, 2014-11-08)	ABCD term name (3.0)	Recommendations
1. CreatedOn - date/time of record creation	*modified*	*hasDateCreated*	Format conforms to ISO 8601
2. RecordCreator	No equivalent	*creators*	Include ORCID ID if known
3. MIDS level (0 – 3)	No equivalent	No equivalent	
4. Physical specimen identifier (catalog number)	*catalogNumber*	*id*	Cardinality 0..n Can be a barcode
5. Institution code, e.g., from GrSciColl	*institutionCode*	*sourceInstitutionID*	
6. Date/time of subsequent modification (if any)	*Modified*	*hasDateModified*	
7. Material type	*preparations*	*preparationsText*	
8. Name (a name appropriate to the specimen)	*scientificName*	*ScientificName*	See also element 23.

MIDS level 2 Regular Information elements	DwC term (latest, 2014-11-08)	ABCD term name (3.0)	Recommendations
Elements 1 – 8 plus:			
9. Continent	*continent*	*NamedArea/areaClass*	Include One World Collection Regions if known
10. Country	*country*	*Country*	ISO 3166 where possible
11. State/province	*stateProvince*	*NamedArea/areaClass*	
12. County	*county*	*NamedArea/areaClass*	
13. Locality	*locality*	*NamedArea/areaClass*	
14. Latitude/longitude, if known	*decimalLatitude, decimalLongitude*	*latitudeDecimal, longitudeDecimal*	Include georeference datum, source and uncertainty if known
15. Altitude/depth, if known	*verbatimElevation or verbatimDepth*	*hasAltitude or hasDepth*	
16. Collector name	*recordedBy*	*Gathering/hasGatheringAgent*	Include person name identifier if known
17. Collector number	*recordNumber*	*No equivalent*	Enter s.n. or equivalent if not present
18. Collection date	*eventDate*	*Gathering/hasDate*	Enter s.d. or equivalent if not present
19. Collection code/name	*collectionCode*	*sourceID*	What standard?
20. Type status	*typeStatus*	*nomenclaturalTypeText*	
21. Geographical region	*higherGeography*	*NamedArea*	Corresponding to One World Collection Regions
22. Deposited / accession date	*No equivalent*	*Accession/hasDate*	
23. Name identifier	*identificationID*	*Identification/hasIdentifiedTaxon*	Uniquely identifying the name in the Name element (10)

MIDS level 3 Extended Information elements	DwC term (latest, 2014-11-08)	ABCD term name (3.0)	Recommendations
Elements 1 – 23 plus:			
24. Quality assertions	*No equivalent*	*No equivalent*	Format/structure to be specified.
25. External link information The following 5 fields, repeated	*No equivalent*	*No equivalent*	Format/structure as below.

Field name	Definition	Example 1 GBIF occurrence	Example 2 DNA sequence	Example 3 Literature reference
External link	A sequence number, 1..n	1	2	3
Domain	Internet or other administrative domain	gbif.org	ebc.ac.uk	Unknown
Relation	Type of external link or 'Unknown'	hasOccurrence	hasSequence	hasLiterature
Type	Type of the identifier e.g., doi, etc. or 'Proprietary'	gbifId	Proprietary	doi
Identifier	An identifier	https://www.gbif.org/occurrence/1826086349	https://www.ebi.ac.uk/ena/data/view/FJ788436	10.11646/zootaxa.1426.1.3

MIDS level 0 Catalogue Information elements	DwC term (latest, 2014-11-08)	ABCD term name (3.0)	Recommendations
1. CreatedOn - date/time of record creation	*modified*	*hasDateCreated*	Format conforms to ISO 8601
2. RecordCreator	No equivalent	*creators*	Include ORCID ID if known
3. MIDS level (0 – 3)	No equivalent	No equivalent	
4. Physical specimen identifier (catalog number)	*catalogNumber*	*id*	Cardinality 0..n Can be a barcode
5. Institution code, e.g., from GrSciColl	*institutionCode*	*sourceInstitutionID*	

Level 0 is equivalent to simply creating a catalogue record containing a physical specimen identifier / barcode. It is more for operational reasons than for research uses. It often precedes fuller digitisation. Hence, level 0 is termed a pre-level. Nevertheless, is still useful minimum information for advertising/knowing the existence of specimens.

An image can be present, as can a higher classification rank related to the specimen's storage or placement in a collection.

MIDS is a ‘minimum standard’

- As a general principle, institutions are encouraged to publish the fullest available data about their collections and individual specimens at the earliest opportunity, expecting that such data is likely to become enriched and annotated over time.
  - The information elements expected at any level should be the minimum amount of information to be made available.
- Level 2 Regular is the *minimum standard to aim for* as a best practice recommendation, but publishing information in accordance with levels 1 Basic and 0 Catalogue is also acceptable. Level 3 Extended is the gold standard to which DiSSCo aspires.

Questions?

- Several institutions have expressed their interest in trying out MIDS and evaluating it.

- Next steps?

Meaning of ‘digitized’ in MIDS context *

“on-line representations that are publicly available”

Notes:

- Includes the publishing/sharing aspect, which is often treated separately from and after digitization.
- Includes elements of FAIR, so ‘publicly available’ means findable and accessible. Accessibility is an essential component of MIDS.

ICEDIG/ DiSSCo definition of digitization: *The process of converting analog information about physical specimens to digital format, which includes electronic text, images and other representations.*

*Discussed and agreed at meeting on 2nd April 2020.

But what else?

Would add from the FAIR Digital Object Framework.

- On-line representations must:
  - Be both human- and machine-readable.
  - Support machine-actionability i.e., must possess entity and referential integrity and have explicit semantic relations.

Definition of Digitisation

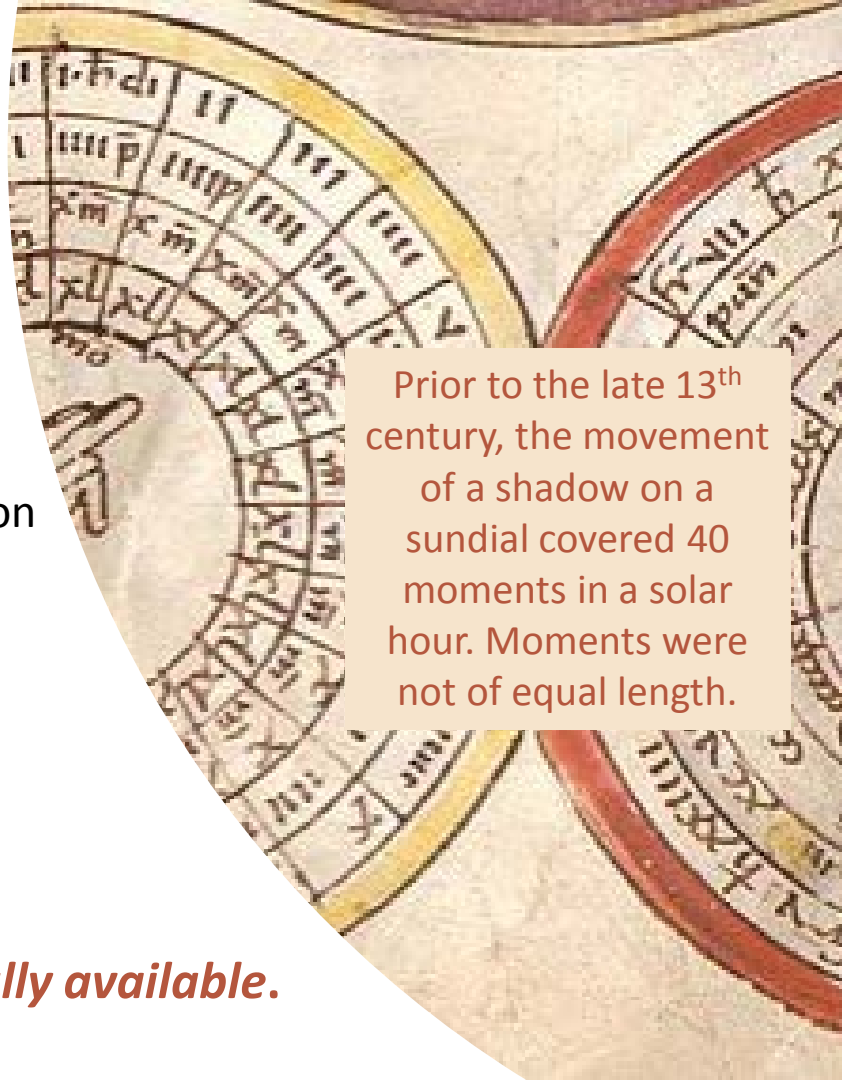
Digitisation means different things to different people.

Practices and procedures vary across different collection types and institutions.

Although the totality of resulting information is often similar, the moments when its components becomes available in digital form varies widely.

Hence, DiSSCo has adopted the simplest definition:
(*proposed by Luc Willemse, Naturalis*)

The process of making physical objects digitally available.



Prior to the late 13th century, the movement of a shadow on a sundial covered 40 moments in a solar hour. Moments were not of equal length.

The process of digitization varies in details. But broadly speaking is the following:

1. Attach an identifier to a physical specimen and create a digital catalogue record.
2. Digitise information about the specimen, including:
   - Imaging the specimen and/or its labels;
   - Extracting, processing and encoding information (from labels, images, etc.);
3. Enrich with supplementary information from third-party sources (links to literature, genetic sequences, etc.)

Aims of MIDS

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