Research Ideas and Outcomes 10: e118244
doi: 10.3897/rio.10.e118244

Project Report

Systematic Design of a Natural Sciences Collections Digitisation Dashboard

Laura J. Tilley[‡], Matt Woodburn[§], Sarah Vincent[§], Ana Casino[‡], Wouter Addink[|,¶], Frederik Berger[#], Ann Bogaerts[¤], Sofie De Smedt[¤], Lisa French[§], Sharif Islam[|,¶], Patricia Mergen[¤,«], Anne Nivart[»], Beata Papp[^], Mareike Petersen[#], Celia Santos[˅], Edmund K. Schiller[¦], Patrick Semal[?], Vincent S. Smith[§], Karin Wiltschke[¦]

‡ Consortium of European Taxonomic Facilities, Brussels, Belgium
§ Natural History Museum, London, United Kingdom
| Naturalis Biodiversity Center, Leiden, Netherlands
¶ Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
# Museum für Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany
¤ Meise Botanic Garden, Meise, Belgium
« Royal Museum for Central Africa, Tervuren, Belgium
» National Museum of Natural History, Paris, France
^ Hungarian Natural History Museum, Budapest, Hungary
˅ National Museum of Natural Sciences, Madrid, Spain
¦ Naturhistorisches Museum Wien, Vienna, Austria
? Royal Belgian Institute of Natural Sciences, Brussels, Belgium

Corresponding author: Laura J. Tilley (laura.tilley@cetaf.org), Lisa French (lisa.french@nhm.ac.uk)

Reviewable v 1

Received: 04 Jan 2024 | Published: 29 Feb 2024

Citation: Tilley LJ, Woodburn M, Vincent S, Casino A, Addink W, Berger F, Bogaerts A, De Smedt S, French L, Islam S, Mergen P, Nivart A, Papp B, Petersen M, Santos C, Schiller EK, Semal P, Smith VS, Wiltschke K (2024) Systematic Design of a Natural Sciences Collections Digitisation Dashboard. Research Ideas and Outcomes 10: e118244. https://doi.org/10.3897/rio.10.e118244

Abstract

This paper describes the design and build of a pilot Natural Sciences Collections Digitisation Dashboard (CDD). The CDD will become a key service for the Distributed System of Scientific Collections Research Infrastructure (DiSSCo) and aims to improve the discoverability of natural science collections (NSCs) held in European institutions, both digitised and undigitised. Furthermore, it will serve as a dynamic visual assessment tool for strategic decision-making, including the prioritisation of digitisation. The CDD pilot includes high-level information from nine European NSCs, covering the number of objects, taxonomic scope, storage type, chronostratigraphy (Earth Science Collections), geographical region and level of detail in digitisation. This information is structured through

a standardised Collection Classification Scheme, which uses high-level categorisation to describe physical natural science collections.

Keywords

data dashboard, natural science collections, biodiversity, geodiversity, collection classification scheme, collections coverage, digitisation metrics, data visualisation, discoverability, prioritisation, decision-making

Introduction

Context

Natural Science Collections (NSCs) hold biological, geological and palaeobiological specimens that are containers of rich data (e.g. geochemical, taxonomic, molecular). They are viable references to past and present biodiversity and geodiversity from across the globe and contain unique specimens that can no longer be collected in the field or are extinct. However, currently, European NSCs are scattered geographically and there is insufficient awareness of the extent of their content due to limited metadata access and a lack of harmonisation in data standards which reduces interoperability and meaningful interpretation. This impedes the collective usage of such valuable data in science, policy and technology for helping to understand and mitigate large scale societal and environmental challenges (e.g. biodiversity loss caused by climate change and habitat destruction). The Distributed System of Scientific Collections Research Infrastructure (DiSSCo) aims to address this impediment by digitally unifying all European collections via a massive digitisation effort under common curation, policies and data standards to ensure that all data are Findable, Accessible, Interoperable and Reusable (FAIR) (www.dissco.eu). This is a gradual process and a service in the form of a Data Dashboard is needed to facilitate this large scale effort. A data dashboard is an information management tool that visually tracks, analyses and displays key performance indicators (KPI), metrics and key data points, for example, to monitor health and development of an organisation and/or specific processes. They often aggregate and reduce voluminous or complex data into a series of summary statistics, sometimes in real time and in a visually appealing way.

This paper focuses on the design and construction of a pilot Collections Digitisation Dashboard (CDD) developed under the SYNTHESYS+ project (a DiSSCo-linked project, funded by the Horizon 2020 Framework Programme of the European Union, www.synthesys.info). It provides an abridged version of the formal project deliverable report (D2.2 'Joint dashboard of collections assessment tools'). The pilot CDD aims to provide a dynamic window for stakeholders to discover the contents, coverage and strength of NSCs held in European institutions, both digitised and undigitised. In addition, it will help track current digitisation progress, to aid in the prioritisation of digitisation and to

support high-level decision-making. The system needs the capability to be embedded into other DiSSCo e-services, a major one being the European Loans and Visits System (ELViS, https://elvis.dissco.eu/welcome). ELViS will be a one-stop-shop that will enable global users (e.g. Scientific researchers, end-user communities etc.) to request loans, virtual access (digitisation on demand) and physical access to collections held in DiSSCo institutions. The CDD will provide support by providing access to information about the physical and digital holdings of NSCs in order to support users in planning visits and arranging loans of material (Islam et al. 2021). In addition, it will also track usage of collections accessed via ELViS (i.e. number of visits and loans, digitisation status etc.).

The work described herein builds upon a previous 'proof of concept' design, which was developed as part of the DiSSCo-linked project ICEDIG (Innovation and consolidation for large scale digitisation of natural heritage) (www.icedig.eu), funded by the Horizon 2020 Framework Programme of the European Union. The ICEDIG project focused on developing the blueprint of DiSSCo. Moreover, work under ICEDIG demonstrated the potential of a collections dashboard, through the development of a set of user stories, the evaluation of technical solutions and the creation of a Collections Classification Scheme. The Collections Classification Scheme uses high-level categorisation to describe physical natural science collections and was developed via a gap analysis of existing data standards (van Egmond et al. 2019).

Project Description

The aim of the SYNTHESYS+ project is to produce an accessible and integrated European resource for users of natural science collections and facilities. Moreover, it is focused on providing the technical foundation for DiSSCo. The work towards the design and construction of the CDD, described herein, was conducted over 14 months (July 2019 - September 2020) and was led by the Consortium of European Taxonomic Facilities (CETAF, https://cetaf.org) with the support and collaborative effort of nine official partner institutions: London Natural History Museum (NHM), Museum für Naturkunde Berlin (MfN), Hungarian Natural History Museum - Budapest (HNHM), National Museum of Natural History - Paris (MNHN), National Museum of Natural Sciences - Madrid (MNCN), Naturalis Biodiversity Center - Leiden (Naturalis), Royal Belgian Institute of Natural Sciences – Brussels (RBINS), University of Copenhagen (UCPH) and Meise Botanic Garden (MBG). There were also valuable contributions from two other SYNTHESYS+ partners who were not officially involved in the task: Natural History Museum Wien - Vienna (NHMW) and the Royal Museum of Central Africa - Tervuren (RMCA).

Collections Classification Scheme

The Collection Classification Scheme was developed to describe physical collections using a high-level categorisation and also for collections information to be presented in a standardised way in the CDD. It will also be the classification used in ELViS and other DiSSCo-linked e-services. The classification scheme presented here is an enhanced

version of the scheme developed in the ICEDIG project. In ICEDIG, the scheme was initially identified by conducting a crosswalk analysis of existing collection related vocabulary in order to delimit existing terminology that could be used in the CDD. The main defined classification dimensions are Institution, Taxonomy, Storage, Geographic Region, Stratigraphic Age and Minimum Information about a Digital Specimen (MIDS) level (van Egmond et al. 2019).

Classification dimensions

In this section, each of the main dimensions of the Collections Classification Scheme are presented and described in terms of purpose, categories they contain, hierarchical level and enhancements that have been made.

Institution

Institutions are identified by their official name and institution acronym, as documented in the Global Research Identifier Database (GRID) and a 2-digit ISO country code. Collections are defined per institution and can only belong to one institution.

Taxonomy

The Taxonomy dimension describes the collections by taxonomic references, i.e. disciplines and their main types of taxonomic areas, for enabling the discovery of the extent of biodiversity and geodiversity covered by DiSSCo participant institutions. The highest level of categorisation are the Natural Science disciplines (Table 1). Enhancements were made to the classification from van Egmond et al. (2019) regarding the addition of Anthropology-specific categories, together with significant amendments to Geology, Palaeontology and Extraterrestrial categories; and Mycology was merged into Botany.

Storage

The Storage dimension (Table 2) is considered essential not only for researchers, but also for collection managers, mainly regarding space and facility planning, because it describes how a collection is preserved (e.g. in fluid jars, dried and pinned). Amongst others, it could be useful for planning research and digitisation workflows and for identification of space needs either for renovating or building facilities. For instance, how objects are preserved may dictate what techniques/methodologies need to be used. Within this dimension, Palaeontology, Geology and Extraterrestrial storage categories from van Egmond et al. (2019) have been enhanced.

Geographic Region

Geographic Region refers to where a specimen/object was collected and not its natural distribution of occurrence in the wild. This dimension adds another layer to the discoverability of collections and information delivery specifically regarding which

biodiversity/geodiversity is represented globally within DiSSCo institutions. It also allows for identification of the uniqueness of collections on an institutional and country level.

The geographic region dimension has been divided into marine and terrestrial (Table 3). The marine regions are based on the ‘International Hydrographic Organisation (IHO) World Seas – Version 3’ (Flanders Marine Institute 2018). In relation to van Egmond et al. (2019), the North Pacific, South Pacific, North Atlantic and South Atlantic have been further divided into 'deep sea', 'shelf area and adjacent seas' and 'unknown'. The term 'adjacent seas' was implemented to better define smaller marine territories/seas, such as the Mediterranean, Red Sea etc.

The terrestrial regions are based on the TDWG World Geographical Scheme for Recording Plant Distributions (WSRPD - level 1) (Brummit 2001). In this case, there were no changes from van Egmond et al. (2019).

Table 1.

Collections Classification Scheme Taxonomy dimension.

Discipline	Categories
Anthropology	Human Biology Archaeology Other
Botany	Algae Bryophytes Fungi/Lichens (including Myxomycetes) Pteridophytes Seed plants
Extraterrestrial	Collected on Earth Collected in space Other
Geology	Mineralogy Petrology Loose sediment Other
Microorganisms	Bacteria and Archaea Phages Plasmids Protozoa Virus - animal / human Virus - plant Yeast and fungi Other
Palaeontology	Botany & Mycology Invertebrates Vertebrates Trace fossils Microfossils Other

Discipline	Categories
Zoology invertebrates	Arthropods - insects (Lepidoptera, Diptera, Hymenoptera, Coleoptera) Arthropods - other insects Arthropods - arachnids Arthropods - crustaceans & myriapods Porifera (sponges) Mollusca (bivalves, gastropods, cephalopods) Other
Zoology Vertebrates	Fishes Amphibians Reptiles Birds Mammals Other
Other Geo/Biodiversity	Other biological or geological objects which fit into none of the other defined categories

To the Terrestrial and Marine region the category 'World/NA' was added for specimens/ objects that could not be assigned to more specific regions. A Region-related ‘unknown’ sub-category was included for objects that had an unknown collection origin.

Table 2.

Collections Classification Scheme Storage dimension.

Domain	Origin	Discipline	Categories	Examples
Biology	Biology Preserved (dead)	Anthropology	Unspecified	
			Dried assemblage	Not in fluid
			Dried - not assembled	Not in fluid, human remains bones, (not recent)
			Fluid preserved	
			Microscope slides	
			Cryopreserved / frozen - 80^{o}C	
			Artefacts: climate controlled conditions	Air conditioning / climate controlled units/rooms
			Artefacts: non climate controlled conditions	Not air conditioned / climate controlled units / rooms can include mummies
			Other	Anything that does not fit into the above
		Botany	Unspecified	
			Pressed and dried	Herbarium specimens
			Dried	Fruits wood samples, not preserved in fluid
			Fluid preserved	Flowers / fungi in alcohol / formalin / glycerine

Domain	Origin	Discipline	Categories	Examples
			Microscopic slides	Microscopic slides
			Cryopreserved/ frozen 80^0C	DNA/RNA, tissue
			Spore print	Spore print
			Other	
		Microorganisms	Unspecified	
			Dried	Not preserved in fluid
			Microscope slides	
			Cryopreserved DNA/RNA	DNA / RNA, tissue
			Other	
		Zoology vertebrates	Unspecified	
			Dried - assembled	Multiple animal parts or entire organism skeletons, stuffed animals
			Dried - not assembled	Animal part: tanned skin, egg shell, etc.
			Fluid preserved	Animals in alcohol/formalin/glycerine
			Microscope slides	Microscopic slides
			Cryopreserved / frozen -80^0C	DNA / RNA, tissue
			Other	
		Zoology invertebrates	Unspecified	
			Dried and pinned	Pinned insects
			Dried - assembled	Not pinned. Multiple animal parts of entire organism
			Dried - not assembled	Animal part, shell, bone, etc.
			Fluid preserved	Animals in alcohol / formalin / glycerine
			Microscope slides	Microscopic slides
			Cryopreserved / frozen -80^0C	DNA / RNA, tissue
			Other	
	Biology Fossilised	Palaeontology	Unspecified	
			Macrofossils (dry preserved)	Hand specimens / slabs / matrix support (i.e. surrounded by original sediment), matrix free (free from original sediment) - botanical, vertebrates, invertebrates, trace fossils etc.

Domain	Origin	Discipline	Categories	Examples
			Mesofossils (dry preserved)	Small fossilised parts of plants such as fruits, leaves and seeds contained in jars, Franke cells - i.e. a paper container, the size of a preparation glass with a circular space covered by a lid-covering glass.
			Microfossils (dry preserved)	Dry samples, in jars, trays (i.e. not preserved in fluid) etc.
			Macrofossils (fluid preserved)	Preserved in a fluid in a jar, a concealed unit.
			Mesofossils (fluid preserved)	Preserved in a fluid in a jar, a concealed unit.
			Microfossils (fluid preserved)	Preserved in a fluid in a jar, a concealed unit
			Fossils preserved in Amber, natural resin	required to be kept in humidity and light controlled storage units.
			Microscope slides	Microscope slides of microfossils, messofossils and macrofossils for either binocular or petrographic microscopes
			Oversized fossils	Too large to be fit into standard storage units
			Other	Sieving residue, other microscopic preparations (SEM stubs) etc.
Geology	Geology	Geology	Unspecified	
			Macro-objects	Hand specimens / hand-held objects / slabs that can be contained in standard units (drawers, shelves, cabinets). For example, rocks, minerals, gems (rough natural form) and ores.
			Micro-objects	Can only be handled/observed with the aid of a microscope. Contained in jars
			Cut/polished gemstones	High-expense/rare/precious stones that need careful handling and contained in secure units
			Microscope slides	Binocular or petrographic microscope slides of rocks, minerals, gems, ore, alloys etc.
			Cores	Rocks, Ore, Sediments (soil, mud etc.) etc.
			Fluids	Hydrocarbons, oils etc.

Domain	Origin	Discipline	Categories	Examples
			Oversized objects	Requires extra space because objects are too large for standard units/containers
			Hazardous material/objects	Material or fluids that are hazardous to health - radioactive, toxic etc.
			Other	Does not fit into the above subcategories, for example, crushed rocks, other microscopic prepared objects (e.g. SEM stubs) etc.
Extraterrestrial	Extraterrestrial	Extraterrestrial	Unspecified	
			Macro-objects	Hand specimens / hand-held / slabs Meteorites, moon rock etc.
			Micro-objects	Can only be handled/observed with the aid of a microscope. contained in jars, sample bags etc.
			Oversized objects	Requires extra space because objects are too large for standard units / containers
			Microscope slides	Thin sections of meteorites etc.
			Other	Anything that does not fit the above
Other geo / biodiversity	Other geo / biodiversity	Other geo / biodiversity	Other geo / biodiversity	

Chronostratigraphy

This dimension is specifically devoted to Palaeontology collections and addresses the fact that these collections include species that lived in ancient times of our Planet. Chronostratigraphy is one of the crucial data for palaeontology collections together with taxonomical and geographical data. This also adds another level of detail for discovering biodiversity, in this case, extinct (Table 4). The categories agreed upon follows the standards of the International Commission on Stratigraphy (ICS) (Cohen et al. 2013).

Minimum Information about a Digital Specimen (MIDS)

To define the level of digitisation of objects in the CDD, the draft specification of the ‘Minimum Information about a Digital Specimen’ (MIDS) was adopted because it defines four levels of digitisation (MIDS-0 being the lowest and MIDS-3 being the most complete), along with the requirements for each level, allowing for a more harmonised and specific understanding of what ‘digitised’ means (Table 5). The Minimum Information about a Digital Specimen (MIDS) specification was used to describe the digitisation level of objects in each dimension of the classification. The MIDS is still under development under TDWG and the version used here was v.0.9 which was developed in 2020, but the specifications are likely to have changed since this study (Haston et al. 2023).

Table 3.
Collections Classification Scheme Geographic Region Classification.

Main category	Regions	Subcategory
Terrestrial	Africa	
	Antarctica	
	Asia Temperate	
	Asia Tropical	
	Australasia	
	Europe	
	North America	
	Pacific	
	South America	
	World/NA	
Marine	Arctic Ocean	
	Indian Ocean	
	North Atlantic	unknown
		deep sea
		shelf area & adjacent seas
	South Atlantic	unknown
		deep sea
		shelf area & adjacent seas
	North Pacific	unknown
		deep sea
		shelf area & adjacent seas
	South Pacific	unknown
		deep sea
		shelf area & adjacent seas
	Southern Ocean	
	World/NA	

Data architecture for acquisition and curation of collection data

Data granularity and aggregation

The first step towards acquiring data for the CDD was to define the level of data granularity and aggregation required. This refers to the extent to which the institutional collections should be quantified according to the different dimensions of the classification scheme (i.e. Taxonomy, Geographic Region, Storage and Chronostratigraphy) and how the dimensions

might need to be combined to support users' needs. It required the consideration of data utility against the effort needed to generate and maintain data by institutions. The decision had to consider a balance between two extremes:

Table 4.

Collections Classification Scheme Chronostratigraphy Classification.

Eon	Era	Period	Epoch
Stratigraphy unspecified			
Phanerozoic	*Any era*		
	Cenozoic	*Any period*	
		Quaternary	*Any epoch*
			Holocene
			Pleistocene
		Neogene	*Any epoch*
			Pliocene
			Miocene
		Paleogene	*Any epoch*
			Oligocene
			Eocene
			Paleocene
	Mesozoic	*Any period*	
		Cretaceous	
		Jurassic	
		Triassic	
	Paleozoic	*Any period*	
		Permian	
		Carboniferous	
		Devonian	
		Silurian	
		Ordovician	
		Cambrian	
Proterozoic	*Any era*		
	Neo-proterozoic		
	Meso-proterozoic		
	Paleo-proterozoic		
Archean	*Any era*		
	Neo-archean		

Eon	Era	Period	Epoch
	Meso-archean		
	Paleo-archean		
	Eo-archean		
Hadean			

Table 5.

A brief description of the four MIDS levels v.0.9 (Haston et al. 2023).

MIDS level	Record extent	Purpose
0(Note)	Bare	A bare or skeletal record making the association between an identifier of a physical specimen and its digital representation, allowing for unambiguous attachment of all other information.
1	Basic	A basic record of specimen information.
2	Regular	Key information fields that have been agreed over time as essential for most scientific purposes.
3	Extended	Other data present or information known about the specimen, including links to third-party sources.

- Collection objects quantified by the four dimensions independently: This is the simplest and requires least effort to contribute the data to the CDD, but it provides low utility of data because questions can only be answered within the individual classification dimensions. For example, users can see how many objects are from South America or how many objects are fungi, but not how many fungi are from South America.
- Collection objects quantified using a combination of all the dimensions in one single breakdown: This would allow users to answer any question related to any combination of the classification schemes used, thus generating a high level of data utility (e.g. how many fluid preseserved fungi specimens are from South America). However, in this case, the amount of effort required would not be feasible for many (if any) institutes, especially within the timeframe of the task, as they would have had to complete up to 50,000 object counts in addition to digitisation level assessments and confidence indicators.

In order to find a middle ground between these two extremes of granularity and aggregation, user stories were analysed to see what dimension combinations were essential. From this analysis, the only combination of dimension that appeared to be useful and achievable was that of the ‘Geographic Region’ and ‘Taxonomy’ dimensions, since ‘Geographic Region’ has a relatively small number of categories compared to the other dimensions. However, it was also agreed that, while collecting object counts for each combination of ‘Geographic Region’ and ‘Taxonomy’ should be feasible, asking for MIDS level assessments in addition to those would be an unrealistic expectation.

Although no other dimensions were combined in their entirety, the highest level of the ‘Taxonomy’ hierarchy (‘Discipline’) was incorporated into each of the dimension breakdowns. The dimension ‘Discipline’ consists of just nine disciplines (‘Zoology invertebrates’, ‘Botany’, ‘Geology’ etc.), so did not greatly increase the amount of data that needed to be contributed. However, it provides a top layer of classification that is common across all breakdowns, which is important for aggregation within the dashboard and for basic interoperability with collections data in other platforms, such as the CETAF Registry of Collections and GBIF Collections Catalogue. A summary of the four breakdown schemes is shown in Table 6.

Table 6.

A summary of the four breakdown schemes used for the CDD dataset.

*only applicable to the 'Palaeontology' discipline.

		Breakdown schemes			
		1: Taxonomy	**2: Taxonomy and Geographic region**	**3: Storage**	**4: Chronostratigraphy**
Dimensions	**Taxonomy level 1 (Discipline)**	yes	yes	yes	yes*
	Taxonomy level 2 (Category)	yes	yes		
	Geographic region		yes		
	Storage			yes	
	Stratigraphic age				yes
Metrics	**Object count**	yes	yes	yes	yes
	MIDS assessment	yes		yes	yes

Data Metrics

Two types of metrics were captured for each breakdown of collections: a count or estimate of the number of physical objects and a measure of the completeness of digital records representing those objects.

Object Count

This is a numeric figure that represents the total number of physical objects (whether digitised or not) within the categories defined by the Collection Classification Scheme. This may be a precise count, but, in most cases, represents an approximation based on curatorial knowledge of the collections or other sources, such as existing collections audit data.

Data providers were also given the option of adding a confidence measure for each object count to show their degree of certainty in the figure. These measures were captured as percentage deviation - for example, +/- 0% would indicate a precise count, whereas +/-

30% would suggest that the count could be up to 30% greater or less than the value given. In practice, many of the confidence figures were left blank due to time constraints or were invalid due to misinterpretations of the methodology, so were not used in the first version of the prototype dashboard. However, there is potential to refine and expand these in future data collection, which would give the opportunity to incorporate statistics, such as upper and lower bounds for collection sizes into future dashboard iterations.

Calculation of MIDS

The data were captured as percentages of the total number of objects with digital records corresponding to each MIDS level. From these percentages, the sum of objects at each MIDS level was calculated, with the quantity of undigitised objects then represented by the remainder. The method by which MIDS percentages were calculated for each collection breakdown was left to the discretion of the contributing institution. Feedback from institutions suggested a range of methods, including queries against the collection management system (or systems), mapping from institutional data standards and rough estimations using curatorial knowledge of the collections and their data.

Data Structure

The data model underlying the dashboard was designed with close reference to the data standard and model under development by the TDWG Collection Descriptions Data Standard Task Group (Latima Core) (Woodburn et al. 2022). This is intended to become the global standard for Natural Science collection descriptions data and so its early adoption for the CDD promotes future interoperability of CDD data with other collection descriptions datasets, such as the CETAF Registry of Collections, ELViS and GBIF Collections Catalogue. The TDWG data model is also being designed to provide for the structured, quantitative collection data that support the dynamic reporting and visualisation offered by the CDD. A simplified representation of the TDWG data model is shown in Fig. 1
.

In the CDD back-end database, the collection classification scheme is represented as dimensions in the TDWG Collection Description Standard data model. The construct is used to differentiate between the multiple breakdowns of each institutional collection according to the different dimensions. This prevents the same object from being counted more than once in any of the dashboard visualisations.

Representing the NSCs as object groups attached to an institution (rather than a fixed hierarchy of institutions, collections and subcollections) means that metrics can be dynamically aggregated and visualised across institutions and also within (and to a degree, across) dimensional hierarchies like Taxonomy and Geographic region.

For the purposes of the pilot CDD, the data model was implemented as a MySQL relational database and the complete data model is shown in Fig. 2.

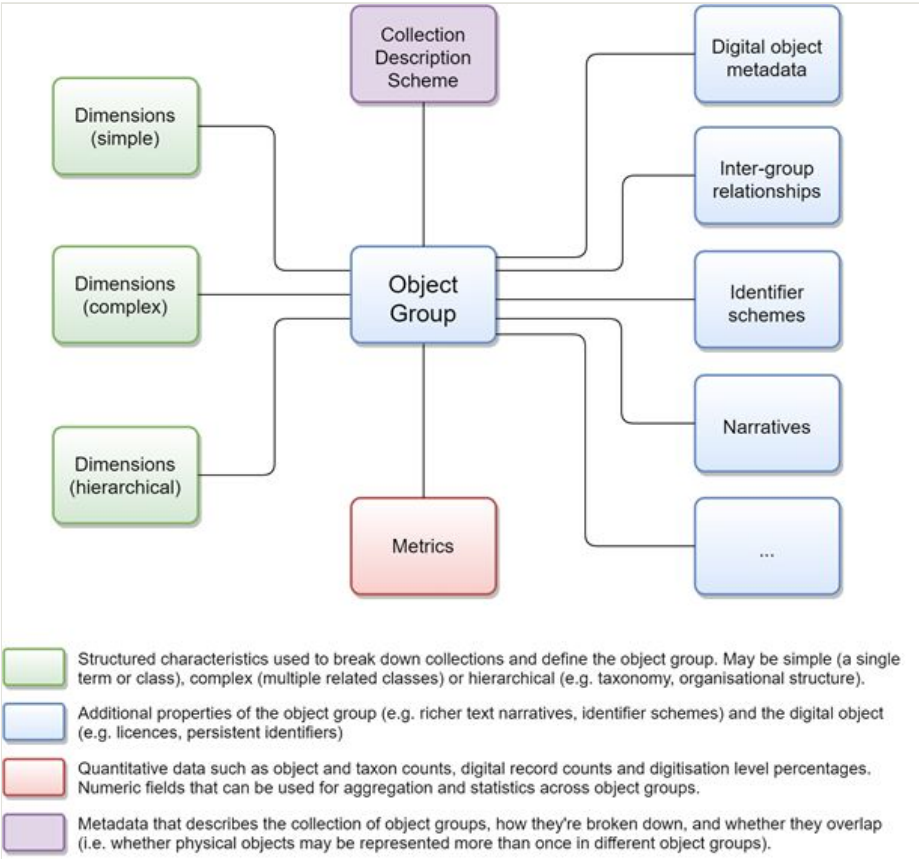

Figure 1. doi

A simplified conceptual view of the TDWG Collections Description data model.

Data collecting and processing

For the pilot dashboard, a Google Sheet survey was considered to be the best tool to test the feasibility for the nine partners to collate data using the Collection Classification Scheme. Data were extracted from the completed surveys through a semi-automated process, involving downloading the individual Google sheets as Excel spreadsheets and using VBA code to generate the SQL queries needed to insert the data into the database in the correct format and structure. The relatively small number of pilot institutions made this a more appropriate method within the timeframe of the task. However, if the pilot framework is scaled up to a much larger number of institutions or more regular updates of the data, then methods for further automation should be explored. Options for this might include more extensive, robust scripting (using Python, for example) to extract and validate data in the survey sheets and directly interact with the database to insert and update data. Alternatively, an ETL (Extract, Transform and Load) tool, such as Pentaho Data Integration, could be employed to achieve similar ends via a more automated workflow.

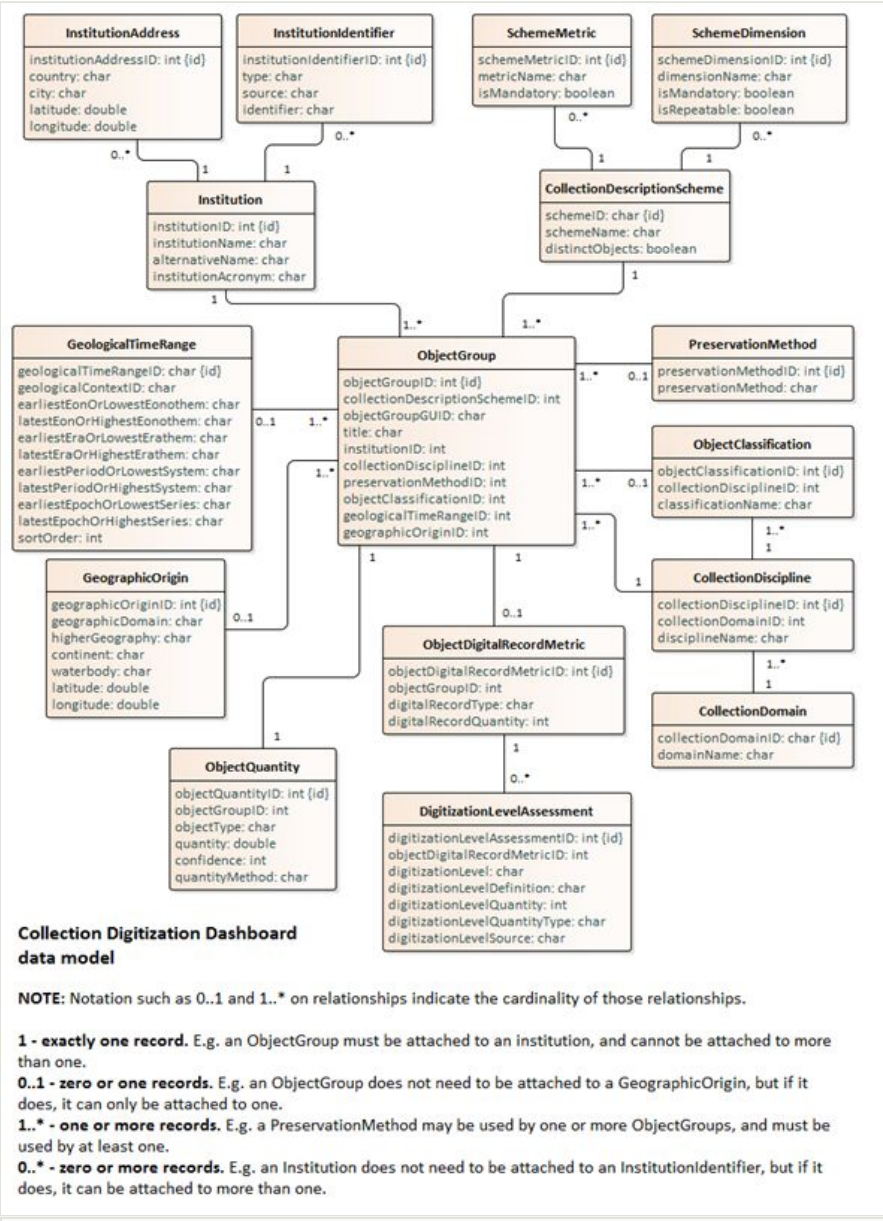

Figure 2. doi

CDD relational data model.

Data validation was carried out both in the source spreadsheets, to ensure there would be no data integrity issues in loading into the database and after each load to ensure that it had been successfully executed. While many common data quality issues were avoided by adding data validation to cells in the survey sheets, some were still encountered in the returned surveys. The most common issue was missing or partial data for object counts

and MIDS assessments for one or more dimensions, reflecting the challenge for institutions in generating and collating this information within the allotted timeframe. Wherever possible, these gaps were handled and the data loaded, but in cases where the integrity of the database or the dashboard might be compromised, then some data were excluded until the issues could be resolved with the contributing institution.

Collections Digitisation Dashboard

Design Process

The design of the CDD was completed in the following phases:

Phase 1: review existing User Stories

The following sources were consulted for providing user stories for CDD requirements:van Egmond et al. (2019); SYNTHESYS+ task partners, the European Loans and Visits System (ELViS) design requirements and the DiSSCo Prepare work packages concerned with collecting Earth Science and Life science user cases for DiSSCo.

The resulting list of 40 user stories was evaluated against the prototype dataset (i.e. the data collected from partners); requirements that were unachievable were identified and flagged as out-of-scope (approximately 40%). The majority of these cases were excluded because the data necessary to meet the requirement was not available within the limits of this project: change-over-time data, granular taxonomy, collection usage, associated research.

To provide information for the high-level CDD structure, 22 in-scope user stories were grouped into five themes (Table 7) that broadly categorised the granularity of data required and provided the initial page-by-page structure of the CDD. The exception to this was the ‘non-functional requirements’ theme, which was used to tag use cases focusing on accessibility, performance etc. that applied to the dashboard as a whole. To avoid duplication between CDD pages, themes 3 and 4 (institution-level and consortium-level overviews, respectively) were ultimately combined into a single page with functionality provided to allow navigation between data at different levels of aggregation.

All partners were asked to prioritise the 22 user stories by using the MoSCoW scoring method (M = Must have, S = Should Have, C = Could Have, W = Will not have currently) (Clegg and Barker 1994). The 'Must Haves' and 'Should Haves' included the ability for the user to find something specific, for example, the size of a certain collection, a collection of a certain taxonomic category and storage category or from a particular geographic region. It also included the ability for institutions to showcase the strengths of their collections and to visually gather information on the overall 'state-of-the-art' of DiSSCo partner collections. Another need was to compare collections within institutions (i.e. digitisation status and size) and across different institutions. Finally, it was essential for the dashboard to be intuitive and user-friendly.

Table 7.

Themes based on user story data needs.

* The examples provided are not real requirements, but are illustrative of the requests provided in the user stories.

Theme	Detail	CDD page	Examples*
1. Find something specific	Similar to search queries: define several parameters and be presented with data that fulfils them.	Locate	‘I want to know which institutions hold collections of type x' 'I want to know which institutions hold both DNA and dried collections'
2. Compare institutional collections	More exploratory: investigate the data in a way that highlights the strengths and weaknesses of a particular collection or group of collections when compared to another collection or group of collections	Compare	‘I want to see what's unique about my collection in the context of the rest of the DiSSCo partners' 'I want to see which institution has the largest digital collection in my country'
3. See an aggregated view of DiSSCo collections	Priority is on a view of the data at a combined/ consortial collection, not institutional: Used to identify high-level areas of weakness/strength and to provide collection stats suitable for use at the policy/national/continental level.	Overview	‘I want to be able to identify gaps across DiSSCo digital collections, so I can prioritise/ fund digitisation more effectively' 'I want to be able to showcase/provide summary status for Natural History collections at the European level'
4. See collection details for a single institution	Single-institution profile view only: suitable for embedding on an institutional website or as a profile within CETAF, GBIF etc.	Overview	‘I want to know what each institution holds so I can market my product/service to them' 'I want to be able to easily share high-level information about my institution's collection to media/policy-makers'
5. Non-functional requirements	Requirements that focus on **how** the dashboard should work, not what it does. Can include security, accessibility, speed etc.	All	‘I want the data to be up-to-date' 'I want the dashboard to be accessible to people with visual impairments'

Phase 2: Prototyping

An agile approach was adopted for building the CDD, carried out by NHM. Each week during the last couple of months of the project, a new version of the CDD was released and stakeholders were asked to provide feedback by the end of the week. Amendments were

then incorporated in the next version, wherever possible. A static view of the dashboard was used in early versions to focus the feedback on the CDD structure and relevance of visualisations used for particular elements of the data, rather than dashboard interactivity and metrics. The fourth version was shared as a live dashboard in order to conduct a user-acceptance test of interactive elements and check performance and display across different systems.

After the structure and content of the CDD had been fine-tuned during the prototyping phase, non-functional requirements were reviewed and changes applied to the dashboard where needed. This process entailed a further three versions of the CDD and they focused on incremental improvements to the user experience (e.g. formatting, performance and accessibility).

Pilot Dashboard

The live and interactive SYNTHESYS+ Pilot CDD is published online: **Collections Digitisation Dashboard**. The licence applied is Creative commons Attribution licence (CC-BY).

The pilot CDD includes data from MBG, UCPH, MfN, HNHM, MNCN, MNHN, RBINS, Naturalis and NHM, which are aggregated and organised within three pages, based on the themes defined during the systematic design. The data are graphically displayed using multiple impactful and appealing visuals per page (e.g. graphs and tables) for addressing the different identified user needs. The CDD has a user-friendly interface with several interactive aspects that make it dynamic, engaging and interesting for users. These aspects include easy access to guidelines/background information on each page (via an 'i' in the top right corner icon, which explains the project, the Collection Classification Scheme and the MIDS). There are data filters that allow the user to choose the granularity of data, as well as specific institutions and parameters of their interest. Visuals can be expanded to whole page views, which provides more detail and allows the users to take quality screenshots for incorporation into presentations and reports.

The first page (Fig. 3) addresses Theme 3 'see an aggregated view of DiSSCo collections'. This page provides an aggregation of data on the total number of objects in all collections and total number of objects digitised in accordance with the MIDS levels. The user can explore the total size of collections as defined by discipline and Taxonomy categories, geographic region against the location of the holding institute and by chronostratigraphic age. The data can be filtered by country and/or institution, thus addressing Theme 4 of the user stories (see collection details for a single institution). More information about a specific collection is given when the mouse cursor is hovered over the item of interest.

The second page shows collection comparison (Fig. 4) and addresses user stories under Theme 2 'Compare institutional collections'. This page allows users to select multiple institutions to compare strengths and uniqueness in terms of disciplines and taxonomy represented and the level at which they are digitised. This information is displayed

graphically in the form of radar charts with percentages. These comparisons are also given as actual numbers within a summary table at the bottom of the page.

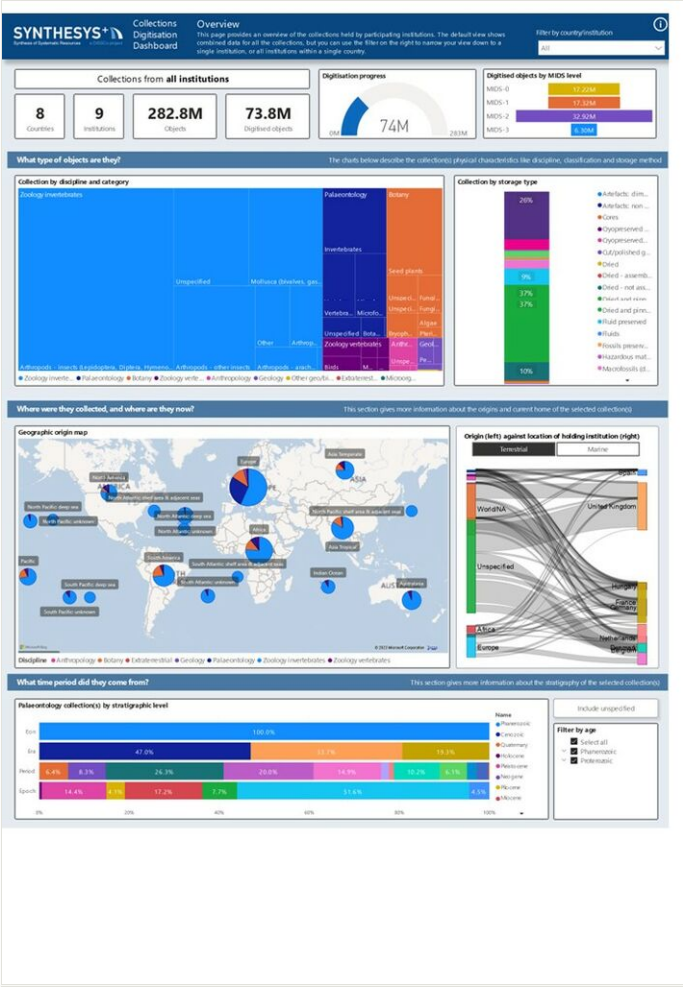

Figure 3. doi

First page of the Pilot CDD showing a collection overview (Licence: CC-BY).

Page three which presents collections location (Fig. 5) addresses the user stories under Theme 1 ‘Find something specific’. It allows the user to locate collections, based on storage type, digitisation level, taxonomy, geographic region and stratigraphic age (for palaeontology only). Only the Taxonomy and Geographic Region dimensions were combined; thus, apart from these two dimensions, users can only infer information and not exactly pinpoint a collection that is, for example, of a certain taxonomic category, from a certain geographic region and additionally preserved in a specific storage type. In order to make this fact clear for the user, this page provides separate choices to view possible combinations, for example: ‘Discipline, Storage and digitisation level’ and ‘Discipline,

Taxonomy and Geographic Region'. When a user clicks on one of these views, they can further filter each of the three options. This page helps the user to visually see which institutions are predominant for their chosen parameters via a map with the location of the institution indicated by a circle of variable size, which refers to the size of the collection it holds. Actual numbers for the size of a collection are provided on a separate page in the form of a table, which is accessed by clicking the 'See data' button (Fig. 6).

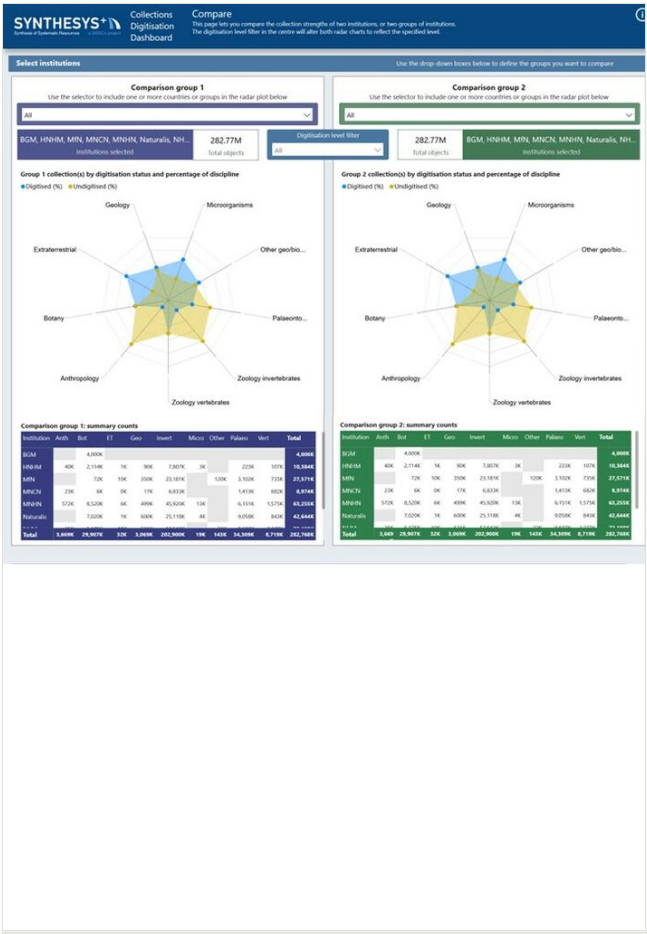

Figure 4. doi

Second page of the CDD which provides a comparison view between the different institutes (Licence CC-BY).

Conclusions

A fully functional pilot **Collections Digitisation Dashboard** has been developed, under the SYNTHESYS+ Project, based on standardised and high-level data from nine NSCs. This work has also led to the enhancement of a Collections Classification Scheme, initiated

within the ICEDIG project, which is being further developed through the TDWG community (van Egmond et al. 2019).

A number of issues remain outstanding and need to be developed under other future endeavours to finally achieve the end goal of delivering a sustainable, dynamic overview of the state of Natural Science Collections and to support the full range of DiSSCo activities. These issues are presented below along with future recommendations for next steps.

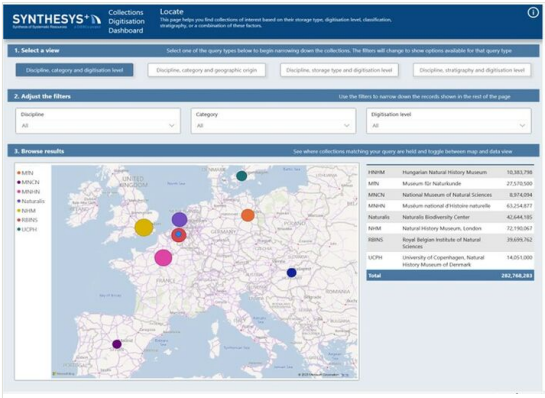

Figure 5. doi

Third page of Pilot CDD, which helps users find the location of collections in Europe based on criteria: Storage, Digitisation level, Discipline, Taxonomy, Stratigraphy (Licence CC-BY).

Source Data Collation

The trade-off between gathering structured high quality data about an NSC, versus the institutional effort involved in provisioning the source data, has long been the primary barrier to delivering an overview of global natural science collections. Additional factors, including the absence of associated data standards, the fact that the data are often held by multiple individuals within an organisation (if held at all), the need to provide regular updates of the data and the absence of any technical agreements on how to provide the

data, all compound to create a complex and potentially insurmountable challenge for a goal that, from the outset, seemed to be a simple problem.

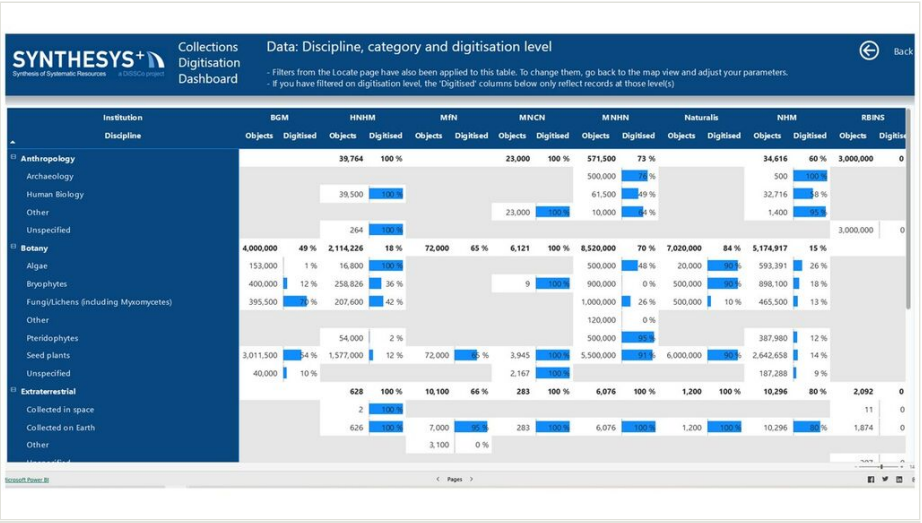

Figure 6. doi

Presenting the page that appears when the 'See data' button is clicked on the third page of the Pilot CDD. The page provides actual object numbers for the different classification dimensions (Licence CC-BY).

Several regional or thematic efforts have been established to solve this problem. Index Herbariorum (IH) is the directory of information on the world's herbaria (addresses, contacts, specialities, size etc.). It is a well-managed resource and highly regarded as a tool by the botanical community. No full equivalent exists globally for other natural history collections, although national/regional infrastructures, such as the Atlas of living Australia (ALA) collections pages, the iDigBio US Collections List and the CETAF Institution profiles serve similar roles. GBIF has recently integrated the Global Registry of Scientific Collections (GRSciColl) into its registry as a framework that can be extended with richer information curated by collections communities.

The GBIF Online international consultation (April 2022) examined the issues associated with the use, information, technology and governance of a global collections catalogue (Hobern et al. 2020). Recommendations from this consultation included that each institution should have primary responsibility and control for information on its collections. However, it may be appropriate to delegate full or partial responsibility for ensuring quality and standardisation of collections descriptions to thematic, regional or national communities that have qualified data curators and/or automated quality checks in place. In this regard, communities, such as Index Herbariorum, ALA , iDigBio and CETAF, play an important role supporting collections and promoting standards-based practices.

Within a European context, CETAF is presently in the final stages of redeveloping the CETAF Institution Profiles as the CETAF Registry of Collections. This has the potential to

provide a more automated approach to the provision of data to the CDD and increase the reliability and accuracy of the sourced information by engaging directly with the CETAF community. The intention is that this registry will be a data entry and management interface for the DiSSCo European Collection Objects Index (ECOI), an internal service for the cataloguing and curation of digital objects by the DiSSCo hub. More specifically, the CDD will hold high-level information about European NSC institutions, including different institutional features apart from collections, such as facilities and laboratories. Navigation is both ‘human readable’ via a user-friendly interface and ‘machine readable’ to facilitate data exchange and harvesting of data. In this regard, the DiSSCo ECOI has the potential to become a stable source for information for European collections feeding into the GBIF Collections Catalogue.

Interoperability with Other Collection Descriptions Data Initiatives

TDWG Collection Descriptions Data Standard

The pilot CDD data model has been designed in alignment with the development of the TDWG Collection Description data standard and model (Latima core) (Woodburn et al. 2022). This early adoption means that CDD data should be interoperable with other key platforms that are intending to adopt the standard, including the CETAF Registry of Collections and DiSSCo ECOI.

As the CDD and Collection Classification Scheme was being developed in parallel with the TDWG standard and to some extent moving at different speeds, there was a degree of divergence at the point where the CDD database needed to be finalised. Work will continue to make sure that outstanding CDD requirements are incorporated into the CD standard and that the CDD database continues to conform to the standard as it develops.

Common hierarchies and vocabularies

While the use of a common data standard provides a base layer of technical interoperability between different collections datasets, data are made truly comparable by the use of common hierarchies and vocabularies (such as the classification schemes used for the CDD). Greater harmonisation of these across initiatives is a longer term challenge, especially for those already well established, but there is potential for incremental gains in this area. For example, it has been agreed that the nine categories specified in the ‘Discipline’ layer of the CDD ‘Taxonomy’ hierarchy will be harmonised across the CDD, CETAF Registry of Collections and DiSSCo ECOI (including the European Loans and Visits System (ELViS), providing a common top layer of collections breakdown across these platforms.

Global persistent identifiers

A core requirement for interoperability between datasets is the use of globally unique persistent identifiers (PIDs) to identify collections and their subsets. This is a framework that needs to be addressed at the global community level, rather than duplicated by

individual platforms and conversations are progressing on this topic within and between DiSSCo, CETAF, GBIF and other contributors. For the CDD database, a temporary solution using GUIDs (Globally Unique Identifiers) has been used with the intention of adopting a wider community PIDS framework for collections when it is available.

Alterations and Additions to the Collections Classification Scheme

There was significant effort towards trying to develop a comprehensive and representative Collection Classification Scheme for Natural Science Collections. Nevertheless, time constraints on the input from some disciplines mean that further alterations/expansion may be required in certain areas. For example, future work should include the development of a schema and identifiers for living collections.

After wider dissemination of the Collections Classification Scheme to the CETAF Earth Science Group, we received useful feedback about the classification of minerals and meteorites. Due to the late stage of receipt, this feedback could not be incorporated into the CDD, but we have included the recommendation that mineralogy, as an independent discipline, should not be classed within geology, since minerals have their own complex classifications and are often curated separately from geology collections. Suggested categories within the Discipline Mineralogy are ‘Minerals’ and ‘Gems’. The proposed storage type categories for ‘Mineralogy' are 'cut/polished gems’, ‘Powder in vials’, ‘Radioactive’, ‘Humidity controlled containers’ and ‘Asbestos form in Perspex boxes’. The renaming of the taxonomic categories for the discipline ‘Extraterrestrial’ was also recommended. The recommended new categories are as follows: 'Terrestrial finds/falls’, ‘Terrestrial Impacta’ and ‘Sample returns’. For geology, it was suggested that the category ‘Petrology’ would be better replaced by ‘rocks’; and loose sediment replaced by sediment. Along with other improvements in the collections classification scheme, the future development of an age classification for Anthropology collections was also a mentioned prospect.

Living collections (notably the outdoor and indoor botanical collections) could also be added into future developments of a DiSSCo Dashboard that should go then beyond the “Digitisation” scheme and model that the current CDD pilot proposes. The same applies to a certain extent to some collections hosted by zoos and aquaria. However, neither one of those collections fall under a global digitisation endeavour and the development should rather focus on establishing interoperability standards and sharable flows of information, whenever possible.

Future needs from the NSC user community

Some of the feedback and requirements received from partners during the agile build and design of the CDD could not be incorporated in the pilot version. This is due to limits on the data collected and MS power BI’s licensing model that controls the publishing and implementation mechanisms used for the CDD. The feedback includes such requests as embedding the institutional logos; the need to see the rate of progress in digitisation; the

ability to export the underlying data for analyses; and the ability for institutions to embed, in their own website, a pre-filtered view of the overview page (e.g. pre-filtered to their institution). These requests should be considered in further work of enhancing the CDD, especially in helping to explore alternative software solutions to construct the dashboard. A solution that has more features and functions, especially with regard to configurability and fine-grained control of dashboard functionality may be needed to meet these requirements. Further work will be conducted under CETAF and the construction phase of DiSSCo (2024 - 2025).

Glossary

CETAF (Consortium of European Taxonomic Facilities): CETAF is a network of European natural science collections, which supports and advocates the value of taxonomy to science and society.

Collections Classification Scheme: describes physical natural science collections in a standardised/harmonised way using a high-level categorisation. The main classification dimensions are Institution, Taxonomy, Storage, Geographic Region, Chronostratigraphic Age and Minimum Information about a Digital Specimen (MIDS) level.

CDD (Collections Digitisation Dashboard): This is an interactive dashboard which summarises the digitisation status and content of natural science collections.

DiSSCo (Distributed System of Scientific Collections): DiSSCo is a pan-European Research Infrastructure for natural science collections, which aims to digitally unify European natural science assets.

Dashboard: Data dashboards are an information management tool that visually track, analyse and display key performance indicators (KPIs), metrics and key data points in order to monitor the health and development of an organisation and/or specific processes. Dashboards often aggregate and reduce voluminous or complex data into a series of summary statistics, sometimes in real time and in a visually appealing way.

ICEDIG (Innovation and consolidation for large scale digitisation of natural heritage): This was an EU-funded project (now finished) which supported the design (blueprint) phase of DiSSCo and designed some of the technical, financial, policy and governance aspects required to operate DiSSCo.

MIDS (Minimum Information about a Digital Specimen): This is a specification that defines different levels of digitisation (MID-0 being the least complete level and MIDS-3 being the most complete), along with the minimum data requirements for each level. The specification allows for a more harmonised and specific understanding of what 'digitised' means. It is being developed into a standard by TDWG.

SYNTHESYS+: This project aims to create a high quality approach to the management, preservation and access to European natural history collections. It helps to lay the

foundations for DiSSCo by creating an accessible, integrated European resource for research users in the natural sciences.

TDWG (Biodiversity Information Standards): TDWG develops data standards and guidelines for recording data about organisms, including the Darwin Core standard.

TDWG Collections Description Data Standard: This standard is under development and describes groups of natural history objects.

Acknowledgements

A special thank you to all the staff at partner institutions (Naturalis, MNHN, HNHM, MNCN, MfN, RBINS, NHM, UCPH, MBG, NHMW, RMCA) who contributed and provided support in the NSCs data acquisition process for the CDD. Thank you to Karsten Gödderz for providing valuable administrative, organisational and communication support for the CETAF Secretariat, which facilitated efforts in completing D2.2. The CETAF Earth Science Group are gratefully acknowledged for providing expertise to support the Collection Classification Scheme regarding Palaeontology, Geology and Extraterrestrial materials. Thank you to Rachel Walcott (National Museum of Scotland) for providing recommendations for improvement of the Collection Classification Scheme regarding Mineralogy, Geology and Extraterrestrial material. Finally thank you to Alex Hardisty (Cardiff University) for providing support and advise on the usage of the MIDS in the data acquisition process of D2.2.

Author contributions

Laura Tilley: Conceptualisation, Investigation, Methodology, Project administration, Supervision, Writing - original draft, Writing - review & editing**. Matt Woodburn**: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing**. Sarah Vincent**: Data curation, Formal analysis, Methodology, Software, Validation, Visualisation, Writing - original draft, Writing - review & editing**. Ana Casino**: Conceptualisation, Project administration, Supervision, Writing - original draft, Writing - review & editing**. Wouter Addink**: Conceptualisation, Investigation, Writing – review & editing. **Frederik Berger**: Investigation. **Ann Bogaerts** Investigation. **Sofie De Smedt** Investigation. **Lisa French** Writing – review & editing **Sharif Islam**: Conceptualisation, Investigation. **Patricia Mergen**: Conceptualisation, Investigation, Writing – review & editing. **Anne Nivart**: Investigation. **Beata Papp**: Conceptualisation, Investigation. **Mareike Petersen**: Investigation, Writing – review & editing. **Celia Santos**: Conceptualisation, Investigation**. Vince Smith**: Writing - review & editing**. Patrick Semal**: Conceptualisation, Investigation, Writing – review & editing. **Edmund Schiller**: Conceptualisation, Writing – review & editing. **Karin Wiltschke**: Conceptualisation, Writing – review & editing.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Brummit RK (2001) World Geographical Scheme for Recording Plant Distributions. Edition 2. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/109
- Clegg D, Barker R (1994) Case Method Fast-Track: A RAD Approach. Addison-Wesley [ISBN 978-0-201-62432-8]
- Cohen K, Finney S, Fan J (2013) International Chronostratigraphic Chart v2016/04. International Commission on Stratigraphy. http://www.stratigraphy.org/ICSchart/ChronostratChart2020-03. Accessed on: 2020-3-10.
- Flanders Marine Institute (2018) International Hydrographic Organisation Sea Areas. Version 3. URL: https://www.marineregions.org/
- Haston E, Hardisty AR, Addink W, Dillen M, et al. (2023) The Minimum Information about a Digital Specimen (MIDS) specification for Natural Science Collections (draft).
- Hobern D, Asase A, Groom Q, Paul D, Robertson T, Semal P, Thiers B, Woodburn M (2020) Advancing the Catalogue of the World's Natural History Collections. GBIF Secretariat: Copenhagen https://doi.org/10.15468/doc-wnsx-ep77
- Islam S, Hardy H, Wilson S (2021) ELViS is in the Building: The European Loans and Visits System and first experiences with Transnational and Virtual Access. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.75312
- van Egmond E, Willemse L, Paul D, Woodburn M, Casino A, Gödderz K, Vermeersch X, Bloothoofd J, Wijers A, Raes N (2019) Design of a Collection Digitisation Dashboard. Zenodo https://doi.org/10.5281/zenodo.2621055
- Woodburn M, Buschbom J, Droege G, Grant S, Groom Q, Jones J, Trekels M, Vincent S, Webbink K (2022) Latimer Core: A new data standard for collection descriptions. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.91159