D2.2 JOINT DASHBOARD OF COLLECTIONS ASSESSMENT TOOLS

LAURA TILLEY, MATT WOODBURN, SARAH VINCENT, ANA CASINO, WOUTER ADDINK, FREDERIK BERGER, ANNE BOGAERTS, SOFIE DE SMEDT, SHARIF ISLAM, PATRICIA MERGEN, ANNE NIVART, BEATA PAPP, MAREIKE PETERSEN, CELIA SANTOS, EDMUND SCHILLER, PATRICK SEMAL, VINCE SMITH, KARIN WILTSCHKE

Grant Agreement Number | 823827

Acronym | SYNTHESYS PLUS

Call | H2020-INFRAIA-2018-2020

Start date | 01/07/2019

Duration | 48 months

Work Package | NA2

Work Package Lead | CETAF – Laura Tilley, Ana Casino

Delivery date | [17.09.2020]

Table of Contents

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

Summary

Within Task 2.2 the ICEDIG collections classification scheme was enhanced, aligning it with the TDWG Collections Descriptions data model. This paper presents the Deliverable of a pilot (click link to explore) **Collections Digitisation Dashboard** (CDD) developed under Task 2.2 '*Integrate and expand institutional collection assessments*' within Work Package NA2 of the SYNTHESYS+ project. The aim of the CDD is to serve as a dynamic visual assessment tool for high-level decision making (e.g. prioritisation of digitisation) and to improve the discoverability of European Natural Science Collections (NSCs) (both digitised and undigitised). This addresses a critical need for the DiSSCo (Distributed System of Scientific Collections) initiative and DiSSCo's associated infrastructure. Task 2.2 as a pilot endeavour (Technical Readiness Level[1] (TRL) 6), only includes the high-level information from 9 partner NSCs, covering the number of objects, taxonomic scope, collections storage categories, stratigraphy, geospatial range and level of detail in digitisation according to standards. This information is structured through a standardised Collection Classification Scheme and displayed with associated metrics. Task 2.2 builds upon a preliminary design of a CDD (TRL 3) that was itself a Deliverable of the now finalised EC funded ICEDIG project where the user needs analysis sought to balance between the effort of data acquisition from partners with the data granularity required to address the compiled user stories. Data was acquired from partners *via* an online form that fed into an accompanying database conforming to the agreed Collection Classification Scheme.

The design and build of the pilot CDD were split into three main development phases. Phase 1: Agreement on the user stories (building on those collated in ICEDIG); Phase 2: Capture of dashboard requirements (i.e. deciding how to visually present summaries of the aggregated data and agreeing other styling specifications); and finally Phase 3: Agile build of the CDD using Microsoft Power BI, including a series of iterative consultations with stakeholders.

This Deliverable includes a discussion of next steps for operationalising the CDD, discussions on the sources for gathering these data (including the CETAF Registry of Collections), the role of GBIF's Collections Catalogue (as determined through the associated consultation in SYNTHESYS+ Task NA5.1), the integration of the classification schemes into the emerging TDWG Collections Description data standard, and lastly, the link between the CDD and the

---

[1] Technology Readiness Levels (TRLs) are indicators of the maturity level of technologies, from 1(lowest) to 9 (highest), see also: https://enspire.science/trl-scale-horizon-2020-erc-explained/

European Loans and Visits System (ELViS) that is being developed in SYNTHESYS+ Task JRA1 as a relevant component of the DiSSCo RI.

Keywords: Data dashboard, Natural Science Collections, biodiversity, geodiversity, Collection Classification Scheme, collections coverage, digitisation metrics, visual tool, discoverability, prioritisation.

1. Introduction

1.1 Context

This Deliverable focuses on the construction of a pilot Collections Digitisation Dashboard (CDD), developed under Task 2.2 of the SYNTHESYS+ project. This work was conducted over 14 months (July 2019 - September 2020) and was led by the Consortium of European Taxonomic Facilities (CETAF) with the support of 9 partner institutes (see section 2.1).

The following sections of this report provide further insight into the scope of the SYNTHESYS+ Pilot CDD; the processes of collaboration; methods of development, data acquisition; proposals for future work and a statement on author contributions.

1.2 Scope

Data dashboards are an information management tool that visually tracks, analyses and displays key performance indicators (KPI), metrics and key data points 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. **The aim of the CDD is to provide a dynamic window for stakeholders to discover the contents, coverage and strengths of European NSCs (both digitised and undigitised), as well as a tool for digitisation prioritisation, measuring digitisation progress and high-level decision making.**

Work on the CDD covers three distinct areas:

- **Content:** The CDD intends to visually summarise the status of collections across the community of institutions, starting in Task 2.2 as a pilot endeavour covering 9 participants, with the potential to expand across all collection holding institutions in Europe, including DiSSCo and CETAF members.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

- **Presentation:** It provides a data dashboard with interactive visual elements to provide high-level information about the NSCs (Natural Science Collections) of partner institutions. It provides summaries and comparisons of their number of objects, taxonomic scope (e.g. including both biological and geological), categories of preservation, stratigraphic age, geospatial range, level of digitisation and digital content availability for reuse (as measured through conformity to various digitisation standards). This information is mapped to a standardised Collection Classification Scheme to enable cross-institutional aggregation and comparison of data.

- **User needs:** It provides a sustainable and easy to maintain evidence base and tool to facilitate decision making (e.g. digitisation prioritisation) within DiSSCo and across European NSCs. In this respect the system needs the capability to be embedded into the ELViS platform to support requests for digital loans and physical access requests.

The deliverable, Technical Readiness Level 6 (TRL 6), builds upon a preliminary design of a CDD that was itself a deliverable of the now finalised EC funded DiSSCo linked project called ICEDIG (Innovation and consolidation for large scale digitisation of natural heritage). The ICEDIG deliverable 'Design of a collection digitisation dashboard' (D2.3) was led by Naturalis (van Egmond et al. 2019). This preliminary design (TRL 3) supported the needs of the end-users as determined through a set of user stories, an evaluation of dashboard technical solutions, creation of an initial Collection Classification Scheme *via* a gap analysis of existing data standards; and investigation of methods for data collection. From this work a collection dashboard for European collections was designed to demonstrate its potential, using the software Microsoft Power BI (*Microsoft Corporation*). Data used in the ICEDIG project was collected from a survey that was developed for a different purpose and lacked the information to address the user stories sufficiently. For example, the collections breakdown lacked a rigorous standardised vocabulary and only retained percentage estimates of digital accessibility. This highlighted the need for a more rigorous and standardised approach. As a consequence, a task group (TG CDD) was established under ICEDIG, dedicated to the harmonisation of data requirements for visualisations and providing recommendations to the Biodiversity Information Standards (TDWG) Collection Descriptions Data Standard Task Group, who are developing a global collections description data model for describing entire NSCs. Another relevant outcome from ICEDIG D2.3 was the recognition of the need for an authoritative data source to automate data collection and to ensure that the final CDD will provide the most up-to-date and reliable data.

SYNTHESYS⁺
Synthesis of Systematic Resources   a DiSSCo project

From the ICEDIG CDD design, several shortcomings were detected and SYNTHESYS+ Task 2.2 has made further developments in the following areas:

- Enhancement of the Collection Classification Scheme and its alignment with the TDWG collection description data model (see section 3).

- The need for a dedicated survey specifically for collecting data for the CDD using the Collection Classification Schemes (see section 4).

- Further analyses of the ICEDIG user stories to identify gaps in user-types and data needs, as well as the addressing the need to thematically group data based on user needs. This was to help with the systematic design of CDD (see section 5).

- Planning and development of a sustainable mechanism to capture this new data (e.g. the CETAF Registry of Collections) and integrate this with a global index (e.g. the GBIF Collections Catalogue) (see section 7).

- Improvement of data metrics (see section 4).

This newly developed CDD in NA2 advanced the TRL 3 proof of concept developed in ICEDIG to TRL6 (Technology demonstrated in relevant environment). Integration with ELViS later in the project will further advance the CDD into TRL7 (prototype demonstration in operational environment).

1.2.1 Collections

In the CDD, collections are defined within a hierarchical classification, 'Institution' being the highest-level descriptor, thus collections are defined per institution and can only belong to one institution. Only collection objects that are of natural origin (e.g. biotic or geological specimens) are included, leaving out of scope objects such as paintings and archives. Living collections (e.g. living organisms in zoos and botanical gardens) are out of scope due to the difficulty of assigning stable identifiers to individual specimens with changing life cycles. Human remains originating from medicine are also not included because they are not related to geo and biodiversity in the intentions of DiSSCo.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

1.2.2 Digitisation

Digitisation is defined as the process of making a physical object and its associated information digitally available (van Egmond et al. 2019). To define digitised objects in the CDD, we adopted the draft 'Minimum Information about a Digital Specimen' (MIDS) specification because it defines different levels of digitisation (MIDS-0 being the lowest and MIDS-3 beginning the most complete), along with the minimum data requirements for each level, allowing for a more harmonised and specific understanding of what 'digitised' means. Therefore, the MIDS levels are considered to enable more effective assessment of digitisation priority areas, in terms of identification and planning of digitisation workflows (Hardisty et al. in progress).

The MIDS specification was originally conceived as part of the ICEDIG project, and refinement continues within the DiSSCo Prepare Project (The EU funded implementation phase). In addition, a working group under the TDWG Collection Descriptions Interest Group has been proposed to introduce MIDS as a global standard. At the time of data collection for the CDD, MIDS v0.9 was the current version, and although not yet finalised, it was decided that the specification still represented a more advanced definition of 'digitised' than those currently available. The CDD also provided an opportunity for task partners to test the specification and feed back into the MIDS development process, with a number of suggestions now incorporated into MIDS v0.10.

1.2.3 Prioritisation

In the context of DiSSCo and thus the CDD, prioritisation refers to decision making for effective streamlining of the deployment of resources for digitisation, in accordance to FAIR principles, to facilitate access to and re-use of the collections, and allocate funding for digitisation projects.

2. Work process

2.1 Partners

Development of the CDD was managed by the Consortium of European Taxonomic Facilities (CETAF). This was a collaborative effort and included the significant contributions from 9 official task partners: 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 (MGB). There were also

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

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).

All 9 task partners were expected to provide data for the dashboard. However, the coincidence of the COVID-19 global pandemic led to unexpected closure of institutions, impacted staff working hours and their access to collections. Therefore, 3 partners that could not provide data to the completeness requested for the CDD within the task timeline, although the data can still be added after the task has finished.

2.2 Collaborative work mechanisms

The successful delivery of a fully functioning pilot CDD has required a continuous collaboration with partners involved in the task, with other SYNTHESYS+ work packages (particularly JRA1 and NA5), the CETAF community at large, and also with external linked initiatives such as the Global Biodiversity Information Facility (GBIF) and TDWG. Communication and work between task partners was organised via virtual monthly meetings (Zoom Version: 5.0.2 (24046.0510), Zoom Video Communications, Inc., CA, USA), using Teamwork (Teamwork Projects version 14.4.33, Teamwork Crew Ltd, Cork, Ireland) as the task management platform, and through Google G Suite (Google Ireland Ltd., Dublin, Ireland) documents to co-develop information on which the joint work was built.

Collaboration with leaders of the ELViS platform (JRA1) was important to ensure alignment in terms of technical and data compatibility. A joint kick-off meeting (Milestone 25 'Expert Session for establishing technical structure, basic criteria and design') between JRA1 and NA2 took place in July 2019, at the start of both work packages, to identify common needs for ELViS and the CDD, particularly with regards to the Collection Classification Scheme, and data standardisation. In order to maintain alignment throughout the lifetime of the two tasks, leaders organised ad-hoc meetings that addressed specific technical issues.

Special engagement was sought from the CETAF Earth Science Group (ESG) looking for expertise to support the definition of aspects of the Collection Classification Scheme in relation to geology, palaeontology and extraterrestrial materials. Similarly, close communication was kept with the TDWG Collection Descriptions (CD) group via meetings and workshops (see Table 1 for details).

SYNTHESYS+
Synthesis of Systematic Resources

a DiSSCo project

2.3 External contributions and outreach

The CDD was promoted at international conferences via talks in order to create awareness of the concept in the wider community, to receive feedback and to promote the Collection Classification Scheme and the data standards that underpin the CDD. In the longer term there is an opportunity for the CDD to be adopted at a global scale, to scale up the breadth of the CDD beyond the community participating in SYNTHESYS+ (Table 1).

Table 1. A list of the main workshops and conferences attended

Conference/Workshop	Description	Purpose for attending
Biodiversity Next - Leiden, October 2019	International conference	CETAF Presented title '*Collections Digitisation and assessment dashboard: A tool for supporting informed decisions*' (doi: 10.3897/biss.3.37505), and attended workshops to gain knowledge about the processes of data standardisation and the global NSCs data landscape (e.g. TDWG CD workshop on collection data standards and community use of them).
TDWG CD group Workshop - London, October 2019	A two-day workshop for progressing with the TDWG CD model.	Contribution to the development of the CD model with regards to the needs of the CDD and CETAF Registry of Collections.
MOBILISE ACTION – Warsaw, February 2020	International workshop under COST European Cooperation in Science and Technology. MOBILISE is the COST Action No. CA17106, "Mobilising Data, Experts and Policies in Scientific Collections" (MOBILISE).	(CETAF) Further dissemination of the CDD Collection Classification Scheme and concept via an oral presentation, for feedback from a new audience. Outlining the priority needs of the CDD and SYNTHESYS+ with regards to developments in the CD standard. Working group sessions attended: WG1 Assessment of existing

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

		systems and standards; WG2 Development of standards and guidelines for data gathering and large-scale digitisation of collection objects
TDWG 'BBQs' - online, attended 2 sessions in April and May 2020.	One hour virtual working meeting, whereby participants worked together on the TDWG CD definitions.	To help improve collection description terminology/definitions that are needed for the CDD Collection Classification Scheme, as well as checking alignment.
SYNTHESYS+ T5.1 Online international consultation 'Advancing the catalogue of the world's Natural History Collections', April 2020 (led by Global Biodiversity Information Facility)	The aim of the consultation was to develop a common international vision for the scope, content, and services for a catalogue of world natural history collections.	Presenting the CDD Collection Classification Scheme, promoted the awareness of Earth Science data standard needs. Promoting the CETAF Registry of Collections as a European authoritative data source and how it could be linked and aligned with a global collection catalogue.

3. The Collections Classification Scheme

The Collection Classification Scheme was developed for describing physical collections using high-level categorisation to present the information in a standardised way in the CDD. It will also be the classification used in ELViS to enable full integration of DiSSCo linked tools. The classification scheme presented here is an enhanced version of the preliminary one developed in ICEDIG D2.3 (as mentioned in Section 1.2). Previously, the scheme was identified by conducting a crosswalk analysis of already existing collection related vocabulary in order to delimitate existing terminology that could be used in the improved CDD (van Egmond et al. 2019).

The main dimensions identified in the preliminary ICEDIG scheme have been kept as classification categories: Institution, Taxonomy, Storage, Stratigraphic age and Geographic region. The categories within each of these dimensions are hinged off the main Natural Science disciplines, which are the highest-level categorisation for collections: Anthropology (newly added in the SYNTHESYS+ CDD), Botany, Extraterrestrial, Geology, Microorganisms,

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Palaeontology, Zoology Invertebrates, Zoology Vertebrates and Other Geo/biodiversity. Each dimension of the Collection Classification Scheme is described in detail in the following subsections.

3.1 Institution

For the CDD, it was decided that each institution was to be identified by their official name (in English as documented in GRID) and institution acronym (as defined in GRID or by CETAF Registry of Collections), and 2-digit ISO country code.

3.2 Taxonomy

The Taxonomy dimension describes the collections by taxonomic references, including disciplines and categories, 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 (see Appendix A Table A1). Enhancements were made from the ICEDIG D2.3 regarding the addition of Anthropology specific categories, together with significant amendments to Geology, Palaeontology and Extraterrestrial categories. Also, Mycology was merged into Botany.

3.3 Storage

The Storage classification (see Appendix A Table A2) is considered essential for collection managers, majorly regarding space and facility planning, because it describes how a collection is preserved (e.g. in fluid jars, dried and pinned). Among others, it could also 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 classification Palaeontology, Geology and Extraterrestrial storage categories were enhanced from the ICEDIG references.

3.4 Geographic region

Geographic region refers to where a specimen/object was collected, and not where it naturally occurs in the wild (for example coming from a zoo, a botanic garden or a park). 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 brought in a way to identify the uniqueness of collections on an institutional and country level.

The geographic region dimension has been divided into marine and terrestrial (see Appendix A Table A3). The marine regions are based on the 'International Hydrographic Organisation (IHO) World Seas – Version 3' (Flanders Marine Institute, 2018) (See Appendix A Figure A1 for Marine region boundaries). In relation to the previous ICEDIG CDD, and under the marine realm, further geographic sub-categories to differentiate the North Pacific, South Pacific, North Atlantic and South Atlantic ('deep sea', 'shelf area and adjacent seas' and 'unknown') were added in order to 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) (Brummitt, 2001). In this case, there were no changes from the ICEDIG version of the dashboard.

In general, several categories were added: 'World/NA' for specimens/objects that could not be assigned to a more specific marine or terrestrial regions. A Region-related 'unknown' sub-category was included for objects that had an unknown collection origin.

3.5 Stratigraphic Age

This scheme is specifically devoted to palaeontology collections: In fact, this addressed the fact that palaeontology collections are divided by stratigraphy as well as taxonomy. Also, it adds another level of detail for discovering geodiversity (Appendix A Table A4). The scheme agreed upon follows the standards of the International Commission on Stratigraphy (ICS) (2020).

3.6 Statistics

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

3.6.1 Object count

This is a numeric figure that represents the total 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.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Survey participants 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 lesser 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.

3.6.2 Digitisation levels

As introduced in section 1.2.2, the Minimum Information about a Digital Specimen (MIDS) specification was used to describe the digitisation level of objects in each specimen breakdown. A brief summary of the four levels in the current MIDS specification is shown in Table 2.

Table 2. A brief description of the four MIDS levels (v0.9) (from Hardisty et al. in progress)

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.

Note: Level 0 is equivalent to creating a simple catalogue record containing a physical specimen identifier, such as a barcode number. Level 0 often precedes more complete digitisation steps that yield more detailed information. Hence, level 0 is termed a pre-level. Nevertheless, level 0 data is useful minimum information for advertising or knowing about the existence of specimens.

The data were captured as percentages of the total 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.

4. Data architecture

4.1 Data granularity and aggregation

The first step towards designing the CDD survey was defining the level of data granularity and aggregation requirements. This refers to the extent to which the institutional collections should be broken down according to the four classification schemes or dimensions, and how the schemes 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, the first being where an institution breaks down its collection data by the four dimensions independently. This is the simplest and requires least effort to contribute the data, but it provides low utility of data because questions can only be answered within the individual classification dimensions (e.g. how many objects are from South America, and how many objects are fungi but not how many fungi are from South America). The opposite extreme is based on a combination of all the dimensions into one single breakdown which would allow users to answer any question related to any combination of the classification schemes used, thus generating a high level of data utility. 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, the ICEDIG user stories were analysed to see what combinations were essential. From this analysis, the only combination of schemes 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 by the task participants 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.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Although no other classification schemes were combined in their entirety, the highest level of the 'Taxonomy' hierarchy ('Discipline') was incorporated in each of the breakdowns. 'Discipline' consists of just 9 classifications ('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 ELViS, the CETAF Registry of Collections and GBIF Collections Catalogue. A summary of the four breakdown schemes is shown in Table 3.

Table 3. A summary of the four breakdown schemes used for the CDD dataset

| | | Breakdown schemes | | | |
		1: Taxonomy	2: Taxonomy and Geographic region	3: Storage	4: Stratigraphic age
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

* only applicable to the 'Palaeontology' discipline

4.2 Data acquisition

For the prototype dashboard, a Google Sheet survey was considered to be the best tool to test the feasibility for partners to collate data using the Collection Classification Scheme. This technology was already familiar and available to all partners in the task, and offered flexibility to integrate amendments quickly as requirements evolved as well as basic validation functionality to help to manage the quality of the data in the responses. In addition, this approach gave

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

partners the option of printing or downloading the survey as an Excel spreadsheet and filling it in offline.

A draft survey was first designed and tested with three partners (RBINS, MBG, and HNHM), before a refined version was sent to the other 6 partners. This preliminary stage allowed a first check on the clarity of the guidelines and the manageability of the survey for partners when filling in the data at the granularity requested. The test partners had 5 weeks to complete the survey. From their feedback there was little structural change to be made, but some further clarity in the guidelines was required.

The final survey consisted of 11 sheets: 4 to provide the respondent with information about how to complete the survey, MIDS levels, and participating institutions, 5 for data entry about the institution and collections, and 2 for additional feedback on the process and notes about the collections. Link to final CDD Survey.

Below is a more detailed explanation of the data entry sheets in the google survey.

Institutional information

This sheet collected the institution's name and the country in which it is located, plus its unique acronym (provided by CETAF Registry of Collections) and identifiers (e.g. Index Herbariorum code, GRID identifiers). Where possible, this information was prefilled from existing sources.

Collection overview

This sheet was for recording the size of an institution's collection at the 'Discipline' level, the highest level in the Taxonomy' classification scheme) (e.g. Anthropology, Botany, Extraterrestrial), and the percentage of those objects that have been digitised to different MIDS levels.

Taxonomy classification & Geographic region

This sheet firstly displayed a breakdown of the main disciplines into more detailed taxonomy categories. Secondly, it displayed a matrix for institutions to enter data on the size of collections (Taxonomy classification) broken down by Geographic regions. MIDS were to be provided for the Taxonomy classification breakdown only.

In order to solve the problem of differing abilities between institutions to provide this level of granularity (i.e. Taxonomy defined by Geographic Region) an 'unspecified' category was

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

added. It was mandatory for institutions to fill in data to at least discipline level (e.g. Anthropology: Unspecified, Botany: Unspecified), with these compulsory fields highlighted in red. Institutions were strongly encouraged to fill in to the highest granularity possible to increase the utility of the final data shown in the CDD.

Storage

This sheet presented the storage collections as collected in the Collection Classification Scheme without any other dimension combination to minimise the burden on the data provider and ensure that the task was feasible. MIDS level categories were added to the Storage classification sheet because it was considered both useful for digitisation prioritisation and for wider decision making (as different storage types may require different techniques or equipment).

Stratigraphic age

This sheet contained the Stratigraphic age classification (Appendix A Table A4), without any combination with other dimensions. In this respect, some issues were highlighted since some objects may span multiple time periods. An attempt to mitigate this concern was for partners to enter data into the 'Any...' category for the level above in the hierarchy. For example, if objects in a collection have a span from Paleocene and Eocene, these could be quantified as 'Paleogene - Any epoch'. However, partners were encouraged to provide further information on this subject (i.e. percentages of how many objects within the collection belong to each of the periods/epochs) in the 'Collection Notes Sheet' to help with deeper understanding and further solutions. MIDS levels were also included in this sheet.

Collection Notes

The collection notes sheet allowed partners to provide any additional information on their collections that could not be reflected by using the breakdown categories.

Contributor Feedback

The collection feedback sheet presented questions asking partners about their experience with collating the data and filling in the survey, and any suggestions for improving the process. The questions and responses are shown in Appendix B.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

4.3 Data structure

The data model underlying the dashboard was designed with close reference to the data standard and model in development by the TDWG Collection Descriptions Data Standard Task Group. 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 Figure 1 below.

SYNTHESYS+
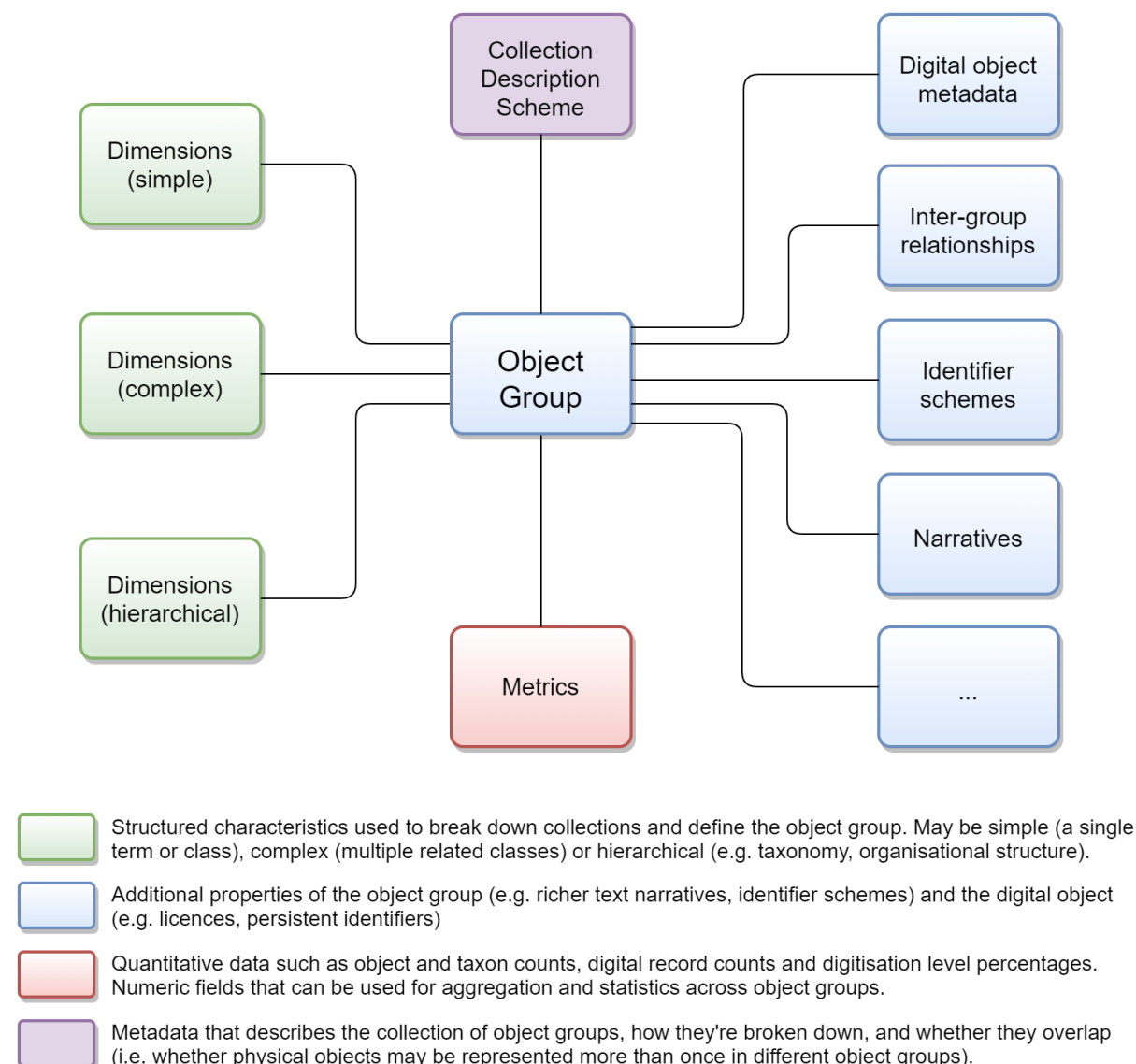Synthesis of Systematic Resources   a DiSSCo project

Figure 1. A simplified conceptual view of the TDWG CD data model.

In the CDD database implementation, the classification schemes are represented as dimensions, and the Collection Description Scheme construct is used to differentiate between the multiple breakdowns of each institutional collection according to the different classification schemes. This prevents the same object from being counted more than once in any of the dashboard visualisations.

Representing the CDD collections as a set of 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

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

across) dimensional hierarchies like Taxonomy and Geographic region. This is an important feature of the data structure for supporting the CDD's breadth of visualisations user cases.

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 Appendix C.

4.4 Data processing

Data was 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.

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 classification schemes, reflecting the challenge for institutions in generating and collating this information within the allotted time frame. 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 was excluded until the issues could be resolved with the contributing institution.

5. Systematic design of the CDD

The systematic design of the CDD refers to the process by which the collated institutional data was visualised, investigated and tested by the Task partners in order to ground truth requirements and frame the data in the most effective manner possible. This comprised:

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

- Deciding how the different data elements and themes should be aggregated and structured as pages.
- The granularity and precision of data.
- The most appropriate number, choice and combination of metrics, charts and maps to use, in relation to the focus of the page and underlying data.
- The range of filterable fields and operations that should be made available to the end user.

5.1 Technologies

The CDD prototype dashboard was built using Power BI, an enterprise analytics, business intelligence and data visualization platform in the Microsoft 365 product ecosystem. A free version of Power BI is available, but excludes functionality essential to the requirements defined for the CDD (e.g., publication to an openly available URL). As such, the Natural History Museum (NHM) London's Power BI enterprise-level Pro subscription was used to build and host the CDD prototype. A copy of the CDD codebase (.pbix) will be deposited in the DiSSCo github repository, to aid reproducibility within another Power BI instance.

It is worth noting that the publication mechanism by which the CDD is made accessible online to all stakeholders, bypassing authentication, also mandates some loss of functionality that would be available if viewing the CDD within the Power BI Service, which is a cloud-based, self-serve Business Intelligence (BI) and analytics platform accessible to Power BI license-holders. Several unmet CDD requirements pertinent to this curtailed functionality are discussed in more detail in section 7.5.

5.2 Design considerations

The scope and aims of T2.2, as defined earlier in this document, state that the SYNTHESYS+ CDD deliverable should serve as a fully-functioning pilot dashboard, with the potential to expand to cover as many DiSSCo-participating institutions as are able to supply the required data. During the early stages of the CDD design process, it became apparent that in some areas there was friction between these two goals:

Lack of temporal ('change over time') data: a dashboard that displays progress of digitisation activities over time was a core requirement expressed in the CDD user stories. The data available to the CDD during the pilot project was a snapshot of each institutional collection at the point of survey completion. Rather than attempt to include visual elements and metrics

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

appropriate for time-series data at the expense of dashboard usability and appearance, a decision was made to focus on snapshot data and categorical visualisations for the prototype, while still collecting and recording additional requirements relevant to change over time data.

Scale: Data from six institutional collections were available to feed the CDD prototype, whereas the upper bound of DiSSCo members is currently in the region of 120 institutions. In order for the CDD prototype to function as a standalone entity, compromises had to be made between designing around the available data vs. designing against potential future data. Filters and other operational dashboard elements used grouping levels suitable for much higher participant numbers (e.g., enabling users to easily view data for a single country) while still providing functionality suited to the scale of the available data (e.g., also enabling users to easily view data for a single institution).

Limited visualisation choice: there are a limited number of visuals available in Power BI out of the box; the available visuals also vary widely in terms of configuration options. User-created visuals are available for Power BI, but these were avoided wherever possible due to concerns about the long-term sustainability and support of non-Microsoft visuals within Power BI.

User profile: A recommendation of the ICEDIG CDD-scoping project was that CDDs should focus on intuitive functionality and avoid more complex end-user operations like drill-downs, outlier identification, customizable parameters, etc. This steer was adhered to during SYNTHESYS+ CDD design and development; as a result some partner feedback was not incorporated. Where possible, complex functionality was included and masked from the end-user perspective (e.g., hard-coding a single-click button to remove unwanted data elements from a visual, rather than making available the more configurable but more complex filtering options).

5.3 Iterative design process

The systematic design was completed in the following phases:

Phase 1 – Review existing user stories against CDD scope and specification

The following sources were consulted during the collection of CDD requirements:

1. ICEDIG user stories: this list comprised the foundation of CDD user stories and was the primary source of requirements

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

2. Task partners and JRA1 leaders were asked to look over the ICEDIG list and add any missing information or requirements

3. Two user requirements-focused DiSSCo Prepare task groups were contacted for input: Task 1.1 'Analyse life sciences use cases and user stories' led by the Finnish Museum of Natural History (LUOMUS); and Task 1.2 'Analyse Earth sciences use cases and user stories' led by Museum für Naturkunde (MfN).

The resulting list of 40 user stories was evaluated against the prototype dataset; 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 fill the requirement was not available: change-over-time data, granular taxonomy, collection usage, associated research.

To inform the high-level CDD structure, 22 in-scope user stories were grouped into 5 themes (Table 4) 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).

Phase 2 – Landscape analysis: existing collection dashboards, style and structure
Partners were presented with existing dashboards from other initiatives so that they could reflect on the many possibilities for visually displaying data in the CDD and provide suggestions on their preferences. They were also asked to provide feedback on other specifications in terms of styling, branding and accessibility (e.g. whether having a mobile view would be important).

Phase 3 – Prototyping
An agile approach was adopted for building the CDD, carried out by NHM. Each week during the period 23rd June to 20th July 2020, a new version of the CDD was released and partners were asked to provide feedback by the end of the week. Amendments/changes were then incorporated in the next version, wherever possible.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

The first three versions were shared as Google Drive PDFs, which allowed task leaders, partners and NHM staff responsible for developing the dashboard to highlight/comment/discuss highly specific areas of the dashboard while keeping all feedback in one place. A static view of the dashboard was used in these early versions to focus the feedback on 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 user-acceptance test interactive elements and check performance and display across different systems.

Phase 4 – Non-functional requirements

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, which were not shared more widely for comment: while the changes within each were non-trivial, they focused on incremental improvements to the user experience (e.g. formatting, performance and accessibility), enabling partners to spend their consultation time on more substantive issues.

Following this phase, a final live prototype was shared with the wider SYNTHESYS+ network by sharing the link via a dedicated message thread in the Teamwork platform in which feedback/responses could be collected.

Table 4. Themes based on the user story data needs

Theme	Detail	CDD page	Examples*
1. Find something specific	Similar to search queries: define several parameters and be presented with data that fulfills 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	Priority is on a view of the data at a combined/consortial	Overview	'I want to be able to identify gaps across DiSSCo digital collections,

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

view of DiSSCo collections	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.		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'

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

6. Description of the pilot CDD deliverable

This section showcases the final pilot CDD and provides an explanation of the features and functionalities of each page. The live and interactive SYNTHESYS+ Pilot CDD is now published online for partners to explore: **Collections Digitisation Dashboard**.

The pilot CDD includes data from 6 partner institutions (MBG, UCPH, MfN, HNHM, MNCN, NHM) that are aggregated and organised within three pages based on the themes defined during the systematic design. The data is graphically displayed using multiple impactful and appealing visuals (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 on each page (via an 'i' in the top right corner icon see Figure 2) which explain the project background, 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.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

First Page: Collections overview (Figure 2)

The first page addresses Theme 3 'see an aggregated view of DiSSCo collections'. This page provides an aggregation of data on the total number of objects in a collection, and total number of objects digitised in accordance to the MIDS levels. The user can explore the total size of collections as defined by discipline and Taxonomy categories, geographic origin, geographic origin against the location of holding institute, and by Stratigraphic age. The data can be filtered by country and/or institution, thus addressing Theme 4 of the user stories. More information about a specific collection is given when a mouse cursor is hovered over the item of interest (see Figure 2 for an example).

Second Page: Collections Comparison (two or more institutes) (Figure 3)

The second page 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. These comparisons are also given as actual numbers within a summary table at the bottom of the page.

Third Page: Collections Location (Figure 4)

This page addresses the user stories under Theme 1 'Find something specific'. It allows the user to locate collections based on storage, digitisation level, taxonomy, geographic region and stratigraphic age (for palaeontology only). As mentioned in section 4, only the taxonomy and geographic region classifications were combined, thus apart from these two classifications, users can only infer information and not exactly pinpoint a collection that is of a certain taxonomy, 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 views of the possible combination choices. For example: 'Discipline, Storage and digitisation level' and 'Discipline, Taxonomy and Geographic Origin'. When a user clicks on one of these views, they can further filter the chosen categories. 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 a various 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 (as highlighted in Figure 5).

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project
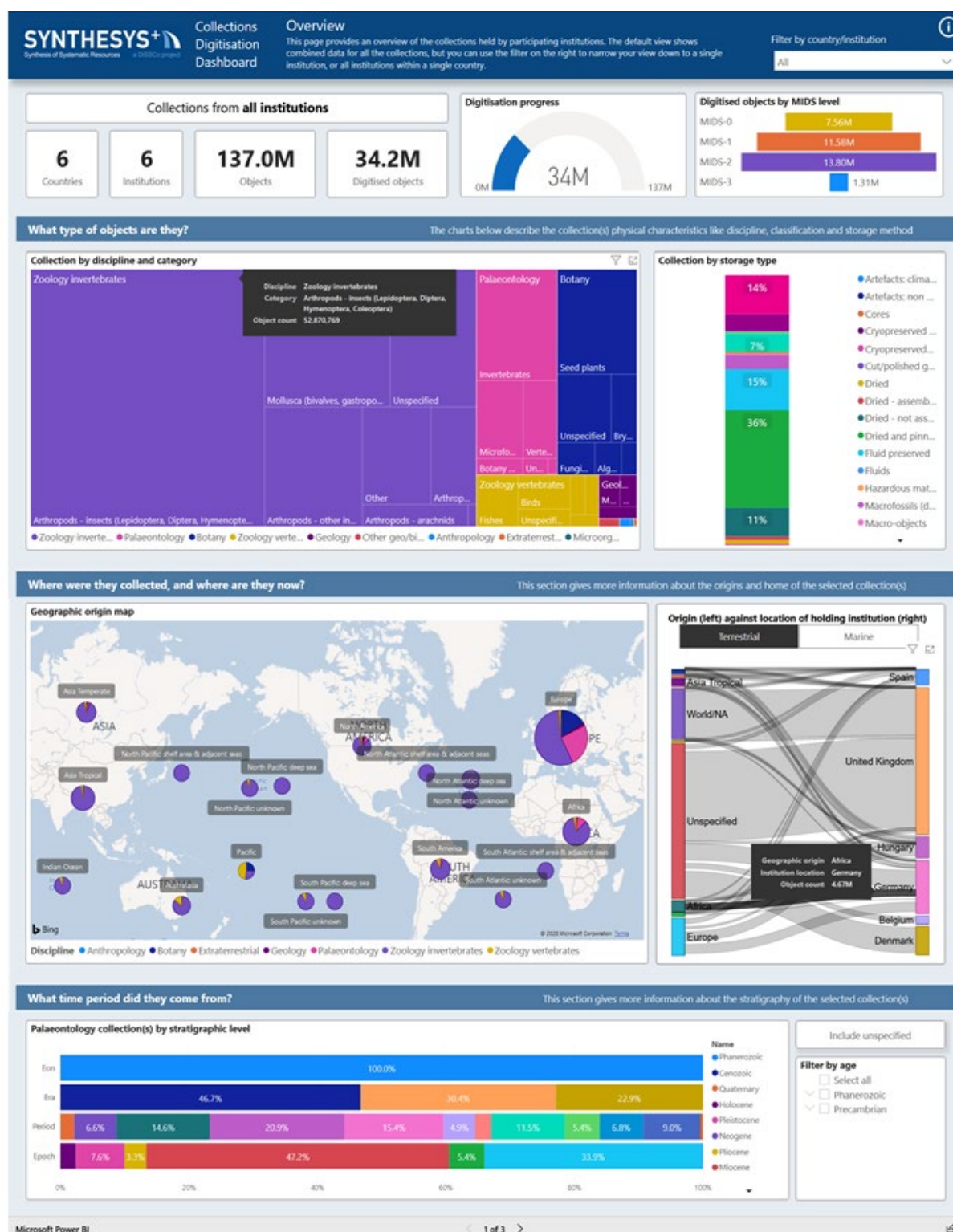
Figure 2. First page of the Pilot CDD showing a collection overview.

SYNTHESYS+\
Synthesis of Systematic Resources   a DiSSCo project
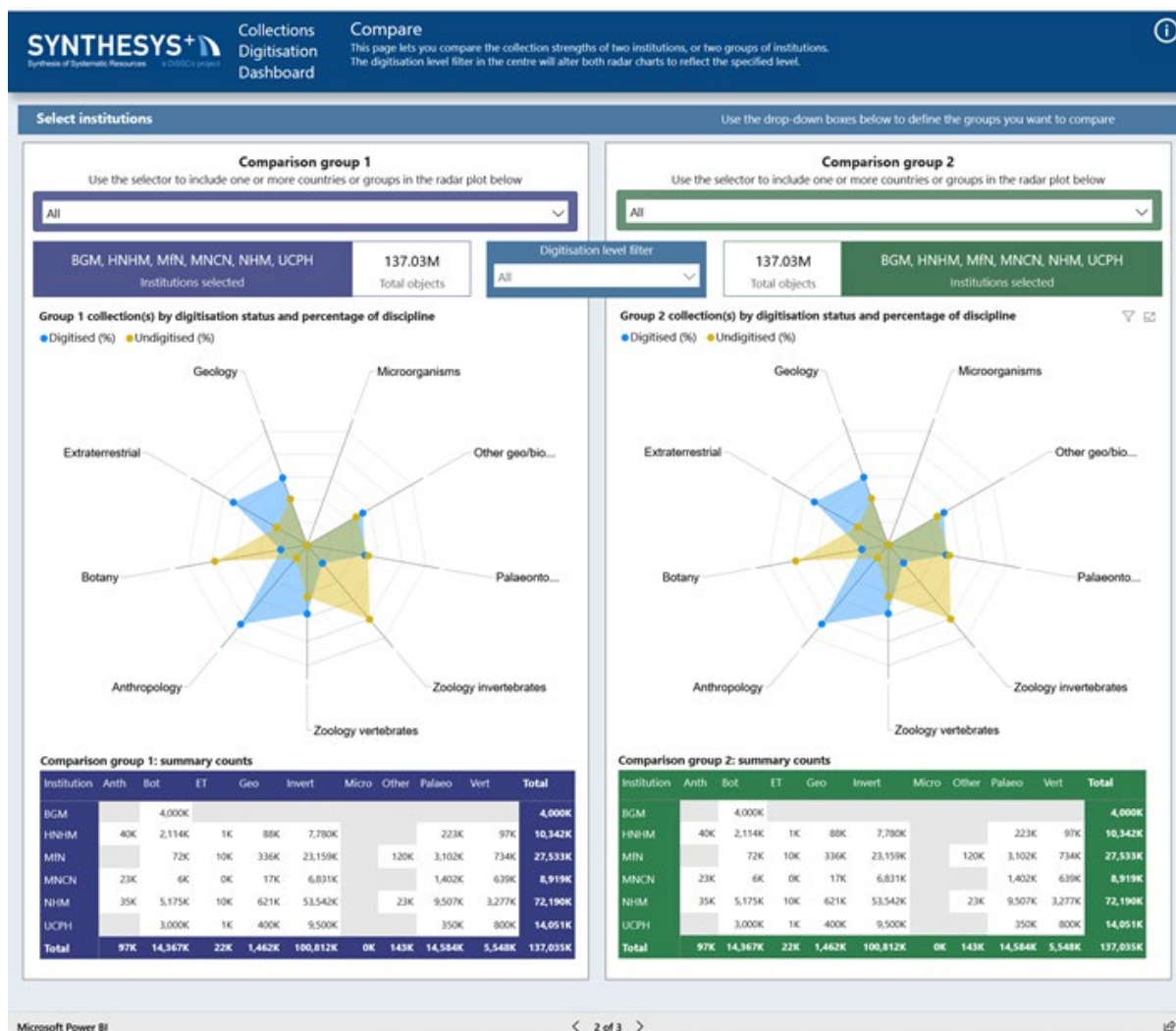
Figure 3. Second page of the CDD which provides a comparison view between the different institutes.
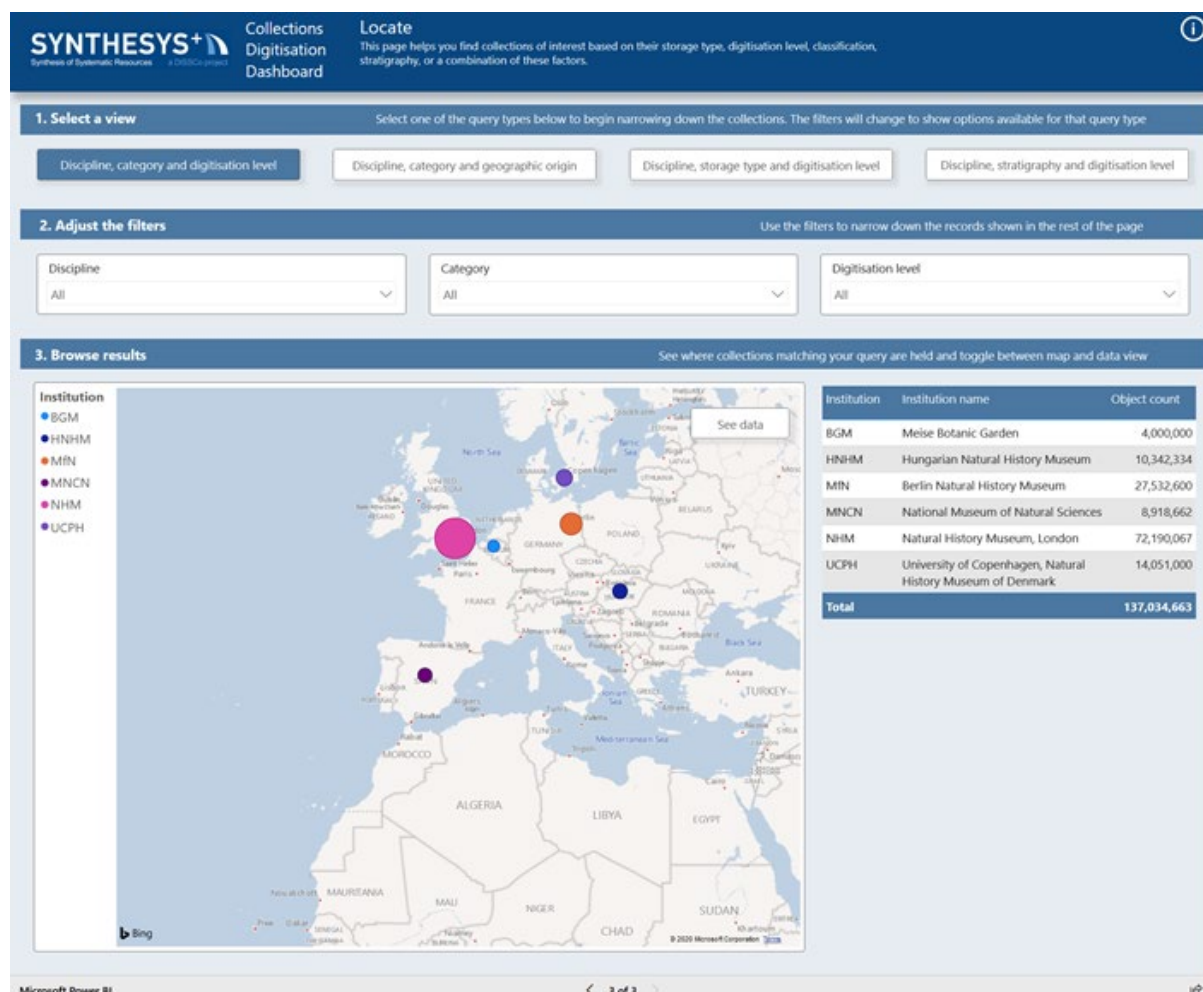
Figure 4. Third page of Pilot CDD, which helps users find the location of collections in Europe based on criteria: Storage, Digitisation level, Discipline, Taxonomy, Stratigraphy.
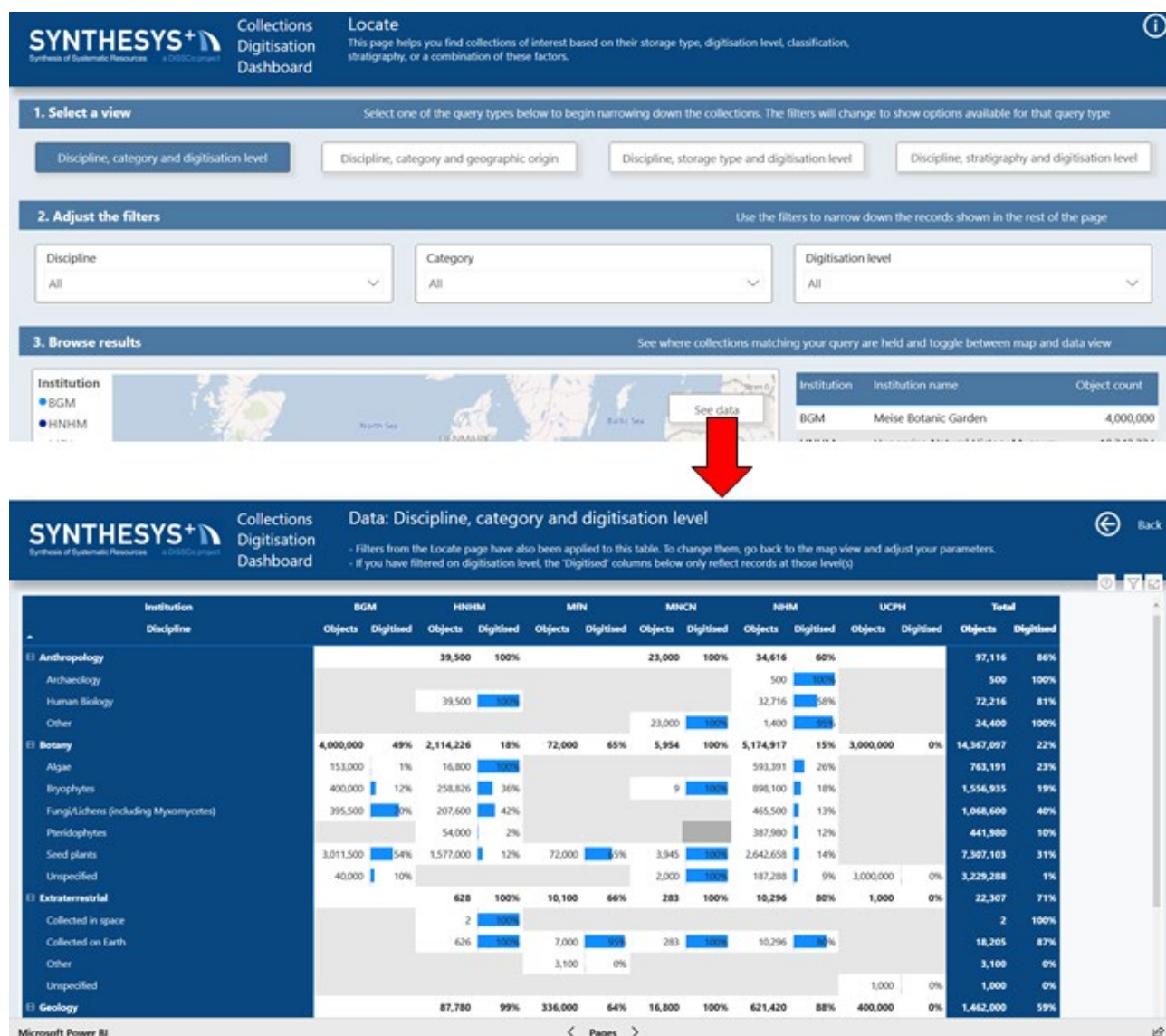
Figure 5. Presenting the 'See Data' button (as highlighted by the red arrow) feature on the third page of the Pilot CDD, which leads users to a separate page that provides actual object numbers for the different classification dimensions.

7. Conclusions and next steps

SYNTHESYS Task 2.2 has delivered a fully functional pilot **Collections Digitisation Dashboard** (CDD), based on standard compliant high-level data from seven NSCs. This work has also led to the enhancement of a Collections Classification Scheme, initiated within ICEDIG deliverable D2.3 that is being further developed through the TDWG community. Despite this achievement as expected under this concrete Task T2.2, a number of issues remain outstanding 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. In these concluding remarks, we review these issues and highlight next steps in the development of a complete DiSSCo Dashboard.

7.1 Source data collation

The trade-off between gathering structured high quality data about a 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 is 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 provision the data, all compound to make what from the outset seems a simple problem, into a complex and potentially insurmountable challenge.

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, specialties, 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. This has the potential to act as a unifying global framework holding baseline information on global collections, and also linking to information services relating to collections.

As part of SYNTHESYS+ Task 5.1, GBIF organised a global community consultation to examine the issues associated with the use, information, technology and governance of a global collections catalogue (see Hobern et al. 2020). Recommendations from this consultation (see

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

integrated summary of the online forum discussions) 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 process of redeveloping the CETAF Institution Profiles as the CETAF Registry of Collections. This has the potential to provide a more automated approach to provision 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 on the DiSSCo ECOI (European Collection Objects Index) holding 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 (Figure 6a,b), and 'machine readable' to facilitate data exchange and harvesting of data. In this regard DiSSCo ECOI has the potential to become a stable source for information for European collections feeding into the GBIF Collections Catalogue.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Figure 6a. A preview of a page in the CETAF Registry of Collections which displays the hierarchy of collections and the information units for each collection / sub-collection, to demonstrate the levels of information that the registry can hold and how it is organised.

Figure 6b. A preview of a page in the CETAF Registry of Collections which displays information on the Royal Belgians Institution of Natural Sciences Amphibian collection, to demonstrate the levels of information that the registry can hold and how it is organised.

7.2 Interoperability with other collection descriptions data initiatives

TDWG Collection Descriptions Data Standard

As described in section 4.3, the prototype CDD data model has been designed in alignment with the development of the TDWG CD data standard and model. 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. The said connection also provided an opportunity for the requirements of the CDD to feed into the design of the TDWG standard.

As the CDD and standard was being developed in parallel 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 9 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 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.

7.3 ELViS as a major use case for the CDD

The European Loans and Visits System (ELViS) is intended to be a major beneficiary of the CDD, as users need access to information about the physical and digital holdings of NSCs in order to plan visits and arrange loans of material. Throughout the development process of ELViS a series of user needs emerged that go beyond the capabilities of the current CDD. These needs have been collated by JRA1 as user stories and are absent from this pilot CDD because the data was not yet available to support them at the time they were needed.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Table 5. Specifications for ELViS with examples of user stories provided by JRA1 (directly cut and pasted without amendments).

Specification	Example User story
Granular views of request, usage (reports and visualisation)	As an Administrator of the ELViS system I want to see in a dashboard - the status of all requests over a certain period, per institution, per researcher/ELViS requester and be able to sort and filter that information.
Customising the view (based on the audience)	As a Researcher I want to be able to see at a glance which depositories hold material of the group I am working on so that I can save time/efforts to get this information.
	Create reports on different aspects of the requests/transactions of all the in the participating institutions in the ELViS system (what kind of reports are used/needed*?), so I can provide detailed overall information on the status of the core business of the ELViS system at all times.
Digitisation demand (from dashboard view, easy visual way to determine the needs)	As an administrator I want to know the demand of digitisation of specimens so that I can evaluate the resources needed.
Compare data from different institutions (loans and visits)	As a Curator I want to compare my statistics with the other institutions (at least average values) so that I can compare the situation of my institution and make report to my general director.#
Display legal information (e.g. Nagoya Protocol, ABS requirements)	As a Researcher I want to see what kind of conditions an institution has for loans (Nagoya related or other), so that I can make sure I am complying with the conditions in my study and publications.
	As a Scientist I want to easily identify which genetic resources I can directly request from a natural history collection so that I can fullfill my due diligence obligations under the EU ABS regulation when requesting tissue or DNA-samples from an *ex-situ* situation inside Europe. For this I need not only information on the sample, but also on the status and reference on the access status without legal doc scan in the www!

These additional requirements involve much richer data that is likely to be available via the CDD. For example, data on current loans or the legal constraints affecting the use of certain specimens, is likely to be only held within institutional collection management systems. Achieving this level of integration for ELViS is potentially complex and is arguably not possible using current technologies for the majority of European NSC.

7.4 Alterations and additions to the classification scheme

Within the time frame of SYNTHESYS+ Task 2.2 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 from a recently joined member, Rachel Walcott, who is a principle curator of Earth Systems at the National Museum of Scotland. 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. Table 6 and Table 7 present new proposed categories for mineralogy within the Taxonomy and Storage classifications to reflect this recommendation. Also, suggestions for new categories for Extraterrestrial are also presented in Table 6. For geology it was suggested that the category 'Petrology' would be better replaced by 'rocks'; and loose sediment replaced with sediment.

Table 6. New suggestion for Mineralogy and Extraterrestrial within the Taxonomy Classification (Rachel Walcott personal communication June 2020).

Discipline	Category
Mineralogy	Minerals
	Gems
Extraterrestrial	Terrestrial finds/falls
	Terrestrial Impacta
	Sample Returns

SYNTHESYS⁺
Synthesis of Systematic Resources
a DiSSCo project

Table 7. New suggestions for Mineralogy categories within the Storage Type Classification Scheme (Rachel Walcott personal communication June 2020).

Discipline	Storage type
Mineralogy	Cut/polished gems
	Powder in vials
	Radioactive
	Humidity controlled containers
	Asbestoform in Perspex boxes

Along with other improvements in the classification scheme (as those mentioned for geodiversity), the future development of an age classification for Anthropology collections was also 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. 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.

7.5 Future data needs and investigation of alternative software

Table 8 presents feedback and requirements received from partners during the agile build and design of the CDD. These could not be incorporated due to the limits on the data collected and MS Power BI's licensing model that controls the publishing and implementation mechanisms used for the CDD. This feedback 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.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Table 8. A collation of partner feedback received during the agile build of the CDD, and the reasons for not incorporating these due to data or technology limitations.

Area	Description of requirement	Issue
data	'For institutional view, should show institutional logo relating to selected institution'	We do not have copies of each institutional logo: which would need to be of similar resolution, size etc. The CETAF Registry of Collections will provide those.
data	'Need to be able to see 'progress' in digitisation: change over time in MIDS levels and overall 'digitized' record count'	As only the initial collection of base data was within the scope of task NA2.2, we don't yet have data representing multiple time points. This data will need to be incorporated into future development of the CDD, along with corresponding enhancements of the data model.
tech	'Need to be able to export the underlying data for local work or analysis'	Not possible in the Microsoft Power BI web view. For this to be possible for the prototype, we would need an additional storage/access solution to hold data downloads and a mechanism to keep data in those sheets versioned and up to date.
tech	'Data/dashboard needs to be citable and versioned'	Not possible in the Microsoft Power BI web view. For this to be possible for the prototype, an alternative storage/access solution and the infrastructure to keep data in those sheets up-to-date, plus a data-versioning process to enable citation of a particular incarnation of the data and/or dashboard.
tech	'Should include a link-out to ELViS in the future to help users with more complex requirements/queries'	Pending ELViS development and implementation.
tech	'Institutions should be able to embed a pre-filtered view of the overview page (e.g pre-filtered to their institution)'	Not possible to do dynamically in the Microsoft Power BI web view, although it might be possible with a front-end developer using the Power BI Software Development Kit (SDK). In the shorter term, it's possible to create a separate copy of the overview page for each institution drawing data from the same database as the main dashboard, and this will be explored. However, this approach creates significant overheads in maintenance of the dashboards, and would not scale beyond a small number of institutions.

SYNTHESYS⁺
Synthesis of Systematic Resources   a DiSSCo project

7.6 CDD publication, maintenance and support

The live CDD is published through the Microsoft Power BI service, and can be directly and openly accessed through a unique url. It can also be incorporated into web platforms like the DiSSCo website (https://dissco.eu) and the ELViS portal either as a simple link, or by embedding in iframes to make it visible and accessible within the page itself.

Following SYNTHESYS+ procedures for a deliverable discoverability, the CDD will be published under open licences. This will be further integrated into the DISSCo repository following the DiSSCo Management Plan.

Further details on the management and maintenance scheme will need to be arranged under the overall DiSSCo structure.

Acknowledgments

Author Contributions

curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Ana Casino:** Conceptualization, Project administration, Supervision, Writing - original draft, Writing - review & editing. **Wouter Addink:** Conceptualization, Investigation, Writing – review & editing. **Fredrick Berger:** Investigation. **Anne Bogaerts** Investigation. **Sofie De Smedt** Investigation. **Sharif Islam:** Conceptualization, Investigation. **Patricia Mergen:** Conceptualization, Investigation, Writing – review & editing. **Anne Nivart**: Investigation. **Beata Papp:** Conceptualization, Investigation. **Mareike Petersen:** Investigation, Writing – review & editing. **Celia Santos:** Conceptualization, Investigation. **Vince Smith:** Writing - review & editing. **Patrick Semal:** Conceptualization, Investigation, Writing – review & editing. **Edmund Schiller:** Conceptualization, Writing – review & editing. **Karin Wiltschke:** Conceptualization, Writing – review & editing.

References

Brummitt, R. K. (2001). World Geographical Scheme for Recording Plant Distributions, Edition 2. Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/109.

Cohen, K., Finney, S., Gibbard, P. and Fan, J.-X. (2013, updated). International Chronostratigraphic Chart v2016/04. International Commission on Stratigraphy. Available from: http://www.stratigraphy.org/ICSchart/ChronostratChart2020-03. [Accessed 10 March 2020]

Flanders Marine Institute (2018). IHO Sea Areas, version 3. doi: https://doi.org/10.14284/323.

Hardisty, A.R., Addink, W., Mathias, D., Groom, Q., Haston, E., Glöckler, F., Paul, D., Petersen, M., Saarenmaa, H. and Güntsch, A. (in progress). Minimum Information about a Digital Specimen.

Hobern, D., Asase, A., Groom, Q., Paul, D., Robertson, T., Semal, P., Thiers, B. and Woodburn, M. (2020). Advancing the Catalogue of the World's Natural History Collections. v1.0. Copenhagen. doi: https://doi.org/10.15468/doc-wnsx-ep77.

van Egmond, E., Willemse, L., Paul, D., Woodburn, M., Casino, A., Gödderz, K., Vermeersch, X., Bloothoofd, J., Wijers, A. and Niels, R. 2019. Design of a Collection Digitisation Dashboard. doi: https://doi.org/10.5281/zenodo.2621055

Appendix A: Collection Classification Scheme

Table A1. Taxonomy classification

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
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

Table A2. Storage classification

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 - 80C	
			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
			Microscopic slides	Microscopic slides
			Cryopreserved/frozen 80C	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 -80C	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 -80C	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.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

			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.
			Unspecified	
Geology	Geology	Geology	Macro-objects	Hand specimens / hand-held objects / slabs that can be contained in standard units

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

				(draws, shelves, cabinets). E.g. 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.
			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. e.g. crushed rocks, other microscopic prepared objects (e.g. SEM stubs) etc.
			Unspecified	
Extraterrestrial	Extraterrestrial	Extraterrestrial	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 doesn't fit the above
Other geo / biodiversity	Other geo / biodiversity	Other geo / biodiversity	Other geo / biodiversity	

Table A3. 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	
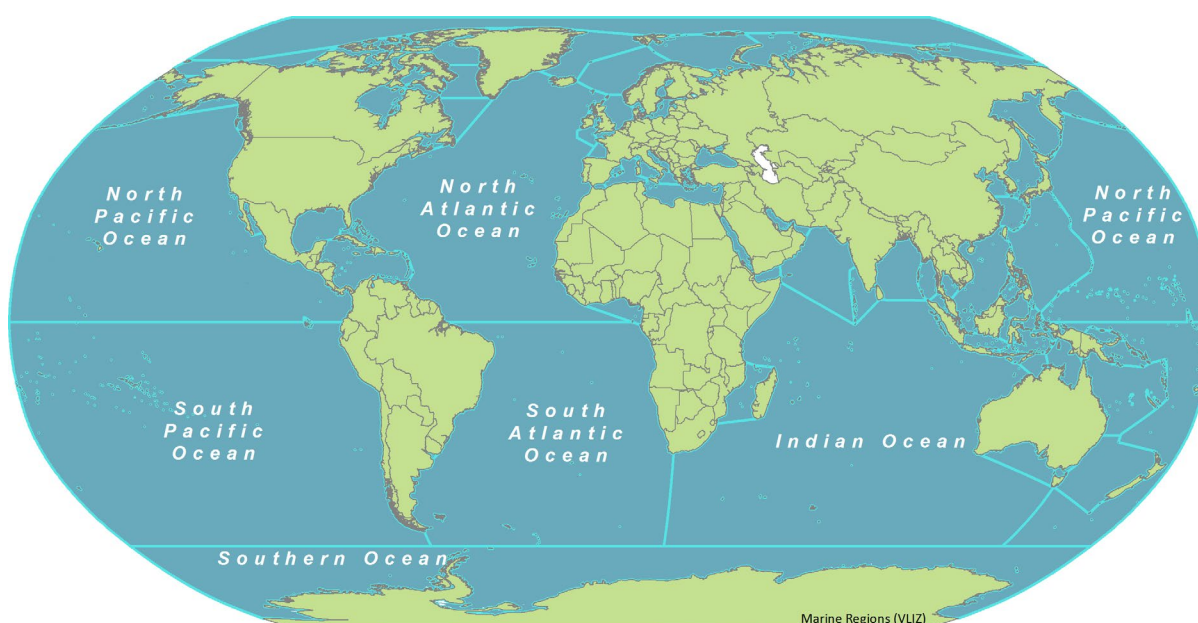
Figure A2. Map showing IHO (World Seas – version 3) marine regions used in the Geographic region classification, and the adjacent seas that occur next to the region boundaries (Flanders Marine Institute 2018).

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project
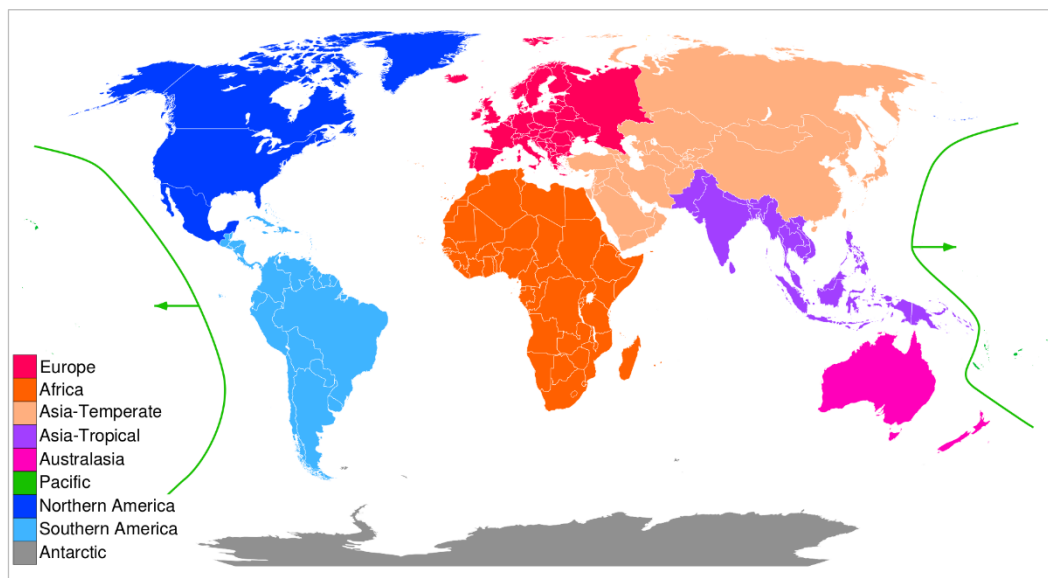
Figure A2. Map showing theTDWG terrestrial (WGSRPD - level 1) regions used in the Geographic region classification,

(https://commons.wikimedia.org/wiki/File:WGSRPD_World.svg).

Table A4. Stratigraphic age classification

Eon	Era	Period	Epoch
Stratigraphy unspecified			
Any era			
	Any period		
Phanerozoic	Cenozoic		*Any epoch*
		Quaternary	Holocene
			Pleistocene
			Any epoch
		Neogene	Pliocene
			Miocene
			Any epoch
		Paleogene	Oligocene
			Eocene
			Paleocene

		Any period
Mesozoic		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	
	Meso-archean	
	Paleo-archean	
	Eo-archean	
Hadean		

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Appendix B: Survey feedback

1. Were the survey guidelines/instructions clear? If not please state why.

HNHM   It is understandable in most parts. The MIDS instructions are still not completely clear. See below.

CSIC   Comments from collectors were the following: Malacology: Not much. The MIDs definitions should be better explained; Entomology: The MIDS level definitions given are not clear. I use those send in the archive "New Mids level for Celia". In general everybody had problems to understand MIDS Level 0.

MBG   Problem with the definitions of MIDS: what about specimens with data but no images. We counted specimens with images & enhanced label data in MIDS 3. MIDS 0 to 2 don't necessary contain an image. We counted the mushrooms, mosses and lichens as dried specimens. We don't have seperate data for Pteridophytes so they are included in the Seed Plants. In the Taxonomic classification Tab. the Object Quantity is in percentages not in numbers. Too difficult to fill out the different fields of the geographical regions. In our herbarium we only use the geographical division: Africa (South of the Sahara), Belgium and General (+ north of the Sahara). The field unknown = dia slides + drawings and watercolor paintings.

MNHN   A lot of collections are counting their specimen as lot not as a single specimen. the dashboard shall introduce these two counting levels which are a reality for collections. for the regions, a already prepare list of countries shall be added. For so we have extracted data from One world collections dashboard, we compiled estimation. So by consequence there is a gap between the different accounts and numbers illustring geographical origin.

2. How was the survey distributed throughout your institute?

HNHM   The survey was sent to the heads of the departments and it was subsequently filled in collaboration with the colleagues working in the departments.

CSIC   By e-mail to all curators and collection managers. Questions had to be answered by mail as well due to the lockdown.

MBG   The curators of BR filled in the survey.

MNHN   2 personnes centralised the data and fulfilled the dashboard. For the storage, we spray the dashboard through the collections managers , around 25 personnes

3. How many people were needed to provide data for the survey?

HNHM   Minimum 6 but if we consider informations obtained from the curators of various collections, it is rather 20.

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

CSIC   Malacology: Four persons. Entomology: One person, helped by a colleague. Herpetology: one person. Arthropoda: one person. Other invertebrates: one person. Tissues and DNA: two persons.
Paleobotany and Fossil invertebrates; Other collections and Added up: one person

MBG   10 people (database managers, curators, biodiversity data scientists

MNHN   around 25

4) How long did it take to obtain data for the survey?

HNHM   Average 8 hours/ collections depending on the databased level of the collection.

CSIC   Three months

MBG   3 half days.

MNHN   more than 2 months

5) Which part/sheet of the survey was the most time consuming to obtain data for? Please state the estimated time taken and reasons why.

HNHM   Filling the "Taxonomy/Classification & Geographical Origin" sheet was the most time-consuming. It took minimum one day to review the inventories in order to categorize the geographical origins. In some cases the geographical origin can be found out by checking the name of the site where it was collected. If we provide exact data, this is the most time-consuming task, because even in the well-databased collections the structure of the databases does not enable easily such searching.

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

CSIC	Malacology: 1. The MIDs. It is not very easy to look for these questions. 2. The distribution by localities. For instance, we do not separate North Atlantic and South Atlantic, the same in other oceans. 3. The malacological collection is ordered by lots, and we have lots with one specimen and lots with 500 specimens. In order to our data, we have changed 1 lot by 6.2 specimens.
Entomology: It has been complicated and difficult to estimate number of specimens according the preservation mode, mainly in specimens preserved in alcohol whose data are known poorly. Also it is difficult estimate the confidence level in geographical provenance.
Break –up data for paleontology in "Taxonomy" sheet. -In paleontology, the "Taxonomy" sheet includes a non-taxonomic break-up, by piece size (that in principle is storage), which implies in several categories adding data from different collections ("Microfossils" includes botany, invertebrates and vertebrates; and "Trace fossils" includes vertebrates and invertebrates). As a result, with this sheet we cannot know the total number of objects of any of the large fossil taxonomic categories (Botany, Fossil vertebrates and Fossil Invertebrates).
Added up data from different collections: -Eg. In MNCN, Tissues and DNA is an independent collection, so that implies that I had to add up data to every zoology taxonomic category; in paleontology for "Microfossils" and "Trace fossils" as I said before; and MIDS in taxonomy where data from several collections must be added up as well.

MNHN	taxonomy and geography

6) Do you have any suggestions for improving the survey?

CSIC	Malacology: For malacology, and other zoological collections, it is very important to separate freshwater from terrestrial. We have put together terrestrial and freshwater. In the case of the Storage data sheet, many lots in our collection (malacology), with the same lot number, are stored in dry, ethanol and microscopic slides; so, the same number of lot, but with many storage ways.
Fossils: Eliminate "size" as a break-up in taxonomy. Eliminate MIDS from storage and stratigraphy.

MBG	It will be very difficult to fill in this survey for institutes who have practically nothing digitised.

7) Are the MIDs level guidelines/definitions easy to follow and apply to your collections? If not please explain why. We would also appreciate any suggestions for improvement.

HNHM	Still it is not easy to follow and not clear, what to do with well-databased (with many metadata) specimens without images. And what are the quality requirements of images? You have to put in the explanation you sent me via e-mail. "For specimens with all/much data but no images, these can be indicated only as MIDS 1 because MIDS 2 and MIDS 3 expect the presence of image(s)."

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

CSIC   Malacology: It would be good to add examples to the MIDs levels. The MIDs definitions should be better explained. Entomology: The initial MIDS level definitions is not clear, because the 0 level includes some characteristics, for example "a list to store any technical metadata such as the quality checks of the digitisation result that have been made" and it seems that this requirement must be in MIDS-1, MIDS-2, ... It is more usual to achieve the requirements include in MIDS-1 or MIDS-2 that those for MIDS-0. The second version of MIDS level definitions is clearer, although MIDS-3 is not developed enough. Fossils: No, because they are not included as a field into databases, and data from different collections must be added up manually. Added-up: for taxonomic collections I have made the effort to include them in the "taxonomy" sheet. It is not possible to include them in "storage", mainly because not all databases incorporate this data. The exception is the DNA and Tissue Collection. Also in the case of Extraterrestrial and Prehistory (Anthropology) because the data is the same as in the "taxonomy" sheet.

MNHN   As we are testing MIDS for the first time, we shall reconsider the definition. For example for MIDS2, not only considering labels but also catalogs or any writing historical record.

8) What CMS does your institute use for collection information? Did you obtain the information manually for the survey or does your CMS allow to retrieve the information automatically?

HNHM   We have partly a Hungarian museological system and a self-developed system. Similar data and statistics are required by our Ministry in the annual reports. Curators have to estimate and provide many various data on their collection. Data of high granularity can be obtained from the curators.

CSIC   Every collection has one or several databases in access, with or without SQL server front end. Some data are in different databases. The information has been obtained querying the whole appropriate databases. Data can be retrieved partially from them, and other must be added up to be included in the correct box. Currently, there are several curators who have not yet joined the museum and do not have access to their databases. I have obtained data from these collections (Geology, Fossil Vertebrates & Prehistory, Fishes, Birds and Mammals) from a book published by CSIC in late 2019, about the MNCN's collections. On the other hand, several curators have not given me the data in the Excel format and I had to do the calculations for the corresponding boxes, which has taken me a long time.
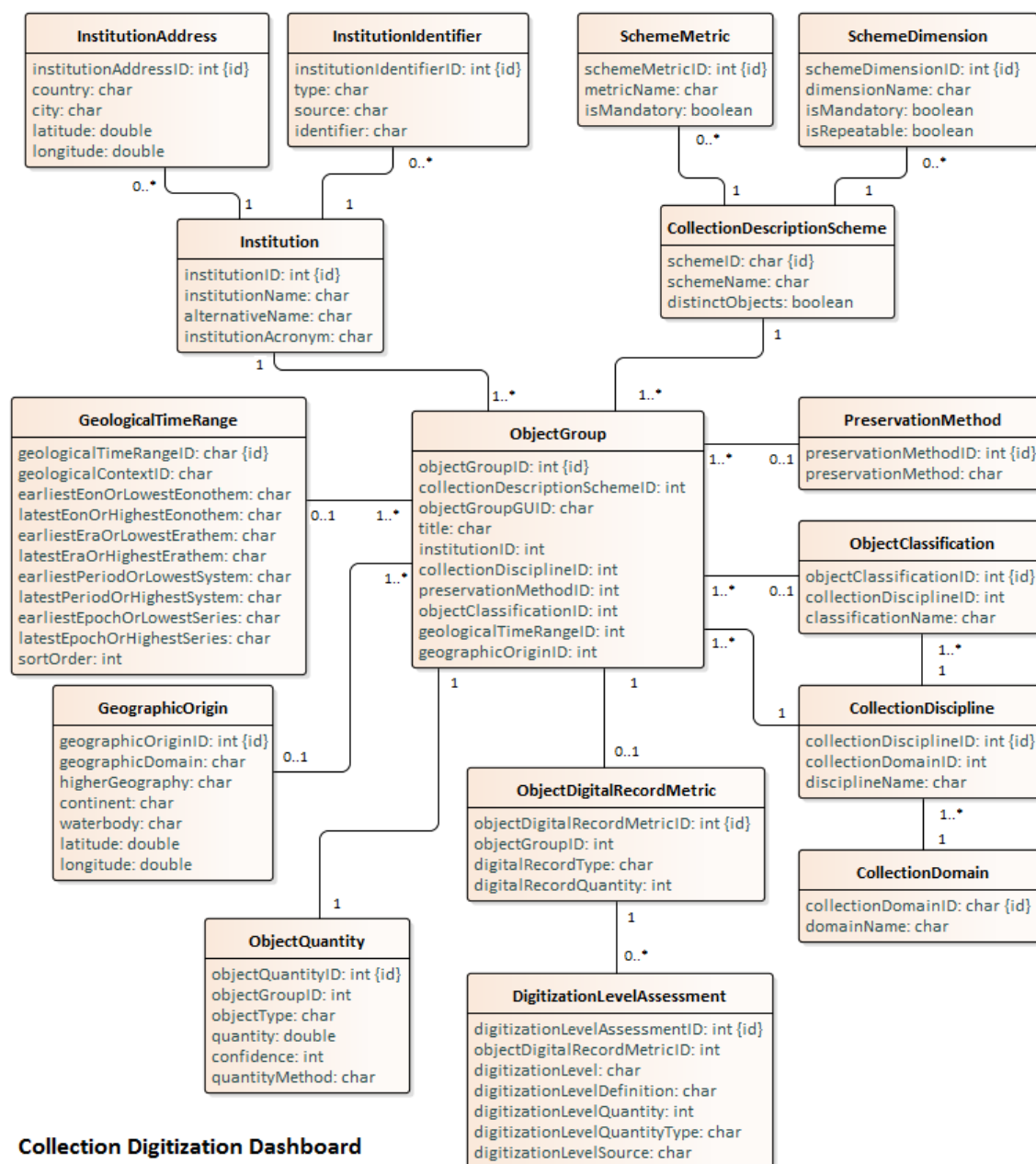
MNHN   Both

SYNTHESYS+
Synthesis of Systematic Resources   a DiSSCo project

Appendix C: The CDD relational data model

InstitutionAddress
institutionAddressID: int {id}
country: char
city: char
latitude: double
longitude: double

InstitutionIdentifier
institutionIdentifierID: int {id}
type: char
source: char
identifier: char

SchemeMetric
schemeMetricID: int {id}
metricName: char
isMandatory: boolean

SchemeDimension
schemeDimensionID: int {id}
dimensionName: char
isMandatory: boolean
isRepeatable: boolean

Institution
institutionID: int {id}
institutionName: char
alternativeName: char
institutionAcronym: char

CollectionDescriptionScheme
schemeID: char {id}
schemeName: char
distinctObjects: boolean

GeologicalTimeRange
geologicalTimeRangeID: char {id}
geologicalContextID: char
earliestEonOrLowestEonothem: char
latestEonOrHighestEonothem: char
earliestEraOrLowestErathem: char
latestEraOrHighestErathem: char
earliestPeriodOrLowestSystem: char
latestPeriodOrHighestSystem: char
earliestEpochOrLowestSeries: char
latestEpochOrHighestSeries: char
sortOrder: int

ObjectGroup
objectGroupID: int {id}
collectionDescriptionSchemeID: int
objectGroupGUID: char
title: char
institutionID: int
collectionDisciplineID: int
preservationMethodID: int
objectClassificationID: int
geologicalTimeRangeID: int
geographicOriginID: int

PreservationMethod
preservationMethodID: int {id}
preservationMethod: char

ObjectClassification
objectClassificationID: int {id}
collectionDisciplineID: int
classificationName: char

GeographicOrigin
geographicOriginID: int {id}
geographicDomain: char
higherGeography: char
continent: char
waterbody: char
latitude: double
longitude: double

CollectionDiscipline
collectionDisciplineID: int {id}
collectionDomainID: int
disciplineName: char

ObjectDigitalRecordMetric
objectDigitalRecordMetricID: int {id}
objectGroupID: int
digitalRecordType: char
digitalRecordQuantity: int

CollectionDomain
collectionDomainID: char {id}
domainName: char

ObjectQuantity
objectQuantityID: int {id}
objectGroupID: int
objectType: char
quantity: double
confidence: int
quantityMethod: char

DigitizationLevelAssessment
digitizationLevelAssessmentID: int {id}
objectDigitalRecordMetricID: int
digitizationLevel: char
digitizationLevelDefinition: char
digitizationLevelQuantity: int
digitizationLevelQuantityType: char
digitizationLevelSource: char

Collection Digitization Dashboard data model

NOTE: Notation such as 0..1 and 1..* on relationships indicate the cardinality of those relationships.

1 - exactly one record. E.g. an ObjectGroup must be attached to an institution, and cannot be attached to more than one.
0..1 - zero or one records. E.g. an ObjectGroup does not need to be attached to a GeographicOrigin, but if it does, it can only be attached to one.
1..* - one or more records. E.g. a PreservationMethod may be used by one or more ObjectGroups, and must be used by at least one.
0..* - zero or more records. E.g. an Institution does not need to be attached to an InstitutionIdentifier, but if it does, it can be attached to more than one.

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project