**Grant Agreement Number:** 777483 / **Acronym:** ICEDIG

**Call:** H2020-INFRADEV-2017-1 / **Type of Action:** RIA

**Start Date:** 01 Jan 2018 / **Duration:** 27 months

**REFERENCES:**

Deliverable **D2.3** / [R] / [PU]

Work package **2** / Lead: **Naturalis**

Delivery date [**M14**]

**Design of a Collection Digitisation Dashboard**

**Deliverable D2.3**

**Version**: 1.0  
**Date**: 31 March 2019

Emily van Egmond (Naturalis)  
Luc Willemse (Naturalis)  
Deborah Paul (iDigBio)  
Matt Woodburn (NHM)  
Ana Casino (CETAF)  
Karsten Gödderz (CETAF)  
Xavier Vermeersch (CETAF)  
Jeroen Bloothoofd (Picturae)  
Agnes Wijers (Picturae)  
Niels Raes (Naturalis)

**Summary**

There is a growing need to set data-driven priorities when planning for the digitisation of European natural history collections. Currently, there is no single location where the required information is gathered and where it can be easily consulted and used by decision-makers and scientists. In particular, the information on digitised and non-digitised natural history collections can inform digitisation-on-demand and mass-digitisation for certain taxonomic or geographic parts of the collection that are not (yet) digitally available. In this Deliverable D2.3 we aim to prepare a preliminary design for a Collection Digitisation Dashboard (CDD), with the main purpose to make European natural history collections visible and discoverable and to highlight the institutional contributions, strengths and weaknesses.

First, we identified six main user groups of the CDD via workshop discussions: a) institutions harbouring natural history collections, b) (non-)professional researchers and collectors, c) education, d) policy makers and financing bodies, e) NGO nature groups and organisations, and f) the wider community interested in natural heritage. User stories were collected and the data elements that belonged to these stories were summarised. The CDD will primarily be used to present high level collection data for communication purposes and as a digitisation planning and data discovery tool.

Secondly, we propose a set of collection classification schemes to be able to describe and characterise a natural history collection at a metadata level. We distinguished a ‘taxonomic’ and a ‘storage’ classification that exist in parallel and are based on a scientific or a collection managers’ view, respectively. For further description of geodiversity collections we identified a third parallel ‘stratigraphic’ classification. In addition, ‘geographic’ and ‘digitisation’ classifications were identified to further characterize the spatial coverage and levels of digitisation of the collections. The most important parameters to be minimally included in the CDD are institution, country of institute, ‘taxonomy’, geography and digitisation.

Based on these requirements we piloted two different CDDs. The first is based on an initial collection survey among DiSSCo partners, and the second is based on a pilot study with Dutch natural collection institutes based on improved classifications.

In this deliverable we have provided a draft on how to create a collection digitisation dashboard to present collection digitisation data and give recommendations on how to proceed from here.

Table of Contents

[1. Introduction 1](#_Toc4962729)

[1.1 Collection 2](#_Toc4962730)

[1.2 Digitisation 2](#_Toc4962731)

[1.3 Dashboard 2](#_Toc4962732)

[1. Requirements for a Collection Digitisation Dashboard 6](#_Toc4962733)

[1.1 Workshops with user groups 6](#_Toc4962734)

[1.1.1 Round Table CDD - ICEDIG 6](#_Toc4962735)

[1.1.2 Dutch collection overview dashboard – NWO-ALW 6](#_Toc4962736)

[1.2 User communities 6](#_Toc4962737)

[1.3 User stories 7](#_Toc4962738)

[1.4 Parameters of interest for the CDD 10](#_Toc4962739)

[1.5 Visual requirements 11](#_Toc4962740)

[1.6 Technical requirements 13](#_Toc4962741)

[2. Collection description standards 16](#_Toc4962742)

[2.1 Task Group Collection Digitisation Dashboards – TG CDD 16](#_Toc4962743)

[2.2 Inventory of collection description schemes 16](#_Toc4962744)

[2.3 Three parallel collection classifications 18](#_Toc4962745)

[2.3.1 ‘Taxonomic’ classification 18](#_Toc4962746)

[2.3.2 ‘Storage’ classification 20](#_Toc4962747)

[2.3.3 Stratigraphic classification (only geodiversity and paleontology) 23](#_Toc4962748)

[2.4 Geographic classification 24](#_Toc4962749)

[2.5 Digitisation classification 26](#_Toc4962750)

[3. Collection Digitisation Dashboard (CDD) 29](#_Toc4962751)

[3.1 Data acquisition and integration 29](#_Toc4962752)

[3.2 Dashboard visualisation 31](#_Toc4962753)

[3.2.1 DiSSCo dashboard 32](#_Toc4962754)

[3.2.2 Dutch national history collections dashboard 34](#_Toc4962755)

[3.3 Final list of parameters to be included in the CDD 37](#_Toc4962756)

[4. Conclusions and recommendations 39](#_Toc4962757)

[Recommendations 41](#_Toc4962758)

[Acknowledgements 42](#_Toc4962759)

[References 43](#_Toc4962760)

[Appendices 44](#_Toc4962761)

[Appendix 1. Details Round Table 44](#_Toc4962762)

[Summary 44](#_Toc4962763)

[Participant list 45](#_Toc4962764)

[User stories 46](#_Toc4962765)

[Appendix 2. Details Dutch collection overview – NWO-ALW 50](#_Toc4962766)

[Background 50](#_Toc4962767)

[Participant list 50](#_Toc4962768)

[User stories 51](#_Toc4962769)

[Appendix 3. Details Task Group CDD 54](#_Toc4962770)

# Introduction

The digitisation of natural history collections has so far been primarily driven by institutional needs, which has resulted in patchy and incomplete digital information on natural history collections, both in Europe and globally (Berendsohn and Seltmann 2010, Blagoderov et al. 2012, Smith et al. 2018). To be able to set priorities when planning for the digitisation of European natural history collections, information about the volume and scope of these collections and their degree and level of digitisation is needed. Currently, there is no single location where this information is aggregated, and where it can be easily consulted and used for decision-making with respect to focussing on future digitisation projects or studying particular of taxa. In particular, the information on digitised and non-digitised natural history collections can inform digitisation-on-demand and mass-digitisation for certain taxonomic groups or geographic regions of the collection that are not (yet) digitally available. Dashboards are useful tools that summarize and visualize this information on natural history collections. Within [DiSSCo](https://dissco.eu/), the Distributed System of Scientific Collections, such a dashboard may not only be used to indicate that digitisation is needed to feed more data into this research infrastructure but also allows strategic choices regarding which collections should be prioritized for digitisation. On a more political level, a dashboard can show the progress of digitisation of the natural history collections in Europe.

A dashboard is expected to be an online tool that gives reliable, complete and up-to-date information on the taxonomic and geographic scope of collections as well as the degree and level of digitisation. It is of great scientific importance to increase the discoverability of non-digitised parts of the collection by providing taxonomic and geographic information about them. Currently non-digitised collections are almost exclusively accessible by taxon (Berendsohn and Seltmann 2010). This information will be key to set priorities for digitisation and see where progress is being made, i.e. a gap analysis can be performed. Ultimately, this is expected to enhance the data availability on past and present biodiversity to support (scientific) research in a wide variety of scientific domains, and allows for strategic planning of research activities.

In ICEDIG deliverable D2.3 we aim to prepare a preliminary design for a Collection Digitisation Dashboard (CDD), with the main purpose to make European natural history collections visible and discoverable. In order to achieve this aim, we define the following subtasks:

1. Identify the main user communities and their user stories for a CDD.
2. Compare the main collection description standards to describe natural history collections related to a CDD.
3. Identify the parameters that are required to develop a CDD that accommodate the needs of a CDD identified through the main user stories.
4. Propose a method to collect and prepare the required data for the CDD.
5. Propose a visualisation of the data in a CDD.

To be clear about the terms ‘collection’, ‘digitisation’ and ‘dashboard’ and their relationship, we indicate below what we mean by these terms in the context of a Collection Digitisation Dashboard.

## Collection

When we use the term ‘collection’, we refer to natural history collections only. Although an institution may distinguish several to many collections within its institution, we will treat all these collections as one, resulting in one collection per institution. For example, there will not be a distinction in the dashboard separating the botanic collection, insect collection collected by collector ‘X’ and insect collection collected by collector ‘Y’ within an institution. These subcollections are all treated as one. A collection is then further subdivided according to 1a) taxonomic, 1b) storage, and 1c) geological levels, 2) geographic regions, and 3) degree and level of digitisation (see Chapter 3). This approach will simplify the data that feeds into the dashboard, both when obtaining the data from institutions and when combining data of different institutions in the dashboard. A dashboard is functional at the level of detail that can be provided by all participating institutions. At this stage that is only feasible at the highest taxonomic and geographic levels.

## Digitisation

In its essence, digitisation is the process of making physical objects digitally available. This can be broadly interpreted and may include textual information on the object itself, an image, or transcribing all information found on a specimen label. For the CDD, it will be crucial to use a clear, unambiguous meaning of the levels of digitisation to be able to easily combine and interpret the visualised data. In Chapter 3 we will further expand on the degree of digitisation and the levels of completeness.

## Dashboard

A dashboard is an online tool that gives a summary of key information relating to progress and performance towards a certain aim (Hetherington 2009). Others such as Few (2006) defined a dashboard as follows:

*‘A dashboard is a visual display of the most important information to achieve one or more objectives; consolidated and arranged on a single screen so that information can be monitored at a single glance.’*

A dashboard is thus usually in a graphical and easy-to-read form. Dashboard are interactive allowing some filtering of the data. Several graphs and/or more textual representations of the data might be placed together. When several screens are needed to show all the data, these are essentially multiple paged dashboards (Few 2006) that are designed to be interpreted by itself. Each screen is designed based on a subtopic of the larger, overall topic of all dashboard pages combined. Depending on the aims, a dashboard may be based on data that is automatically and frequently (e.g. daily) updated, or on data at a fixed moment that shows what has been happening so far (e.g. in the last half year). This type of tool is often associated with managers, who need more general, high level information.

Examples of natural history collection dashboards (Figures 1 and 2), some of which also show the degree of digitisation, can be found in the list below:

* Atlas of Living Australia (Figure 1)

[https://dashboard.ala.org.au](https://dashboard.ala.org.au/)

* Bluegill prototype

[http://fishfindr.net](http://fishfindr.net/)

* CETAF Passports

<http://nhm-informatics.github.io/cetafstats.html>

* Chicago Field Museum

[http://collections-dashboard.fieldmuseum.org](http://collections-dashboard.fieldmuseum.org/)

* Global Biodiversity Information Facility (GBIF)

<https://www.gbif.org/analytics/global>

* iDigBio  
  <https://www.idigbio.org/portal/collections>
* Natural History Museum, London (NHM)

[http://data.nhm.ac.uk](http://data.nhm.ac.uk/)

<http://nhm-informatics.github.io/dcp-external.html>

* Naturalis Biodiversity Center (Figure 2)

[http://bioportal.naturalis.nl/dashboard](http://bioportal.naturalis.nl/dashboard/)

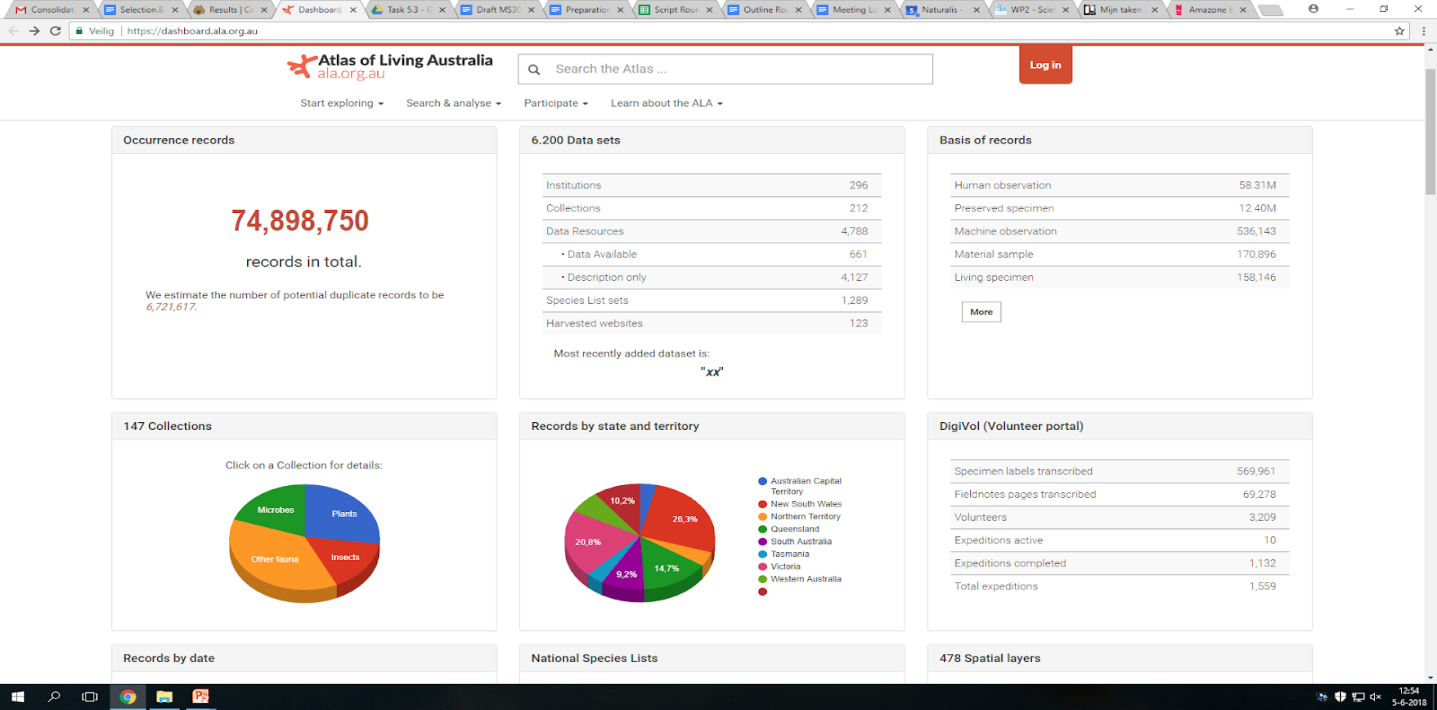
* NCBI Genbank

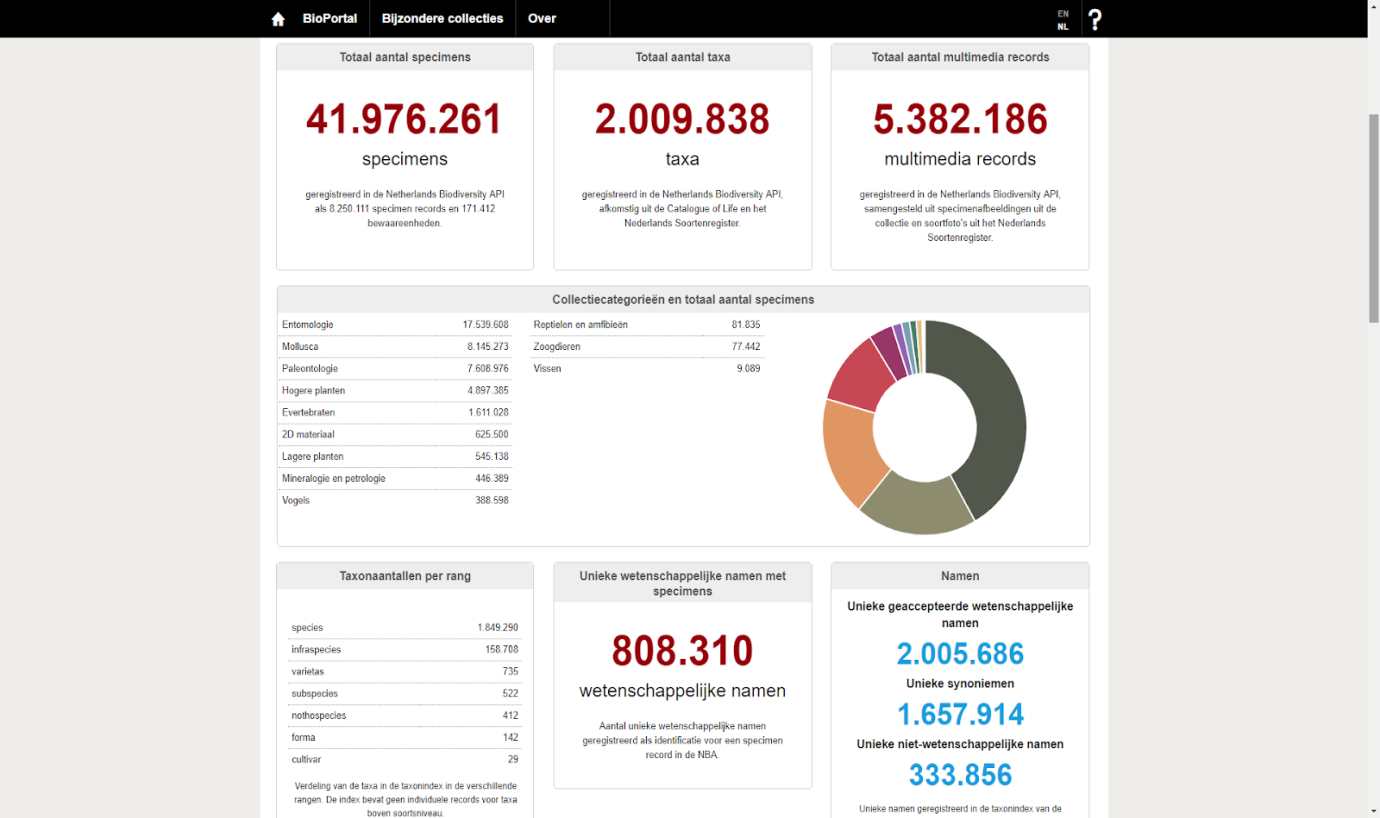
[https://www.ncbi.nlm.nih.gov/genbank/statistics](https://www.ncbi.nlm.nih.gov/genbank/statistics/)

* Smithsonian Institution

<https://www.si.edu/dashboard>

<https://www.si.edu/dashboard/national-collections#collections-digitization>

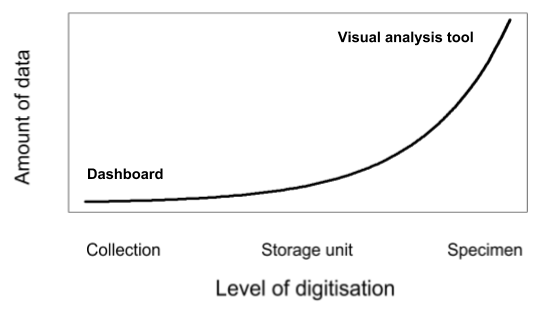
**Figure 1**. Dashboard screenshot from Atlas of Living Australia. <https://dashboard.ala.org.au/>.



**Figure 2.** Dashboard screenshot from Naturalis Biodiversity Center’s BioPortal. <http://bioportal.naturalis.nl/dashboard/>.

It is important to indicate that a dashboard is basically at one end of the continuum of data visualization (more basic), while a visual analysis tool at the other end provides much more details (more advanced). A visual analysis tool is commonly regarded as an online tool that has advanced abilities to select for various date ranges, pick different groups, or drill down to more detailed data (Chiang 2011). Hence, the screen presenting the data is highly interactive and allows searching for patterns and potential outliers in the data. This type of tool is more often associated with researchers as they need more detailed and specific information as well as evolving trends (e.g. [FishfindR.net](http://fishfindr.net/)).

The relationship between the type of visualisation, the level of digitisation and the amount of data needed is presented in a conceptual model (Figure 3). For the CDD, as defined for this task in ICEDIG, the focus will primarily be on collection-level information. There are various reasons to start with a CDD at collection level. First, it is the highest aggregation level of data/information for a collection with the lowest amount of data (Figure 3, lower left). From collection level down to species and specimen levels, the amount of data increases exponentially and along with it the complexity of visualisation tools. One of the issues to be resolved in order to develop a data visualisation tool is how the required information that is distributed across institutes in Europe can be brought together. Another issue involves the standardisation of parameters required for the CDD. Without standardisation, developing a CDD will be difficult and introduce noise into the data. As generating a first design for a European CDD already encompasses quite a number of challenges, the best option is to start with the most attainable endeavour based on data at collection level.

****

**Figure 3.** A conceptual model indicating the relationship between the level of digitisation and the amount of data needed in the visual representation of natural history collection information.

# Requirements for a Collection Digitisation Dashboard

## Workshops with user groups

### Round Table CDD - ICEDIG

On the 11th of June 2018, a Round Table was organised as part of this task on the topic ‘Design of a Collection Digitisation Dashboard for European natural history collections’ during the first ICEDIG All-hands meeting held in Leiden, the Netherlands. The main aim was to prepare a preliminary design for the CDD with the purpose to make digitised and not (yet) digitised natural history collections visible and discoverable across Europe. For this purpose, twenty people attended, consisting of a mix of ICEDIG participants and invited external experts. After the general introduction, two subgroups were formed: the first focused on the end users, user stories and parameters, while the second focused on the technical aspects and unifying data. Discussions in the two subgroups were followed by a short presentation of each subgroup during the final, general discussion. Details on the Round Table can be found in Appendix 1, while the full report will be published as an ICEDIG deliverable (D9.3) together with the other Round Tables at the end of the project.

### Dutch collection overview dashboard – NWO-ALW

As a case study for the European CDD, we have been collaborating with the NWO-ALW (Dutch Science Foundation - Life Sciences) project currently being executed at Naturalis Biodiversity Center. As part of this project, a dashboard presenting a collection overview of the natural history institutions in the Netherlands is being designed (4.2.2). Up until now, three meetings together with Dutch national history institutions have been organised: on the 12th of October 2018, 29th of November 2019, and 25th of January 2019. In these meetings, input was gathered from these institutions to identify their needs regarding a collection overview at national level. At least one more meeting is expected to follow, but will take place just past the due date of this deliverable. Details on the Dutch collection overview meetings can be found in Appendix 2. A full report on the Dutch collection overview will be available mid 2019, which can be provided upon request from Naturalis as that report is not a deliverable within ICEDIG.

## User communities

The following main (potential) user categories for the CDD were identified during the Round Table (in random order): Research, Collection, IT, Governmental, Non-governmental, Education, Industry, Media, Institution and Citizen Science. For each user group, the participants together indicated which level of information (collection, storage unit, species or specimen) would be relevant for each user group (Table 1). This shows that collection and specimen level information are considered to be useful to many of the user groups, while storage unit and species level information is of most value to collection managers.

**Table 1.** Overview of the user groups with indication of presence at the Round Table (RT), and their expected need for each type of data of natural history collections.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **User category** | **Present at RT** | **Collection level** | **Storage Unit level** | **Species level** | **Specimen level** |
| Research | x |  |  | x | x |
| Collection | x |  | x | x |  |
| IT | x | x |  |  | x |
| Governmental |  | x |  |  | x |
| Non-governmental | x | x |  |  |  |
| Education | x |  |  |  | x |
| Industry | x | x |  |  |  |
| Media |  |  |  |  | x |
| Institution | x | x | x | x | x |
| Citizen science | x | x |  | x | x |

During the first meeting on the Dutch collection overview dashboard, the following main user categories were identified:

1. Institutions harbouring natural history collections.
2. (Non-)professional researchers and collectors.
3. Education.
4. Policy makers and financing bodies.
5. NGO nature groups and organisations.
6. The wider community interested in natural heritage.

The first user category - institutions harbouring natural history collections - were naturally considered to be the most important user group, as these were the participants of these meetings. And as providers of data for this collection overview, there needs to be a clear use to encourage their participation, e.g. institutional collection and digitisation progress reports, badging, etc.

## User stories

After identifying the main user categories for a collection digitisation dashboard, user stories were captured at a higher hierarchical level following the format of an epic user story. An epic user story format starts with: ‘as a’ [*user*] ‘I want to’ [*do this; know this*] ‘so that I’ [*can do this*]. For example: “as a *collection manager*, I want to *see all digitised European collections of bees*, so I *can prioritise the digitisation of bee collections*”. In total, 22 user stories were collected related to the CDD during the Round Table (Appendix 1). We have selected the user stories that were identified to explicitly need both digitised and not digitised collection-level information (Table 2). In addition, it was indicated by the participants during the Round Table that an overview of natural history collections at the highest data level would be (to varying degrees) useful to all user groups.

During the first meeting on the Dutch collection overview dashboard, user stories were collected for the different identified user groups (Appendix 3). At an institute, directors would like to be able to evaluate collection donations and therefore would be interested in the taxonomic groups that have already been included in Dutch natural history collections. For a collection manager, different levels of collection information could be useful, but collection-level information is needed to increase the discoverability of (parts of) the collection at national level for several purposes. Also, collection managers are interested in the niche his/her institute holds in the national landscape and use this to see where improvements/enrichments in e.g. geographic scope of the collection can be made. When collection policies are written within an institute, the position of the collection in comparison with other Dutch natural history collections needs to be clear. Researchers and citizen scientists, as well as educational institutions are interested to know which collections from the various main geographic regions are held by an institute. Policy makers and funding bodies need information to be able to determine how to divide funds across institutions/projects but also to create policy for e.g. local nature conservation. Nature-focused non-governmental organisations highlighted the need for specimen level information e.g. to be able to produce distribution maps, although collection-level information may still give relevant, more general information for this group to see what is present in the Dutch natural history collections as a whole. And the wider community and journalists might want to know what to expect when they visit an institute, or which taxonomic expertise is represented by the staff of institutes.

**Table 2.** Selected user stories that were identified during the Round Table that need both digitised and not digitised collection-level information.

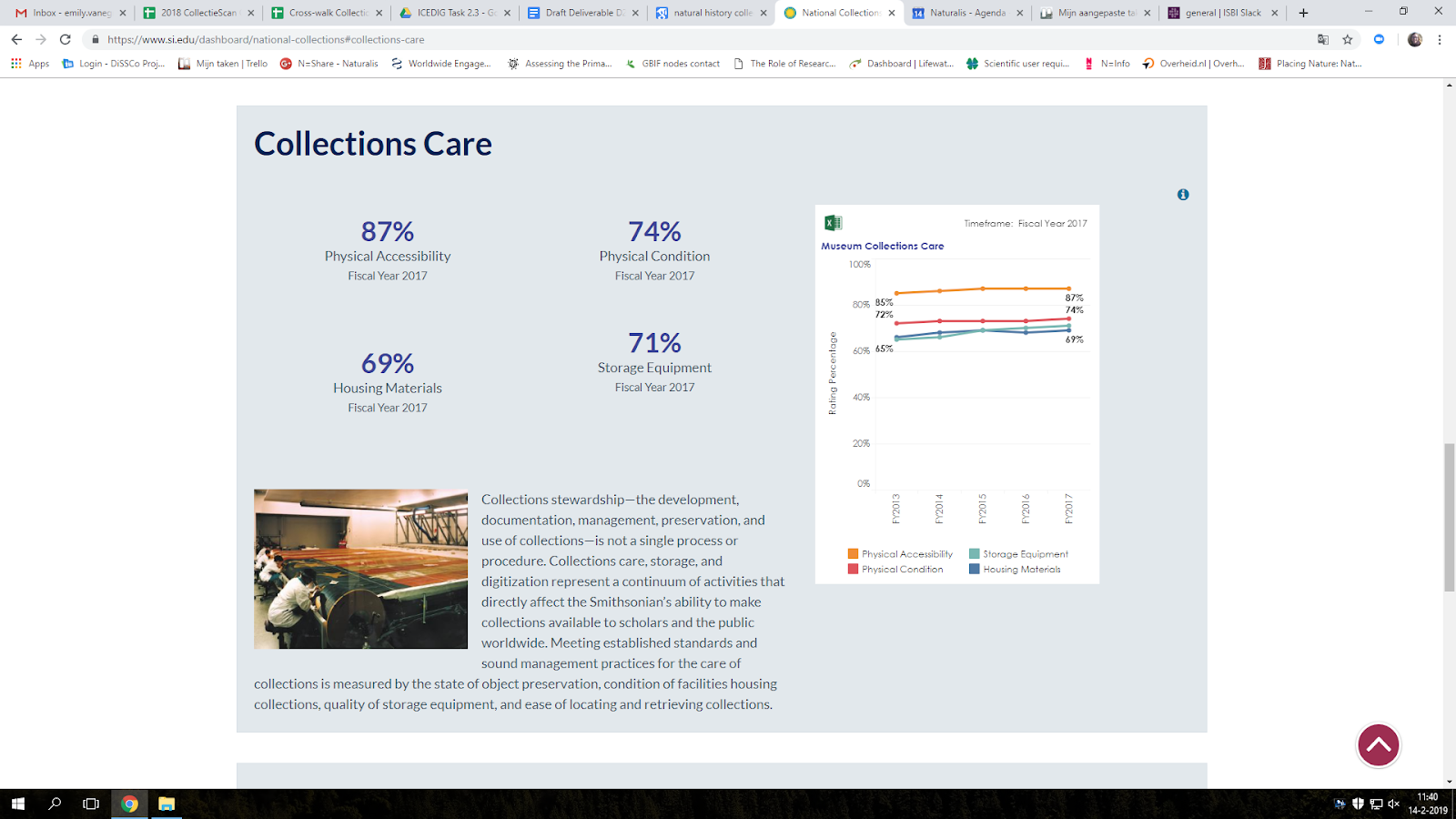
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **User groups** | **As a** | **I want to** | **So that** | **For this I need (data elements)** | **Level of digitization** | **Digitized/non-digitized** |
| Institution | Director | Hire a curator with knowledge of specific groups | I can be sure they have a background that includes knowledge of the main collection | Collection types, importance of collection gauged by size, scope, and time period of collection | Collection | Both |
| Institution | Collection manager, Director, Administrator | Know what the situation is regarding collection size | I can plan for new space / storage needs | I need to know existing size of collections, and amount of new material coming in. Also, I need to know the status / condition (e.g wet, dry) of existing material and collection health information | Collection, species | Both |
| Collection | Collection Manager | Start a digitizing project | I like to digitize a certain group of my collection, I like to do this internationally because of funding | Know where else there are collections of this group | All levels | Both, but mainly digitised |
| Governmental | Policy maker | Know the use of the collections by other domains as a key indicator of its impact globally | I can distribute resources and allocate them in alignment to the strategic priorities of the government that I represent | Access to the collections, virtually and physically, from different types of users | Collection | Both, digitized (publicly available) and non-digitized (to understand the need to bridge the gap) |
| Non-governmental | Association | To gather information to have overall figures representative of partners' state-of-the-art | We can showcase the relevance of the collections to policy makers and attract funds | High-level figures that feature the collections as a whole | Collections | Both, digitized and non-digitized information are valuable (to indicate the progress and the support needed, respectively) |
| IT | Solution provider | Tap into the vast market of digital storage solutions for digital natural collections | I can sell my services and consult | Predictable numbers on collection type, volume and progress in digitization | Collection | Both |
| Industry | Solution provider | Build and provide solutions and related services | The keepers and scientists can work better and easier with their collections for less cost | Volumes, locations and physical sizes plus an insight on what is digitally represented and what not. Even better would be if there is an institutions priority as to what needs to be digital first. | Collection and partly storage level | Both |

## Parameters of interest for the CDD

Each of the main user stories of interest (see Table 2), indicated, in relatively similar terms, that there is a need for information on collection type (taxonomy), volume and its geographical coverage. This information is needed in addition to the required digitised and non-digitised data that is essential to be able to use the CDD as a prioritisation tool for collection digitisation. During the Round Table it became clear from the general discussion that it is necessary to clearly indicate when we consider a specimen to be digitised or not digitised. There appeared to be differences between participants considering these last terms, where some would consider a specimen that is only catalogued as being digitised, while others consider a specimen digitised when e.g. a picture has been added to the record. Overall, this discussion emphasised the importance of having clear definitions for the variables to be presented in the CDD (see 3.5).

Although one institutional user story from collection managers indicated that it would be useful to have an indication of the physical condition of the collection in the CDD (Table 2), the availability of this kind of data is expected to be low, especially when this information needs to be broken down into e.g. taxonomic groups. A few years ago, however, the Smithsonian National Museum of Natural History (US) developed the ‘Move/Join the Dots’ assessment, a tool that can be used by collection managers to indicate among others the physical condition of the collection ([Smith *et. al.* 2018](https://www.idigbio.org/content/shining-new-light-world%E2%80%99s-collections)). With this tool a collection is divided into logical units and for each unit a number of characteristics are inventoried like the physical condition of the specimens, the appropriateness and quality of the storage unit and the physical accessibility. The ‘Move/Join the Dots’ assessment is being explored by the larger national history institutions in the world (One World Collection institutes), and has for example been adopted by NHM London. A link between a ‘Move/Join the Dots’ assessment from an institution and the CDD for aspects regarding the condition of a collection seems therefore interesting to explore in the future. Data on collections care can then be visualised in a dashboard (such as in Figure 4), but a breakdown by taxonomy (3.3), geography (3.4) and digitisation (3.5) levels will be a requirement for a CDD to be used as a prioritisation tool for collection digitisation.

Stakeholder feedback on the Dutch collection overview dashboard indicated that taxonomic information, geographic information, volume and the current state of digitisation were among the most important parameters to include in a dashboard. Additionally, preservation type and state of specimens were indicated as relevant parameters, but more difficult to implement. This corresponds to the outcome of the Round Table. In addition, information about collection expertise of institutional staff (collection managers/researchers) and specialisations of institutions at a national level was highly valued and mentioned multiple times in the user stories. Most likely, information on staff expertise needs to be presented as a list, due to expected detailed levels of taxonomic and/or geographic expertise (e.g. the *Curculionidae* of Germany). To our knowledge, there is no example of a dashboard that represents the fields of expertise of individual collection employees or expertise at institutional level (or even at national level). A first overview of all collection institutes with a basic characterisation of their collection has been developed for the United States in the [iDigBio project](https://www.idigbio.org/portal/collections), however. At the European level more exploration is required on how to implement the expertise/knowledge parameter in a dashboard in the future.



**Figure 4.** Dashboard snapshot presenting collections care information from the [Smithsonian National Museum of Natural History](https://www.si.edu/dashboard/national-collections).

## Visual requirements

As there is much data that potentially can be shown in the CDD, it will be necessary to prepare not a single dashboard, but dashboards consisting of multiple pages. The first dashboard page could present a couple of individual numbers that give an impression of the total collection, such as the total number of specimens within all institutions combined. Each following page can then present data related to a certain theme (e.g. taxonomic group).

Ideally, a CDD presents figures that can be interpreted quickly, easily and unambiguously by a wide variety of people. For clarity and to prevent misinterpretation of any of the dashboard figures, it might be useful to add a simple explanation of the variables used. This can be done in different ways, such as placing the explanation directly in sight (Figure 5a) or by displaying the explanation when hovering over the dashboard (Figure 5b). Also, a link to a glossary can be added, which is reached when clicked on the corresponding figure/term on the dashboard.

|  |  |
| --- | --- |
| a) | https://lh4.googleusercontent.com/MkjmCGW2QSLSzOVkR1qT681jY8osTRB0fJGix0_p2ciBKhfTctTijCWBAzKw9671RXoxYP-rflFFcBW7tC-0cmA34H0vhtbpy0zs1V3YvoizTz0nbsHIfyVk51xc_MQV0sTTDh3E |
| b) | https://lh4.googleusercontent.com/1-UXd7Np8kj5d2Idx7Mi_Qt_eW5wtpnAU-d6hD3VzuL5ICwwpbaRbZHvCPELGx6nFjoXhkEzM2bqxZGtdJiWFbLYrmKI0WmcOpU-_vIVVMr5jGhDjZ_cR81lc6-daK029yLKbtrr |

**Figure 5.** Dashboard visualisation snapshots presenting information explaining the visualised variables: a) showing explanation directly in sight (small paragraph at the bottom; Naturalis Biodiversity Center BioPortal), b) showing explanation when hovering over a number (still in dark blue; Smithsonian National Museum of Natural History).

As a CDD is designed to inform users that require high-level collection information, its presentation needs to be visually appealing. This will need to result in dashboard pages that contain not too much nor too little information and/or open space. Different types of data visualisations will keep visitors curious and interested, and if the visualisation fits the type of information it is presenting, it will make interpretation easier. These could for example include maps, (stacked) bar graphs, (stacked) line graphs, pie charts and individual numbers grouped in a metadata table.

A map of Europe indicating the size and scope of the collection for each country would be interesting in that respect. During the first meeting of the Dutch collection overview, a map of the Netherlands with the location of all natural history institutes was indicated to be of high interest (similar as the [iDigBio dashboard](https://www.idigbio.org/portal/collections)). These could perhaps be filtered based on taxonomic group and/or geographic region or the level of collection digitisation. When you click on an institute, information of that institute is shown on a separate institutional page of the dashboard. When creating this visualisation on a European level it is important to keep in mind that countries with either large collections or with many participating institutions have a visual dominance.

Stacked bar graphs are suitable to visualise for example the degree of digitisation, where one bar represents the entire collection as the number of specimens and the stacked parts of the collection that have been digitised to a certain degree. Stacked line graphs can show progress of digitisation (digitised vs. non-digitised) over time, split out to taxonomic and geographic groups. Also, the progress of digitisation can be compared between countries and institutions.

Pie charts can be used to easily filter and present data for a certain subgroup. For example, several pie charts can be shown for the subgroup ‘vascular plants’ or ‘minerals’ to indicate e.g. the countries, institutions and geographic regions that contain (not)digitised specimens for this subgroup. As indicated by both workshops with user groups, it will be necessary to be clear about what answers from the CDD are needed for which questions from the end users. One such question could be: ‘Which taxonomic group contains the lowest percentage of digitised specimens?’.

A table can finally be used to quickly but precisely compare individual numbers across institutions, countries, taxonomic or geographic groups and digitisation degree. Data can be filtered by clicking on rows within this table.

## Technical requirements

The CDD needs to be publicly available online and be rapidly adjusted and/or updated when the underlying data changes. Various technical solutions to create a dashboard exist and an overview was prepared for comparison (Table 3). Of these technical solutions, Microsoft Power BI was selected to create the CDD due to its ease of use, flexibility, professional tools, and it being free of charge / low in costs.

To be able to collect and integrate the data easily, data is ideally collected in one format or agreed standard (e.g. [TDWG collection description standard](https://www.tdwg.org/standards/ncd/)) to simplify integration. Initially this may be done manually, but in the future it is desirable that data collection and integration occurs in an automated way. One possible automated process that was discussed during the ICEDIG Round Table within the technical subgroup is to harvest data from institution sites with an RSS. An RSS file could be placed at the institution’s website showing collection level digitisation. This data can then be harvested by the dashboard. In the end, it would be useful to synchronize APIs, which get data on digitised records from the Collection Management System (CMS) of the institute and/or when contributed to GBIF to feed into the dashboard. Updating of the entire institutional collection holding estimates when new collections arrive will always require manual input, however.

As a more practical requirement that was indicated during the workshops, contact information of natural history institutions should be easy to find within the dashboard, both for the main institution and for experts on different taxonomic and geographic fields (when included). This will make it easier to find and reach people associated with a particular collection or with specific taxonomic expertise.

**Table 3.** Overview of available software applications to create Business Intelligence Dashboards.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Tableau** | **Microsoft Power BI** | **IBM Cognos Analytics on Cloud** | **Shiny dashboards** | **Google Data Studio** | **Kibana** | **Grafana** | **Splunk** |
| **url** | <https://www.tableau.com/> | <https://powerbi.microsoft.com/en-us/> | <https://www.ibm.com/nl-en/marketplace/business-intelligence#product-header-top> | <https://rstudio.github.io/shinydashboard/index.html> | <https://datastudio.google.com> | <https://www.elastic.co/guide/en/kibana/current/introduction.html> | <https://grafana.com/> | <https://www.splunk.com> |
| **Examples** | <https://www.tableau.com/solutions> | <https://powerbi.microsoft.com/en-us/tour/>  <https://community.powerbi.com/t5/Data-Stories-Gallery/bd-p/DataStoriesGallery> | <https://www.ibm.com/nl-en/marketplace/business-intelligence/details> | <https://gallery.shinyapps.io/LDAelife/>  <http://www.dataseries.org/> | <https://datastudio.google.com/u/0/navigation/reporting> | Kibana is an open source analytics and visualisation platform designed to work with Elasticsearch. | <https://grafana.com/dashboards>  Grafana is an open source analytics and visualisation platform designed to work with Elasticsearch. | <https://www.splunk.com/en_us/software.html>  Is used as a dashboard tool by CERN (Andrade et al. 2012) |
| **Price** | <https://www.tableau.com/pricing/teams-orgs>  $70,00 per user per month | Power BI Desktop - Free  Power BI Pro - $9.99 per user per month | Starting at € 1.920,25 per month | Free | Free | Free | Free | Free version available, more advanced versions from $87 per ingested GB per month |

# Collection description standards

## Task Group Collection Digitisation Dashboards – TG CDD

With the establishment of DiSSCo, all partners that signed the MoU were asked to fill in a survey with information about their institutional collection. The initial survey included an estimate of the entire institutional holding of specimens and a breakdown in ten collection categories as percentage classes of 10%, i.e. 0-10%, 10-20%, etc. The ten categories included: 1) Botany, 2) Zoology, 3) Entomology, 4) Mycology, 5) Microbiology, 6) Paleontology, 7) DNA, 8) Living, 9) Seed, and 10) Mineralogy. The results of the survey indicated that the used categories were ambiguous. For example, the Westerdijk Institute in the Netherlands has mycological collections making up 90-100% of its total collection; at the same time most specimens of the Westerdijk Institute are living (90-100%), and from a large percentage of specimens DNA samples were taken, together summing to far over 100%. This indicated the urgent need for an unambiguous collection description standard. This need was also recognised by TDWG (Biodiversity Information Standards) who initiated the Draft Standard for Natural Collection Descriptions which is now being updated by the [TDWG Collection Descriptions Task Group](https://github.com/tdwg/cd).

These outcomes, together with the start of ICEDIG task 2.3, initiated the start of the Task Group Collection Digitisation Dashboards (TG CDD) on the 8th of June 2018, lead by Niels Raes from Naturalis. The TG CDD aims to harmonise data requirements for visualisation of (not yet) digitised natural history collections and the analysis of digitisation progress. These discussions contribute to set collection description standards and provide recommendations to the final TDWG standard for Collection Descriptions. The TG CDD currently consists of 16 members from a wide variety of institutions and international organisations (Appendix 1).

## Inventory of collection description schemes

There are multiple examples available that describe natural history collections at high level. Via an internet search and our network we identified the following collection description schemes which are currently in use or under development:

1. One World Collection (brief summary in <https://www.idigbio.org/content/shining-new-light-world%E2%80%99s-collections>)
2. GRBIO (FAQ: <http://scicoll.org/grbio_error.html>)  
   <https://docs.google.com/spreadsheets/d/1JD3ROc4X6paBlKtmbunF6gG3htSGBU_RVZvfkST0qlM/edit#gid=1813299279>
3. TDWG NCD and CD  
   <https://terms.tdwg.org/wiki/Natural_Collections_Description#CollectionType>  
   <https://github.com/tdwg/ncd/blob/master/NCD-v090_TDWG/NCD-v090_TDWG-NonNormative.pdf>  
   <https://github.com/tdwg/ncd/blob/master/NCD-v090_TDWG/NCD-v090_TDWG-Normative.pdf>  
   <https://github.com/tdwg/cd>  
   [https://github.com/tdwg/cd/blob/master/charters/task­\_group­\_charter/tg­\_charter.md](https://github.com/tdwg/cd/blob/master/charters/task)
4. CETAF digitisation workgroup  
   <https://species-id.net/o/media/c/c8/Digitisation_definitions_for_collections.pdf>
5. CETAF passports  
   <https://cetaf.org/tags/passports>
6. Join the Dots (NHM) (not published)  
   <https://biss.pensoft.net/article/26500/download/pdf/>
7. CSAT Synthesys (Collections Self-Assessment Tool)  
   <http://synthesys3.myspecies.info>
8. CABRI (microorganisms) - Common Access to Biological Resources and Information <http://www.cabri.org/guidelines.html>
9. MIRRI (Microbial Resource Research Infrastructure - microorganisms) <https://www.mirri.org/about-mirri/the-rationale-for-mirri.html>
10. Catalogue of Life   
    higher taxonomy <https://doi.org/10.1371/journal.pone.0119248> and <http://www.catalogueoflife.org/col/browse/tree>
11. iDigBio  
    <https://docs.google.com/forms/d/e/1FAIpQLSfffIIqu9PoAac2FVBCCBNt66O1WQbxvtWn60_1fVtAx23nAQ/viewform>
12. RBINS – Geology  
    <https://www.naturalsciences.be/en/science/collections/overview/542>
13. NHM collection   
    <http://www.nhm.ac.uk/our-science/collections.html>
14. American Museum of Natural History  
    <https://www.amnh.org/our-research>
15. Museum National d’Histoire Naturelle <https://www.mnhn.fr/en/collections/collection-groups/>

The first six collection descriptions schemes were compared in a crosswalk analysis, identifying the differences and commonalities between the different description schemes. The remaining collection data descriptions were used to check for additional components that were potentially lacking. During the TG CDD meetings, the crosswalk analysis was extensively discussed and reiteratively adjusted where needed. The results of the crosswalk analysis informed us to prepare and propose an improved collection description scheme that is based on the currently existing descriptions for collection level information. We aim to contribute this collection description scheme as a community agreed standard via [TDWG collection description task group](https://github.com/tdwg/cd). Also, this gave us detailed information on which parameters and their levels of detail are needed to be minimally included in a CDD.

*It is important to note* that collection description schemes describe physical objects. A DNA sample taken from a herbarium specimen, or a pollen sample taken from drilling cores, which are stored separately from the original specimen, constitute two physical specimens. Linking between specimens is arranged at the level of the CMS, or DiSSCo, and is not part of the current collection description scheme.

## Three parallel collection classifications

Following the discussions in the TG CDD, meetings with the Dutch national history institutions, and the Round Table meeting we came to the conclusion that essentially three main collection classifications exist in parallel, each with collection information relevant to different user groups. The three recognised user groups are: a) ‘scientific biological’ requiring a taxonomic classification, b) ‘collection management’ in need of a storage classification, and c) ‘scientific geological’ requiring a stratigraphic classification. The scientific biological view focuses on the taxonomic division of specimens (e.g. which institutes hold botanical/zoological specimens), while the scientific geological (and paleontological) view focuses on the geological period from where a specimen was collected. The collection management view focuses on the preservation type (e.g. dried, liquid preserved, etc.) and, linked to that, the storage type that is needed for a (set of) specimen(s). For the CDD purpose, we focus on the scientific biological view for now, mainly because quantitative information about the number of specimens for each preservation category is currently lacking, and there are generally more biological than geological specimens. For each classification all subcategories are nested within main categories, ensuring that data can be aggregated to higher hierarchical levels when needed. A full description of the three identified classifications is provided below.

### ‘Taxonomic’ classification

The ‘taxonomic’ classification includes elements from both biodiversity (taxonomy) and geodiversity (paleontology, geology and extraterrestrial), hence ‘Taxonomic’ between quotation marks. In addition, the main category ‘Not bio/geodiversity’ was introduced to assign a label to bio- and geodiversity related objects such as field note books and art works. Further classification for these objects needs to be defined by the appropriate domains and is beyond the scope of this classification. The main group ‘Bio/geodiversity other’ can be used to assign a label to specimens that were impossible to identify or contain a mixture of several specimens from different taxonomic groups (e.g. ecological soil sample, pitfall trap, water sample).

Interestingly, the crosswalk analysis revealed that different collection data description schemes always recognised several subcategories under zoology, while this was not the case for botany. In other words, all algae, vascular plants and bryophytes are placed within the main group ‘botany’ even though these are highly distinct, while for the main group ‘zoology’ multiple divisions were recognised (e.g. insects, birds and mammals are all separated). We therefore introduced subdivisions for ‘botany’ that distinguish ‘vascular plants’ from ‘algae’ and ‘bryophytes’. Also for the main category ‘microorganisms’ multiple subgroups were introduced.

A preliminary proposal for the taxonomic classification has been composed (Table 4), based on the results of the crosswalk analysis of collection data description schemes (section 3.2). Ten main categories were identified with the number of subcategories indicated between brackets: Botany (3), Mycology (-), Zoology Invertebrates (7), Zoology Vertebrates (5), Microorganisms (7), Bio/geodiversity other (-), Paleontology (4), Geology (3), Extraterrestrial (2) and Not bio/geodiversity (-). In total 34 subcategories are recognised.

**Table 4.** Overview of the preliminary proposal for the ‘Taxonomic’ classification, indicating the main category and subcategory of the ‘Taxonomic’ collection description standard.

|  |  |
| --- | --- |
| **Main category** | **Subcategory** |
| Botany | Botany: Vascular plants |
| Botany: Bryophytes (mosses) |
| Botany: Algae |
| Mycology | Fungi, including lichens |
| Zoology Invertebrates | Zoology Invertebrates: Arthropods - insects |
| Zoology Invertebrates: Arthropods - arachnids |
| Zoology Invertebrates: Arthropods - crustaceans & myriapods |
| Zoology Invertebrates: Mollusks (bivalves, gastropods, cephalopods) |
| Zoology Invertebrates: Cnidaria (corals, jellyfish, anemones) |
| Zoology Invertebrates: Porifera (sponges) |
| Zoology Invertebrates: Other (other taxonomic groups) |
| Zoology Vertebrates | Zoology Vertebrates: Fishes |
| Zoology Vertebrates: Amphibians |
| Zoology Vertebrates: Reptiles |
| Zoology Vertebrates: Birds |
| Zoology Vertebrates: Mammals |
| Microorganisms | Microorganisms: Bacteria and Archaea |
| Microorganisms: Phages |
| Microorganisms: Plasmids |
| Microorganisms: Protozoa |
| Microorganisms: Virus - animal / human |
| Microorganisms: Virus - plant |
| Microorganisms: Yeast |
| Bio/geodiversity other | E.g. eDNA, culture/tissue collection, mixed biological collections (virus infected living plant), drilling cores including pollen and plant remains |
| Palaeontology | Palaeontology: Botany & Mycology |
| Palaeontology: Zoology Invertebrates |
| Palaeontology: Zoology Vertebrates |
| Palaeontology: Trace fossils (e.g. footprints) |
| Geology | Geology: Mineralogy (e.g. rocks, ores, gems, minerals) |
| Geology: Sample (e.g. drilling cores, soil, (ocean) sediment) |
| Geology: Other (e.g. fluid) |
| Extra-terrestrial | Extra-terrestrial: Collected on Earth (e.g. meteorites) |
| Extra-terrestrial: Collected in space (e.g. moonstone) |
| Not bio/geodiversity | Not bio/geodiversity - classified by other domain |

### ‘Storage’ classification

The ‘Storage’ classification is focussed on the storage type of a specimen and closely relates to collection management. Identifying the way a specimen has been preserved, such as a dried insect on a pin, also determines how it needs to be stored (in a drawer, in a dry environment). The domain ‘Biology’ was first divided into ‘Preserved (dead)’ and ‘Living’ specimens, allowing not only natural history specimens to be included but also specimens from living culture collections, botanical gardens and zoos. Under ‘Preserved’ the main categories are the same as for the ‘Taxonomic’ classification, but ‘Paleontology’ has been added as a main category to the domain ‘Biology: preserved (dead)’.

A preliminary proposal for the ‘Storage’ classification has been composed based on the first results of the crosswalk analysis for collection data descriptions (Table 5). Fourteen main categories were identified, with the number of subcategories indicated between brackets: Preserved > Botany (5), Preserved > Mycology (5), Preserved > Zoology Invertebrates (6), Preserved > Zoology Vertebrates (5), Preserved > Microbiology (3), Preserved > Paleontology (6), Preserved > Other (-), Living > Botany (3), Living > Mycology (-), Living > Zoology (2), Living > Microbiology (2), Geology (6), Extraterrestrial (2) and Not bio/geodiversity (-). This results in a total number of 48 subcategories. Finally, examples are given for each subcategory to indicate what type of specimen could fit here (see last column of Table 5).

**Table 5.** Overview of the preliminary proposal for the ‘Storage’ classification, indicating the main category and subcategory of the collection description classification.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Domain** | **Origin** | **Main category** | **Subcategory** | ***Examples*** |
| Biology | Biology: Preserved (dead) | Botany | Botany: pressed and dried | *Herbarium specimens* |
| Botany: dried | *Fruits, wood samples* |
| Botany: fluid preserved | *Flowers in alcohol/formalin/glycerine* |
| Botany: microscopic slides | *Microscopic slides* |
| Botany: cryopreserved / frozen -80°C | *DNA / RNA* |
| Mycology | Mycology: dried | *Dried fungi* |
| Mycology: spore print | *Spore print* |
| Mycology: fluid preserved | *Fungi in alcohol/formalin/glycerine* |
| Mycology: microscopic slides | *Microscopic slides* |
| Mycology: cryopreserved / frozen -80°C | *DNA / RNA* |
| Zoology Invertebrates | Zoology Invertebrates: dried - pinned | *Pinned insects* |
| Zoology Invertebrates: dried - assembled | *Not pinned. Multiple animal parts or entire organism* |
| Zoology Invertebrates: dried - not assembled | *Animal part: shell, bone, etc.* |
| Zoology Invertebrates: fluid preserved | *Animals in alcohol/formalin/glycerine* |
| Zoology Invertebrates: microscopic slides | *Microscopic slides* |
| Zoology Invertebrates: cryopreserved / frozen -80°C | *DNA / RNA* |
| Zoology Vertebrates | Zoology Vertebrates: dried - assembled | *Multiple animal parts or entire organism: skeletons, stuffed animals* |
| Zoology Vertebrates: dried - not assembled | *Animal part: tanned skin, egg shell, etc.* |
| Zoology Vertebrates: fluid preserved | *Animals in alcohol/formalin/glycerine* |
| Zoology Vertebrates: microscopic slides | *Microscopic slides* |
| Zoology Vertebrates: cryopreserved / frozen -80°C | *DNA / RNA* |
| Microbiology | Microbiology: dried |  |
| Microbiology: microscopic slides |  |
| Microbiology: cryopreserved DNA / RNA | *DNA / RNA* |
| Palaeontology | Palaeontological: botany | *Dead and fossilized plants* |
| Palaeontological: mycology | *Dead and fossilized fungi* |
| Palaeontological: zoology vertebrates | *Dead and fossilized vertebrate animals* |
| Palaeontological: zoology invertebrates | *Dead and fossilized invertebrate animals* |
| Palaeontological: trace fossils | *Foot prints etc.* |
| Palaeontological: microscopic slides | *Microscopic slides* |
| Other | Other | *Waxblock, SEM stub, surface coating, embedded* |
| Biology: Living | Botany | Botany (in vivo) | *Botanical garden* |
| Botany (in vitro) | *Algae cultured collections* |
| Botany: Seeds & germplasm (dormant) | *Seeds* |
| Mycology | Mycology (in vitro) | *Spores* |
| Zoology | Zoology (in vivo) | *Zoo* |
| Zoology: germplasm (in vitro, dormant) | *Sperm, egg cells* |
| Microbiology | Microbiology: cryopreserved / frozen -80°C (in vitro) | *Dormant* |
| Microbiology: cell and tissue cultures (in vitro) |  |
| Geology | | | Geology: Mineralogy | *Rocks, gems, minerals* |
| Geology: Sample | *Soil, sediment, cores* |
| Geology: Microscopic slide | *Microscopic slides* |
| Geology: Fluid | *Fluids, e.g. water* |
| Geology: Radioactive | *Radioactive materials* |
| Geology: Other |  |
| Extra-terrestrial | | | Extra-terrestrial: Collected on Earth | *Meteorites* |
| Extra-terrestrial: Collected in space | *Moonstone* |
| Not geo/biodiversity | | | Not geo/biodiversity |  |

### Stratigraphic classification (only geodiversity and paleontology)

During the discussions within the TG CDD, it became clear that for geodiversity and paleontology it is essential to have a ‘Stratigraphic’ classification. A specimen such as a rock or mineral can then be assigned a geological time period from which the specimen originates. Three era’s were identified, subdivided into 12 periods. The Quaternary, Neogene and Paleogene periods are further subdivided into either two (former two) or three (latter one) epochs. This results in a total of 16 subcategories that can be used in a dashboard.

A preliminary proposal for the ‘stratigraphic’ classification has been composed based on the first results of the crosswalk analysis for collection data descriptions (Table 6). For the initial CDD we will however focus on biodiversity and can later add geodiversity/paleontology and the related stratigraphic classification.

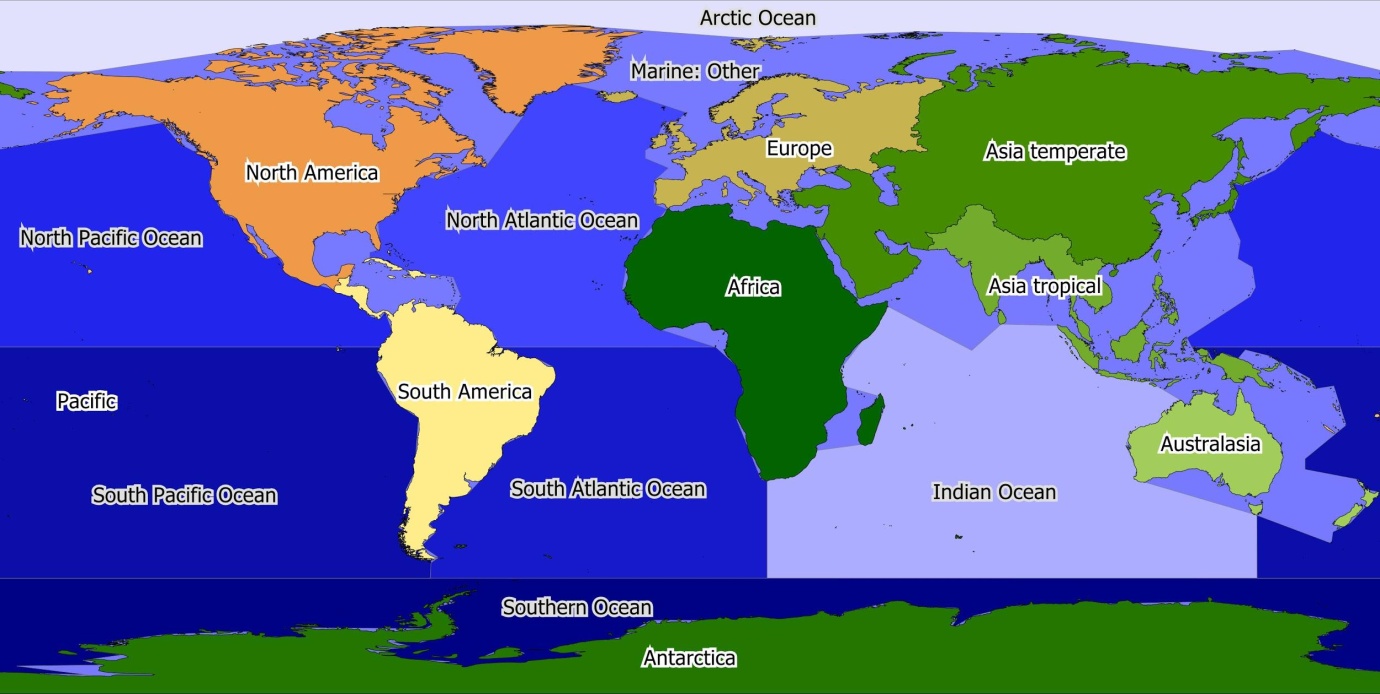
**Table 6.** Overview of the preliminary proposal for the ‘Stratigraphic’ classification, indicating the main and subcategories of the collection description classification.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Eon** | **Era** | **Period** | **Epoch** | **Upper (Ma)** | **Lower (Ma)** |
| Phanerozoic | Cenozoic | Quaternary | Holocene | 0.00 | 0.01 |
| Pleistocene | 0.01 | 2.58 |
| Neogene | Pliocene | 2.58 | 5.33 |
| Miocene | 5.333 | 23.03 |
| Paleogene | Oligocene | 23.03 | 33.9 |
| Eocene | 33.9 | 56 |
| Paleocene | 56 | 66 |
| Mesozoic | Cretaceous | | 66 | 100.5 |
| Jurassic | | 100.5 | 201.3 |
| Triassic | | 201.3 | 251.902 |
| Paleozoic | Permian | | 251.902 | 298.9 |
| Carboniferous | | 298.9 | 358.9 |
| Devonian | | 358.9 | 419.2 |
| Silurian | | 419.12 | 443.8 |
| Ordovician | | 443.8 | 485.4 |
| Cambrian | | 485.4 | 541 |

## Geographic classification

The three parallel collection classifications described above (‘Taxonomic’, ‘Storage’ and Stratigraphic) can each be further broken down according to a geographic classification (Table 7). For example, all vascular plants (Taxonomic classification - Botany: vascular plant) may be subdivided into the vascular plants from Europe (Terrestrial: Europe), North America (Terrestrial: North America) etc. Similarly, all fossils from the Miocene (Stratigraphic classification - Cenozoic: Neogene: Miocene) may be subdivided into the fossils from Europe (Terrestrial: Europe), North America (Terrestrial: North America) etc. Moreover, all pinned insects (Storage classification - Biology: Preserved (dead): Zoology Invertebrates: dried - pinned) may be subdivided into pinned insects collected from Europe (Terrestrial: Europe), North America (Terrestrial: North America) etc. Three main geographic categories were identified (the number of subcategories are indicated within brackets): Terrestrial (10), Marine (9) and Extraterrestrial (-). In total, there are 19 geographic subcategories. Both the terrestrial and the marine main categories contain a subcategory ‘world/NA’ for any specimen that cannot be assigned to one of the categories. The definition of the different terrestrial regions is based on the TDWG World Geographical Scheme for Recording Plant Distributions ([WGSRPD](https://www.tdwg.org/standards/wgsrpd/) - level 1) (Figure 6; Brummit, 2001). The definition of the marine regions is based on ‘IHO World Seas - version 3’ from the International Hydrographic Organisation (IHO) (Figure 6; Flanders Marine Institute, 2018).

Specimens collected from fresh or brackish water bodies will need to be placed under one of the subcategories of the main category terrestrial. For example, a small invertebrate collected from the river Rhine will receive the label ‘Terrestrial: Europe’ from the geographic classification. From the taxonomic classification its aquatic environmental requirements can be derived.



**Figure 6.** Map showing the TDWG terrestrial ([WGSRPD](https://www.tdwg.org/standards/wgsrpd/) - level 1) and marine (IHO World Seas - version 3) regions used in the geographic classification.

**Table 7.** Overview of the preliminary proposal for the ‘Geographic’ classification, indicating the main group and subgroup of the collection description classification.

|  |  |
| --- | --- |
| **Main category** | **Subcategory** |
| Terrestrial | Terrestrial: Africa |
| Terrestrial: Antarctic |
| Terrestrial: Asia Temperate |
| Terrestrial: Asia Tropical |
| Terrestrial: Australasia |
| Terrestrial: Europe |
| Terrestrial: North America |
| Terrestrial: Pacific |
| Terrestrial: South America |
| Terrestrial: World / NA |
| Marine | Marine: Arctic Ocean |
| Marine: Indian Ocean |
| Marine: North Atlantic |
| Marine: South Atlantic |
| Marine: North Pacific |
| Marine: South Pacific |
| Marine: Southern Ocean |
| Marine: Other |
| Marine: World / NA |
| Extra-terrestrial | Extra-terrestrial |

## Digitisation classification

Within the wider collection digitisation community, numerous discussions have proven that it is difficult to settle on a digitisation classification, even at a high level. Previously, a tiered strategy for collection digitisation was proposed and consists of five levels (Krishtalka et al. 2016):

1. Metadata I: collection-level information.
2. Metadata II: species-level or cabinet-level information of the collection.
3. Specimen data I: skeletal-level data which has been checked for data quality.
4. Specimen data II: georeference locality data and adding additional data such as field notes.
5. Specimen data III: data from new collections is immediately digitised and added to the digital, including georeferencing and data quality checks.

Secondly, within the DiSSCo questionnaire on the size and identity of institutes which are members of the DiSSCo consortium (2017-2018), respondents were asked to give an indication for each of the following three levels of digitisation:

* Percentage of collections catalogued; i.e. records exist in in-house collection management system.
* Percentage of collections digitised; i.e. specimen information in collection management system with partly or fully transcribed labels.
* Percentage of collections fully digitised; i.e. specimen information in collection management system, with fully transcribed labels and images.

Although a tiered strategy for digitisation can definitely be useful, there will be few natural history institutions that have completed the first three steps as described by Krishtalka et al. (2016), let alone all five steps. Also, not each tier is unambiguous and can be completely nested (both examples).

To ensure that the different levels of digitisation are more uniform across the community, a definition for a ‘Minimal Information for Digital Specimens’ (MIDS) is proposed (see ICEDIG - MS35; part of WP6), which is hierarchically divided into four levels:

1. MIDS-0
   * The Digital Specimen Object (DSO; a digital representation of the physical specimen in a collection) only contains metadata and one or more media files. This level also includes the following three Darwin Core (DwC) elements that are related to the process of digitisation and collection management rather than the specimen.
     + [DwC:institutionCode](https://terms.tdwg.org/wiki/dwc:institutionCode) – from e.g., Index Herbariorum and other catalogues.
     + [DwC:collectionCode](https://terms.tdwg.org/wiki/dwc:collectionCode) – if exists, given by the institution.
     + [DwC:catalogNumber](https://terms.tdwg.org/wiki/dwc:catalogNumber) – automatically readable from the specimen label; must be attached to the specimen prior to imaging.
2. MIDS-1
   * Includes MIDS-0, but adds basic data elements that can be entered in bulk for a number of DSOs. Most scientific collections include this bulk information in their boxes and folders (plants), or drawers and units (insects). These elements typically are:
     + [DwC:scientificName](https://terms.tdwg.org/wiki/dwc:scientificName) – at some taxonomic level
     + [DwC:higherGeography](https://terms.tdwg.org/wiki/dwc:higherGeography) – at some accuracy such as ‘Europe’
3. MIDS-2
   * Includes MIDS-1, but adds data elements that have been transcribed from the specimen label, literally. These include: location, date, collector name, and scientific name. Many different DwC elements, often using the verbatim variety of the elements, can be used to describe these data elements, which can vary between collection types
4. MIDS-3
   * Includes MIDS-2, but adds interpretations. An example of this is finding the geographic coordinates of the collecting locality through research on gazetteers or field notebooks. Also an interpretation is asserting a taxonomic concept to the specimen ([DwC:taxonID](https://terms.tdwg.org/wiki/dwc:taxonID)) and the currently valid scientific name (MIDS-1 and MIDS-2 level scientific names are not necessarily the valid ones).

***It is important to note*** that additional data can be added at any MIDS level and may for example include images, sounds and DNA-barcodes. Thus, we focus here on levels of data registration and not imaging. As these levels are only used to identify digitised data, we include a separate level to identify that a part of the collection is not digitised at all and has not received a digital record in a CMS. When using the MIDS levels of digitisation in the CDD, it will be best to use a more informative name when presenting this information. This finally leads us to a preliminary proposal for the ‘digitisation’ classification (Table 8).

**Table 8.** Overview of the preliminary proposal for the ‘Digitisation’ classification, indicating the main group and subgroup of the collection description classification.

|  |  |
| --- | --- |
| **Main category** | **Subcategory** |
| Not digitised | Not digitised |
| Digitised | Minimally digitised (MIDS-0) |
| Regularly digitised (MIDS-1) |
| Fully digitised (MIDS-2) |
| Additionally or extensively digitised (MIDS-3) |

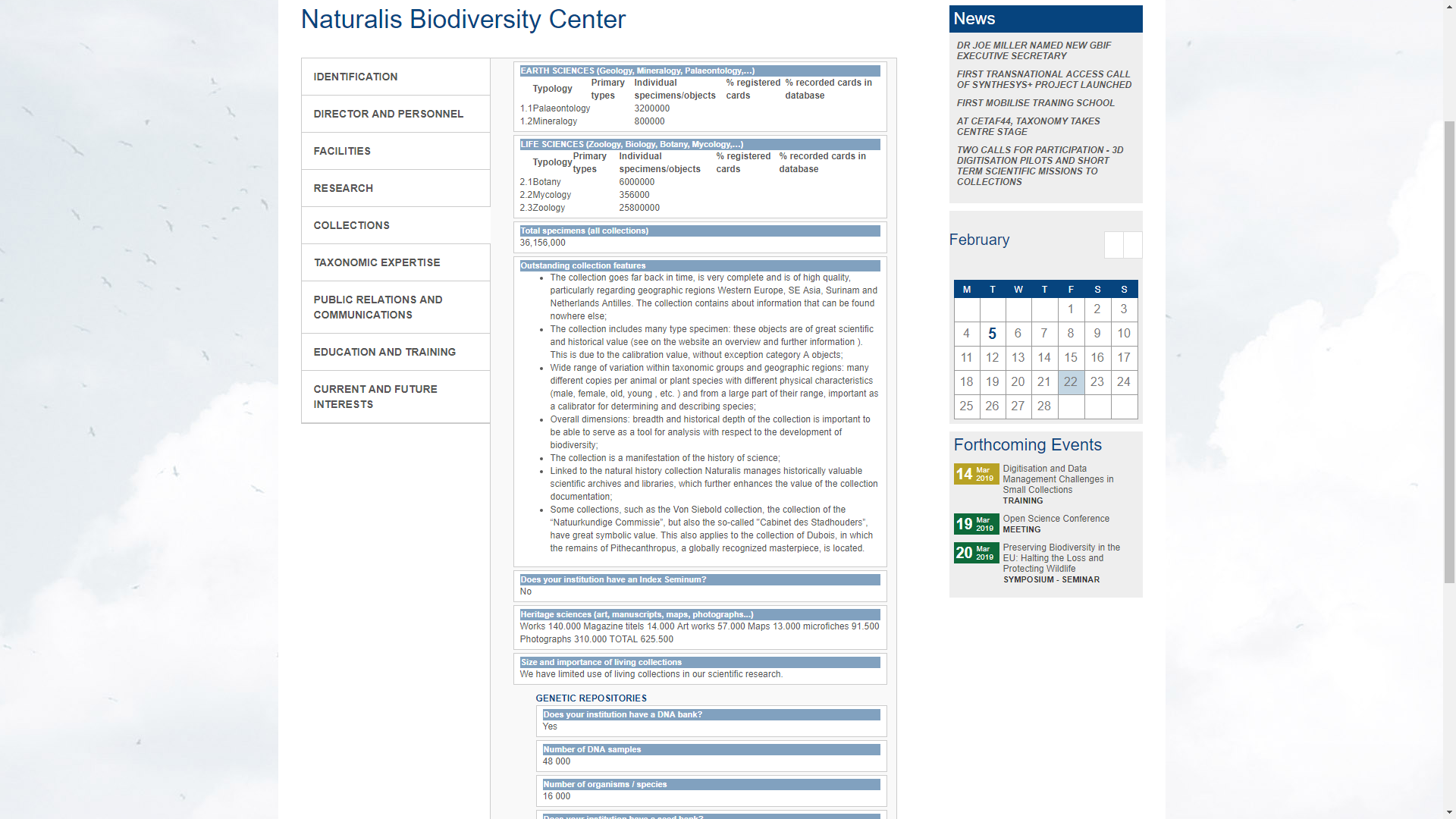
# Collection Digitisation Dashboard (CDD)

## Data acquisition and integration

A functional CDD depends on three quantitative data sources, a) the total estimated number of physical specimens in each classification category, b) the number of digitised records in each classification category, and c) the derived number of not digitised records in each classification category (b subtracted from a). These numbers are subsequently further subdivided to the different geographic regions (3.4 Geographic classification) and digitisation levels (3.5 Digitisation classification). The most challenging part is to obtain an accurate estimate of the number of specimens in each classification category. A specimen is defined as a physical object in an institutional collection that will be entered as a record in its CMS and gets assigned a Universal Unique Identifier (UUID). The number of digitised, and derived not digitised records per classification category can relatively easily be extracted from the institutional CMS.

**Key aspect of an operational CDD is therefore the metadata table** **with the estimated numbers of specimens for each of the three collection classification schemes** (Section 3.3). This provides an estimate of the entire European collection holding when data of all DiSSCo partners are merged. Subtraction of the digitised records allows highlighting collection digitisation gaps, both at taxonomic as well as at geographic levels. It allows identification of institutions with specimens of interest that are not yet digitised to drive digitisation-on-demand requests. For example, a question like ‘Which institution holds a fish collection from the Southern Ocean that is not digitally available yet?’ could be answered in this way. Furthermore, this can even indicate general collecting gaps.

To keep the CDD up to date requires updating the estimated number of specimens per classification category on an annual basis, or when new collections are added to the institutional holdings; and annually subtracting the number of digitised records per classification category. As soon as natural history institutions are fully committed to DiSSCo, we can request from each institution to keep their collection level data that will feed into the CDD up to date. Although we can ask for this data to be send to us every year, it would require quite some time to combine and prepare the data for the CDD. It may be more efficient to set up a system analogue to the [CETAF passports](https://cetaf.org/services/institutional-profiles) on which the CDD can rely for a more stable and easy to reach data input. The CETAF passports contain information on different aspects of the collection and the natural history institution itself, which is publicly and openly available. All CETAF members have an institutional page on the main CETAF website, where members can enter and adjust information regarding their collection themselves (Figure 7). Each CETAF member has a delegate who can update the institutional page whenever it is necessary, but are encouraged to do this at least once a year. The current CETAF passports could be used initially to harvest data on the digitisation of natural history collections to feed into the CDD as many institutions made data available through CETAF. However, the classifications and categories used to by CETAF is dissimilar from what we propose here, and not all DiSSCo partners are necessarily a partner of CETAF, thus we will miss some institutions. In the future, partners of DiSSCo should be united in one platform (e.g. on dissco.eu), where institutional pages similar to CETAF could be established. All DiSSCo partners can then manage their own institutional data that will ideally feed directly into the CDD as presented on the dissco.eu website.



**Figure 7.** A screenshot of a part of the institutional page of Naturalis on the CETAF website (CETAF passport).

## Dashboard visualisation

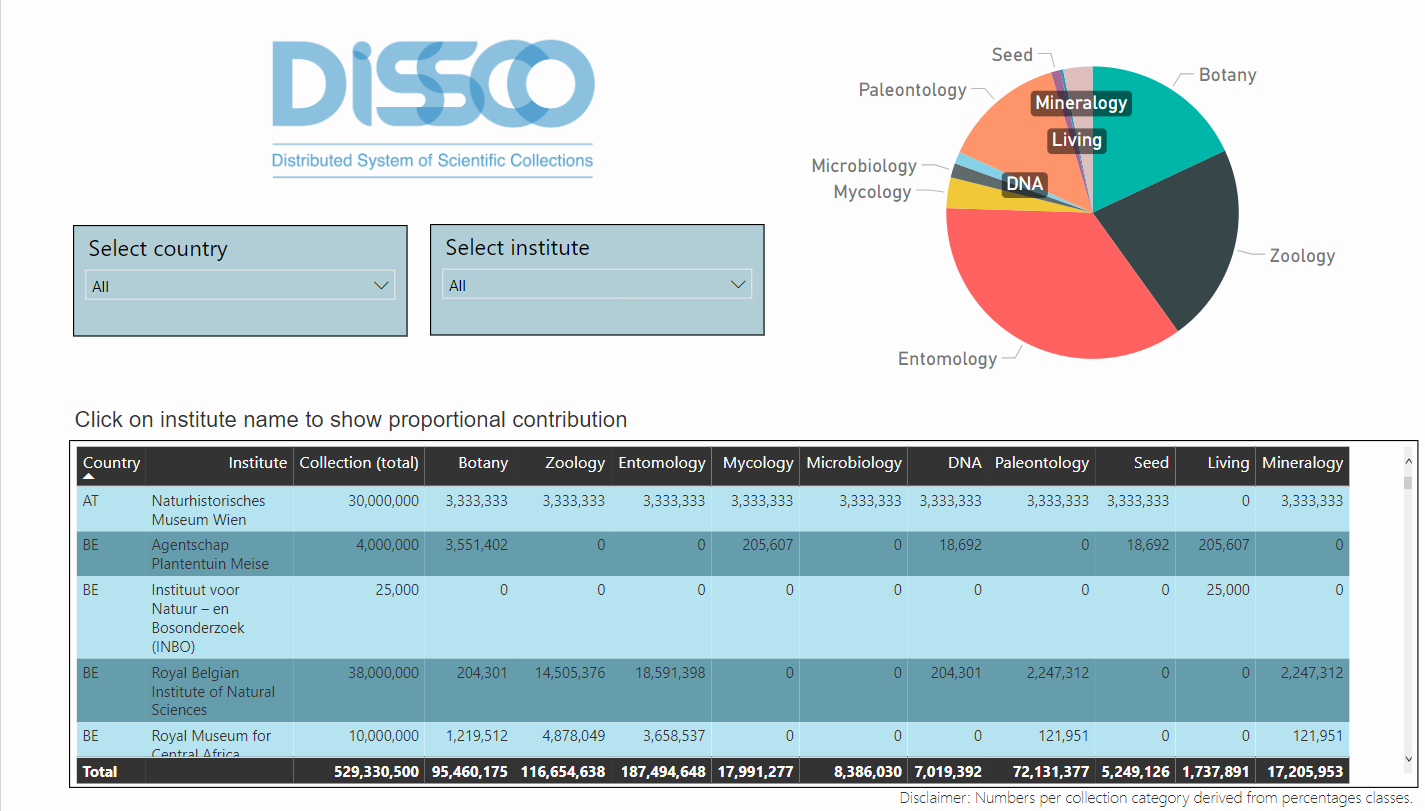
A collection digitisation dashboard allows visualising many different aspects of collection holdings. Based on the three preliminary collection classification schemes (3.3), the geographic classification (3.4), the digitisation classification (3.5) and the data obtained from the DiSSCo partners and Dutch collection institutes up to date (February 28, 2019), we developed two interactive CDDs to showcase different visualisation options. The **first CDD** is based on a survey that was send to all initial DiSSCo partners, including data from 89 collection holding institutes. This survey provides the best estimate of the total holding of European institutes, but does not include any indication of the spatial distribution of the collections, i.e. lacks a geographic classification. The **second CDD** is based on a pilot study held under 13 Dutch collection institutes that used a combination of the ‘Taxonomic’ classification (3.3.1), the ‘Geographic’ classification (3.4) and a ‘Digitisation’ classification (3.5) at the most basic level (digitised/not digitised), indicating the percentage of the collection that is digitised. Given that the digitisation classes MIDS-0 - MIDS-3 (3.5) are nested it should not be difficult to expand the number of digitisation classes. Although identified by the community, through iterative meetings of the TG CDD and workshops with the Dutch collection institutions, that the ‘Storage’ and ‘Stratigraphic’ classification are important, data for these two classifications are currently not available and therefore not included in the CDDs. Once digitised records with their digitisation MIDS levels can automatically be extracted from the institutional CMSs, it will also be possible to monitor digitisation progress through time based on the entry dates of the records in the respective CMSs.

Based on the user stories (section 2.3) and available data from the two surveys we identified the following prioritised visualisations:

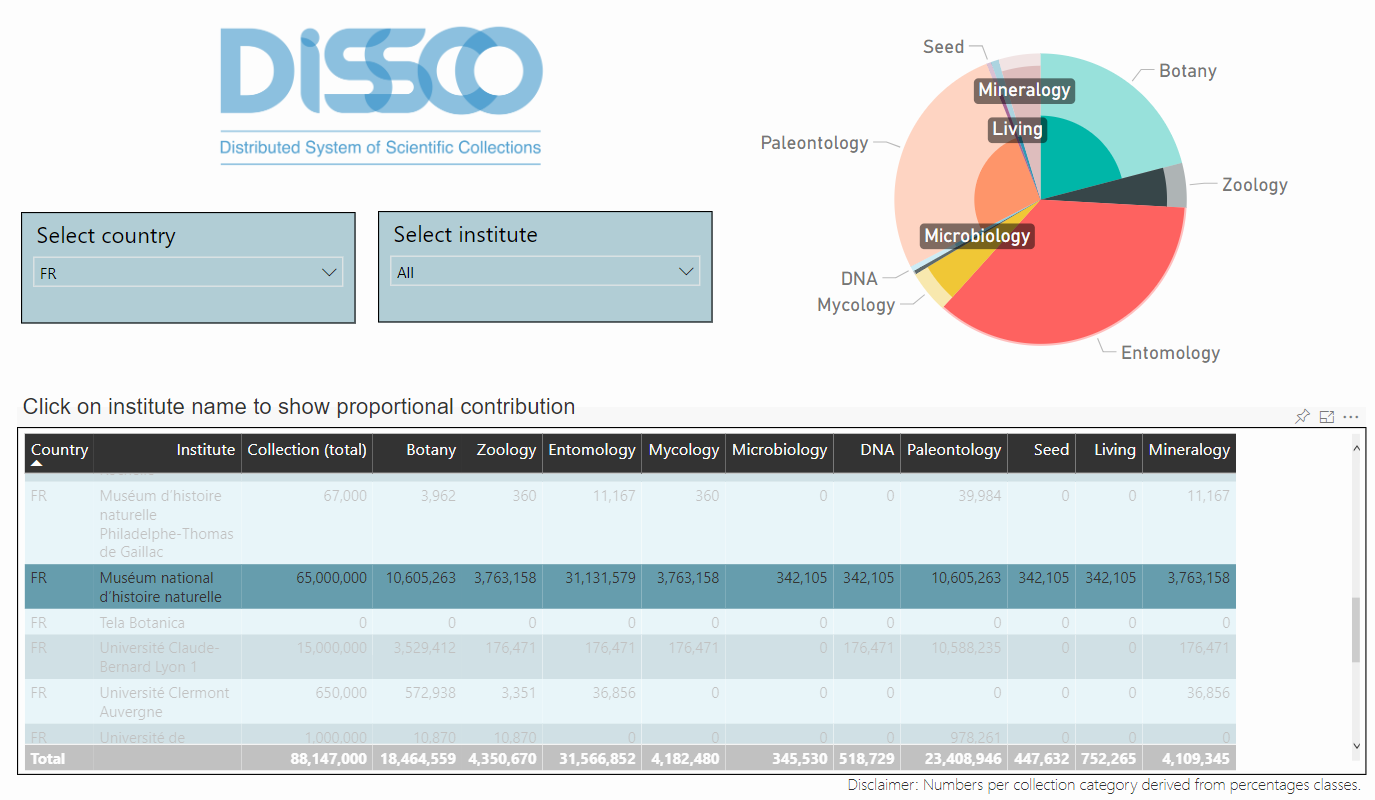
1. Overview of the entire European/Dutch collection holding according to the ‘Taxonomic’ classification. Data for the other two classifications (‘Storage’ and ‘Stratigraphic’) are currently lacking.
   1. Visualise the European\Dutch holding divided over the different taxonomic categories.
   2. Filter national and institutional holdings and visualise these as proportion of the total.
   3. Show quantitative summary statistics.
2. Show proportional and quantitative numbers of digitised versus not digitised data
   1. Filter by biome.
   2. Filter terrestrial and marine regions.
   3. Filter by institutional holdings.
   4. Filter by taxonomic category.  
      And combinations of the above.

### DiSSCo dashboard

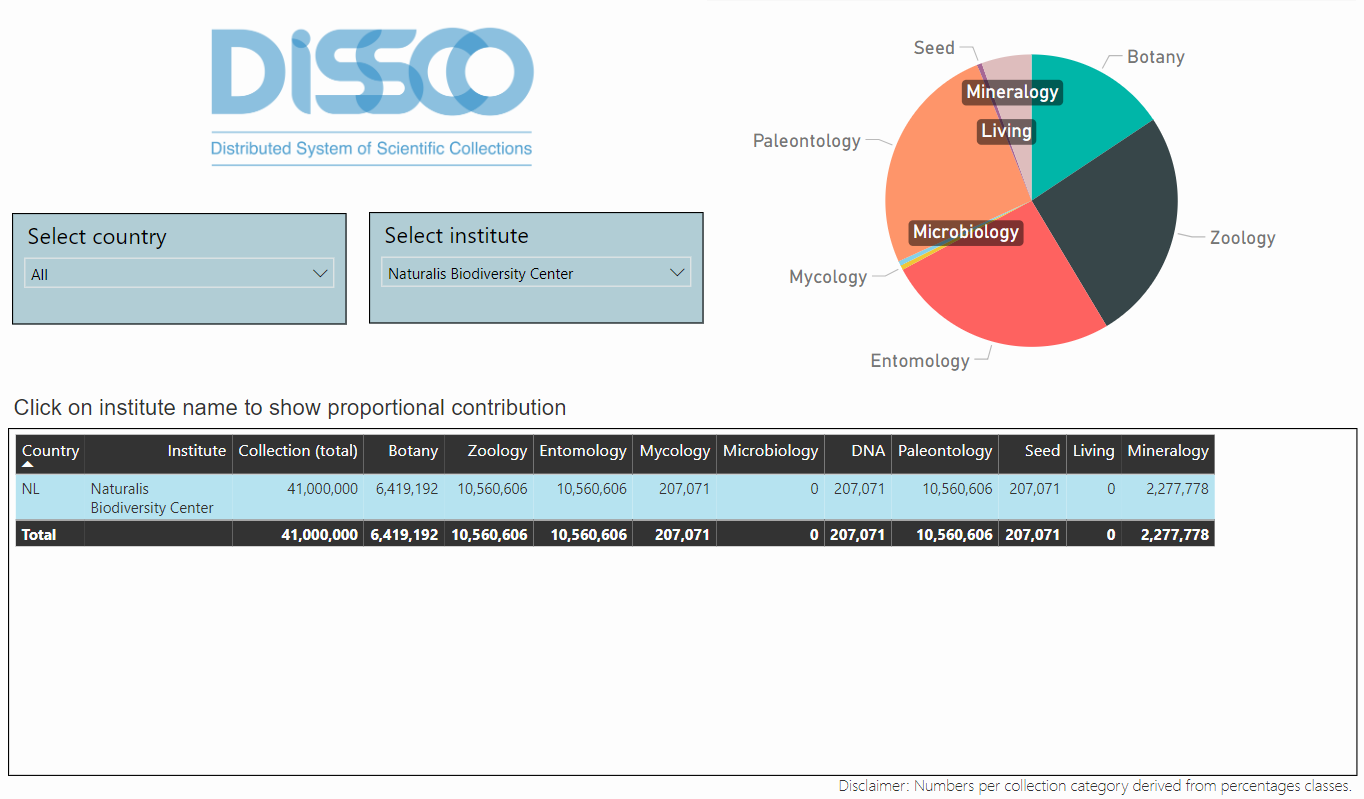
The live and interactive version of the DiSSCo dashboard can be found through this [link](https://app.powerbi.com/view?r=eyJrIjoiNWU5MWQzYmItYTUwZC00ZTE3LThjN2YtNTY0ZWFiOTNiOGZkIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9). Below we provided four screenshots/snapshots (Figures 8-11) of the DiSSCo dashboard based in the initial DiSSCo survey results with contributions of 89 DiSSCo partners. The DiSSCo consortium is continuously growing and the DiSSCo dashboard will be regularly updated. The figure captions describe what is shown in the dashboard snapshots.



**Figure 8**. Snapshot of the first page of the [DiSSCo dashboard](https://app.powerbi.com/view?r=eyJrIjoiNWU5MWQzYmItYTUwZC00ZTE3LThjN2YtNTY0ZWFiOTNiOGZkIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9) showing the distribution of collections over 10 collection categories. The selection boxes allow filtering for country and institutions. CTRL + click allows selecting multiple items. The table shows the approximate number of collections per category for all 89 institutes included in the initial DiSSCo survey.



**Figure 9.** First page of the DiSSCo dashboard showing the distribution of collections over 10 collection categories of the French DiSSCO partners [Select country: FR] with the contribution of the largest French collection institute, ‘Muséum National d’Histoire Naturelle’, highlighted in the pie chart.



**Figure 10.** First page of the DiSSCo dashboard showing the distribution of collections over 10 collection categories of Naturalis Biodiversity Center [Select institution: Naturalis Biodiversity Center].



**Figure 11.** Second page of the DiSSCo dashboard showing national contributions to the European collection held by 89 DiSSCo partners divided over 10 collection categories. The bubble graph shows the relative contribution of each country to the entire holding of the DiSSCo partners.

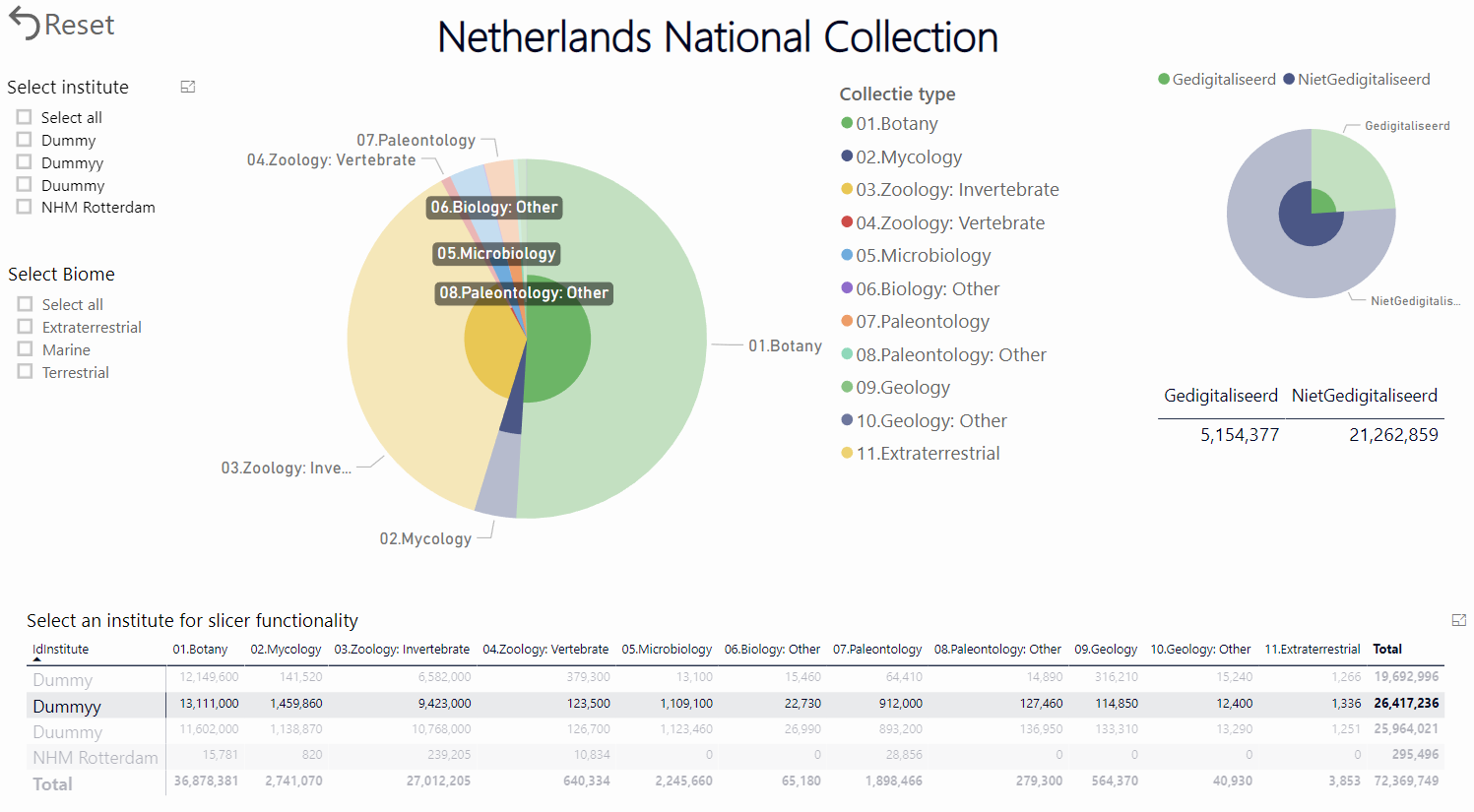
### Dutch national history collections dashboard

At the moment of creating this dashboard, we are awaiting data from the Dutch collection institutions and facing some Microsoft Power BI issues with spatial map selections (page 2 of the dashboard). To demonstrate the main functionality of this dashboard for the purposes of this deliverable D2.3, the current lack of data was overcome by introducing dummy data with a single digitisation level (digitised/not digitised).

Below we provide five snapshots (Figures 12-16) of the Dutch natural history collections dashboard (To see the live and interactive version please visit [dashboard of the Dutch collection institutes](https://app.powerbi.com/view?r=eyJrIjoiOThiNTg0NWYtM2Y4MC00NGJkLWIyNTItODRhODdiYjkyNDMxIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9)). The figure captions describe what is shown in the screenshots.



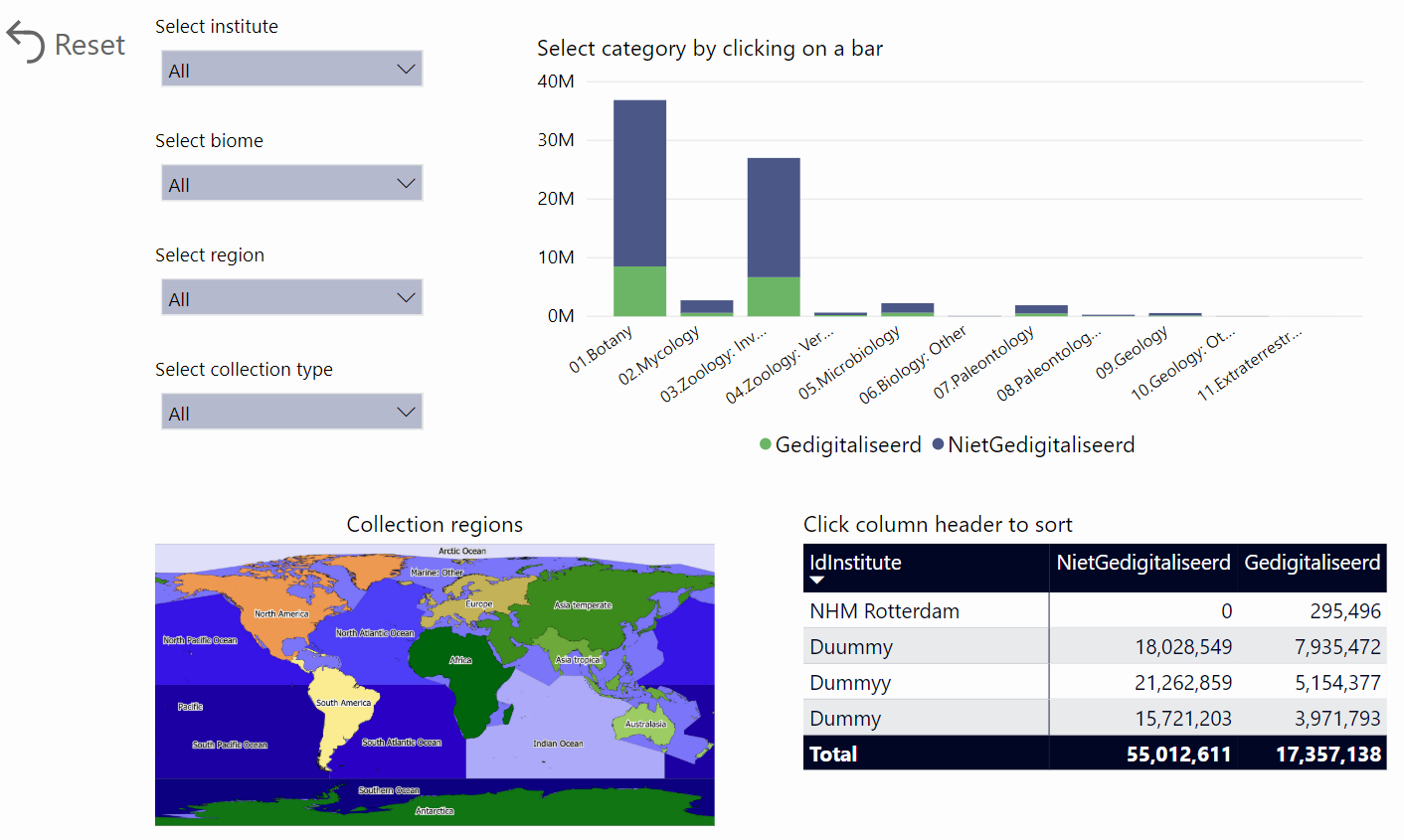
**Figure 12**. Snapshot of the first page of the [Netherlands National Collection dashboard](https://app.powerbi.com/view?r=eyJrIjoiOThiNTg0NWYtM2Y4MC00NGJkLWIyNTItODRhODdiYjkyNDMxIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9) showing the ‘entire’ Dutch collection divided over 11 main taxonomic categories based on the ‘Taxonomic’ classification (3.3.1).



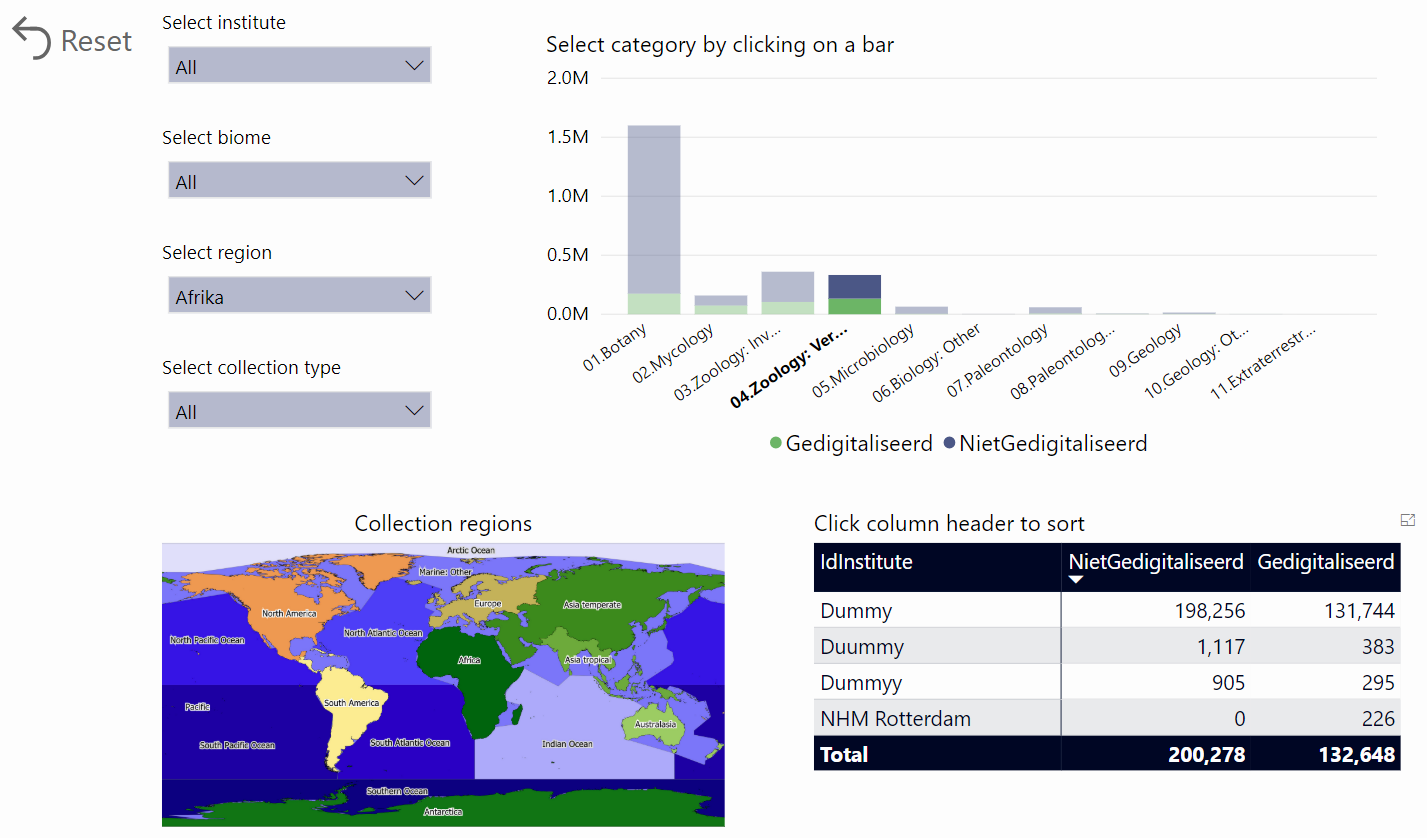
**Figure 13**. Snapshot of the first page of the [Netherlands National Collection dashboard](https://app.powerbi.com/view?r=eyJrIjoiOThiNTg0NWYtM2Y4MC00NGJkLWIyNTItODRhODdiYjkyNDMxIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9) showing the proportional contribution of a single institution [Dummyy] highlighted in the pie charts.

****

**Figure 14**. Snapshot of the first page of the [Netherlands National Collection dashboard](https://app.powerbi.com/view?r=eyJrIjoiOThiNTg0NWYtM2Y4MC00NGJkLWIyNTItODRhODdiYjkyNDMxIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9) showing the summary statistics for a single category [03 Zoology: Invertebrates].



**Figure 15**. Snapshot of the second page of the [Netherlands National Collection dashboard](https://app.powerbi.com/view?r=eyJrIjoiOThiNTg0NWYtM2Y4MC00NGJkLWIyNTItODRhODdiYjkyNDMxIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9) showing the summary statistics of digitised and not digitised data with all filter options.



**Figure 16**. Snapshot of the second page of the [Netherlands National Collection dashboard](https://app.powerbi.com/view?r=eyJrIjoiOThiNTg0NWYtM2Y4MC00NGJkLWIyNTItODRhODdiYjkyNDMxIiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9) showing the number of vertebrate specimens [04.Zoology: Vertebrates] for the African collection holdings [Select region: Afrika]. The summary table is sorted on [NietGedigitaliseerd] records indicating that institution [Dummy] has the largest holding of undigitised vertebrate zoology records.

## Final list of parameters to be included in the CDD

Based on the information obtained, presented and discussed in Chapters 2 and 3 of this deliverable, a final list of parameters to be (minimally) included in the CDD is proposed (Table 9).

The minimal number of parameters to include in the CDD is six: size, digitisation, institution, country, taxonomy and geography (Table 9). The minimal number of levels for the parameters size, digitisation, taxonomy and geography is 1 + 4 + 10 + 3 = 18, while the minimal number of data fields is (1 + 4) × 10 × 3 = 150 when crossing all levels. Depending on the final number of countries and institutions that will provide data to feed into the CDD, the minimal numbers of parameter levels and data fields will increase accordingly.

It is very important to realise that an increase in parameters and/or their levels results in an exponential increase in the number of fields that an institution must fill in and the complexity of the dashboard visualisation. Thus, it will be crucial to keep the CDD as simple as possible to ensure that the data can actually be provided by each institution. For the institution parameter, it will be important to standardise these across the CDD by using their full English institution name. The same holds for the country parameter, which should be standardised and in English as well.

**Table 9.** List of parameters that can (minimally) be included in the CDD.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Description** | **Levels** | **Required ?** |
| Institution | To indicate from which institution a (part of a) collection is | Similar to the number of institutions that provided data | **Yes** |
| Country | To indicate in which European country a (part of a) collection is kept | Similar to the number of countries that provided data | **Yes** |
| Size of collection | Estimated number of specimens within a (part of a) collection | Continuous factor | **Yes** |
| Taxonomy classification (3.3.1) | To indicate to which taxonomic category a (part of a) collection belongs | 10 main, 34 sub | Main: **Yes** Sub: No |
| Storage classification (3.3.2) | To indicate how a (part of a) collection is stored | 14 main, 48 sub | Main: No Sub: No |
| Stratigraphic classification (3.3.3) | To indicate to which stratigraphic group a (part of a) collection belongs | 3 main, 16 sub | Main: No Sub: No |
| Geographic classification (3.4) | To indicate to which geographic group a (part of a) collection belongs | 3 main, 19 sub | Main: **Yes** Sub: No |
| Digitisation classification (3.5) | The number of digitised records in a (part of a) collection | 4 nested MIDS levels | **Yes** |

# Conclusions and recommendations

Within this deliverable D2.3, we set out to prepare a design for a Collection Digitisation Dashboard (CDD), with the main purpose to make European natural history collections visible and discoverable. Below, we will identify our main conclusions and recommendations for the different aspects discussed in this deliverable.

Based on the Round Table on the CDD held within ICEDIG and the workshop for the Dutch collection overview dashboard (case study), we identified the following four main user communities for the CDD:

* Institution (director and collection manager)
* Government (policy maker)
* Non-government (nature association)
* Research (scientist)

The CDD is expected to be mostly used by these user groups for communication purposes, as a digitisation planning tool and the identification of key collections held by institutes. For example, a collection manager may be interested in the niche his/her institute holds in the (inter)national landscape and use this to see where improvements/enrichments in e.g. geographic scope of the (digital) collection can be made. Nevertheless, all user communities are expected to benefit in some way from a collection-level overview presented as a dashboard.

When preparing a collection data standard to be used as a guideline for what parameters to present in the CDD, we identified that a ‘Taxonomic’ classification would be most useful for its overarching purpose. In addition, a ‘Geographic’, ‘Stratigraphic’ and ‘Digitisation’ classification are used to further subdivide and characterise a natural history collection at a metadata level. At a minimum, the CDD must contain five parameters, each with a number of levels:

1. Institution
2. Country of institution
3. Taxonomy
4. Geography
5. Digitisation

This means that the total *size* of a natural history collection from an institution, given by the number of specimens, can be presented as a subset defined by these parameters (e.g. 100.000 out of a total of 250.000 digitised specimens of mosses from Europe in a Belgium natural history institution). Both main and subcategories were identified. It will be very important to keep in mind that the more the data is being subdivided into smaller categories, the more work it will be to obtain and maintain the data from these institutions. Also, it will become difficult to present the data in a clear way as the data become increasingly dense with an increasing number of smaller categories. A trade-off therefore exists between the desired granularity of the data and its presentation and the feasibility of creating a dashboard.

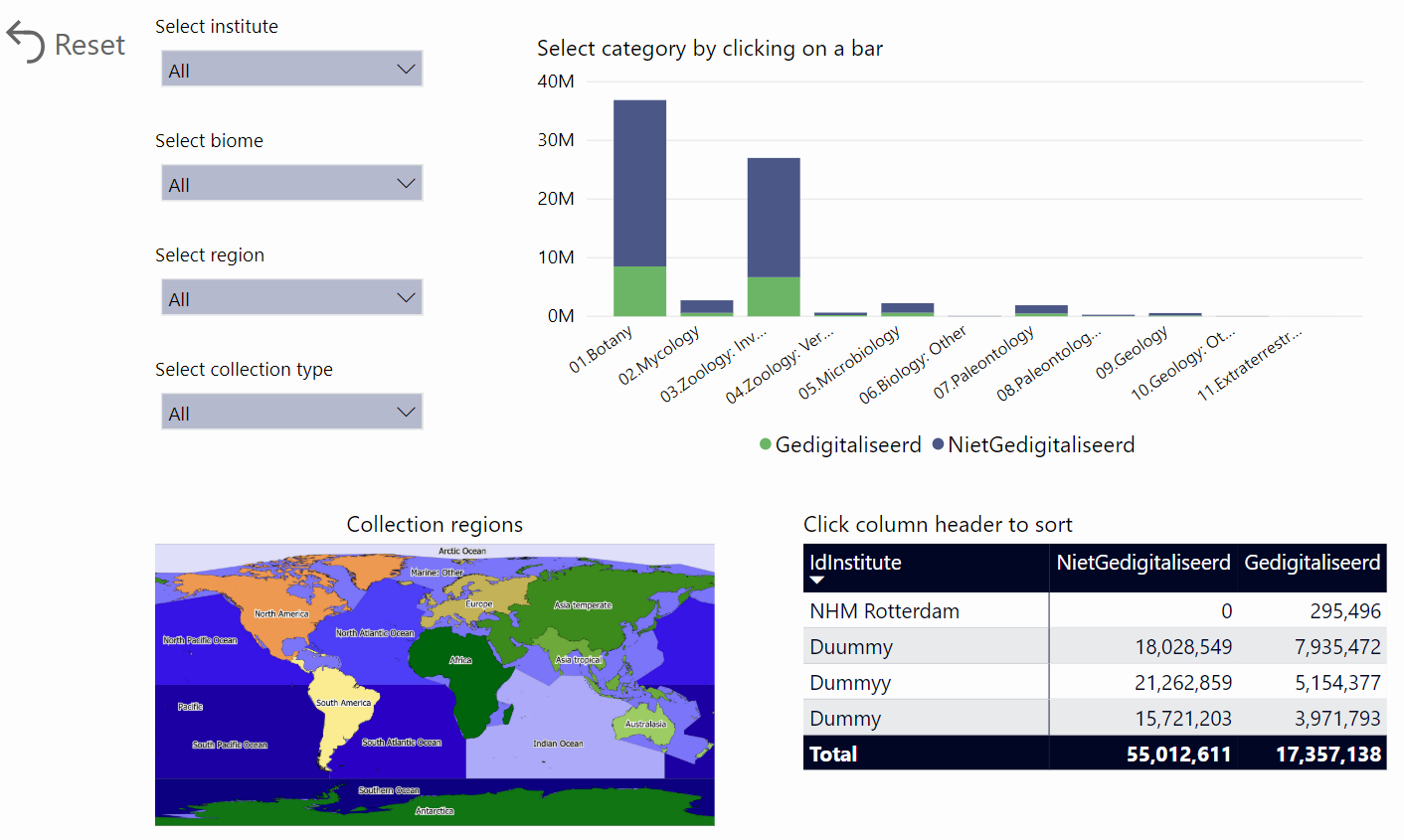
The most efficient way to collect the required data in a harmonised way is to ask natural history collections directly for the data needed by providing a template. For the dashboard on DiSSCo partners this was done through a survey, while for the Dutch collection overview dashboard information was obtained by sending an Excel file containing the required fields to fill in.

Regarding the dashboard visualisations, a dashboard may contain two important visualisations. The first visualises the size of the entire collection divided over different taxonomic categories and the individual institutional contributions (Figure 17).



**Figure 17**. The contribution of [Duummy] to the entire Dutch collection. The pie chart in the upper right corner shows the level of digitisation.

The second visualisation shows the spatial filtering options and allows identifying institutes with largest not digitised and digitised collection holding, filtered by biome, region and collection types (Figure 18).



**Figure 18**. The distribution of the total and individual institutional and taxonomic collections over different terrestrial and marine regions.

## Recommendations

Our recommendations and proposed actions to be taken up next are indicated in the following list:

* Propose the above collection descriptions as a community-accepted standard via TDWG CD task group. A first step is taken by having a TWDG workshop entitled ‘TDWG CD Task group meeting’ (GT51) during the biodiversity\_next conference to be held in October 2019. In this workshop the [TDWG Collections Descriptions (CD) Data Standard Task Group](https://github.com/tdwg/cd/blob/master/charters/task_group_charter/tg_charter.md) aims to provide a data standard for describing natural scientific collections facilitating 1) automated metrics using standardised collection descriptions and/or data derived from specimen datasets (e.g. counts of specimens), and 2) a global registry of physical collections (either digitised or non-digitised).
* Continue testing and fine-tuning the current two dashboards on a technical level to ensure all data is correctly and clearly presented.
* Ask for feedback on the dashboard visualisations from the main user groups as indicated above to identify if the CDD answers their main user questions. This is likely best done by another Round Table or via a series of short interviews with representatives from the main user groups. Alternatively, a targeted questionnaire could be sent out. In addition, the recommended parameters to be minimally included in the CDD can be evaluated as part of this effort.
* Publish the dashboards online (at the dissco.eu webportal at the European and national levels) as soon as they are finished.
* Prepare a platform for DiSSCo partners (e.g. on dissco.eu) where they can log in to enter and adjust data on their collections and institutions on an institutional page, in analogy to the CETAF passports. When this platform is in place, collection-level data can feed directly in the CDD and be presented on the dissco.eu website. Ideally, when data is adjusted on an institutional page, the CDD is automatically updated.
* Beyond this Deliverable 2.3, the work will be continued under the SYNTHESYS+ project which has a task dedicated to “integrate[ing] and expand[ing] institutional collection assessments”.
* Recognizing the need to automate as much of this process as possible, ensure that any discussion about harmonizing CMS across the DiSSCo network, includes a conversation about how to get this metadata more easily and how to engage the DiSSCo community about the importance of this resource.
* Recognizing the time constraints of those doing museum collections work, ensure that SYNTHESYS+ helps to contribute to a design that is scalable, elegant, easy to link to CMS, and employs the use of people identifiers (ORCIDs) (and their roles) to enhance the usability and automation possible. These developments are also furthered by the MOBILISE Cost Action.

# Acknowledgements

We would like to thank all participants of the two workshops for their committed time and valuable input. Furthermore, we are thankful for the active participation of the members of the Task Group CCD (Appendix 3) in preparing a collection classification scheme.

# References

Andrade, P., Bell, T., van Eldik, J., McCance, G., Panzer-Steindel, B., Coelho dos Santos, M., Traylen, S. and Schwickerath, U. (2012). Review of CERN Data Centre Infrastructure. *Journal of Physics: Conference Series*, 396 <https://doi.org/10.1088/1742-6596/396/4/042002>.

Berendsohn, W.G. and Seltmann, P. (2010). Using geographical and taxonomical metadata to set priorities in specimen digitization. *Biodiversity Informatics*, 7: 120-129.  
<https://journals.ku.edu/index.php/jbi/article/view/3988>.

Blagoderov, V., Kitching, I., Livermore, L., Simonsen, T., Smith, V. (2012). No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133-146.  
<https://doi.org/10.3897/zookeys.209.3178>.

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

Chiang, A.S. 2011. What is a Dashboard? Defining dashboards, visual analysis tools and other data presentation media. Accessed 9 July 2018 at <http://www.dashboardinsight.com/>

[articles/digital-dashboards/fundamentals/what-is-a-dashboard.aspx](http://www.dashboardinsight.com/).

Few, S. 2006. Information dashboard design: the effective visual communication of data. Beijing: O'Reilly.

Flanders Marine Institute (2018). IHO Sea Areas, version 3. Available online at <http://www.marineregions.org/>. <https://doi.org/10.14284/323>.

Hetherington, V. 2009. The Dashboard Demystified: What is a Dashboard? Accessed 9 July 2018 at <http://www.dashboardinsight.com/articles/digital-dashboards/fundamentals/the->

[dashboard-demystified.aspx](http://www.dashboardinsight.com/articles/digital-dashboards/fundamentals/the-).

Krishtalka, L., Dalcin, E., Ellis, S., Ganglo, J.C., Hosoya, T., Nakae, M., Owens, I., Paul, D., Pignal, M. and Thiers, B. 2016. Accelerating the discovery of biocollections data. Copenhagen: GBIF Secretariat. Available online at: <http://www.gbif.org/resource/83022>.

Smith, V., Paul, D., Woodburn, M., Grant, S., Singer, R., Love, K. 2018. Shining a New Light on the World’s Collections. Available online at:  
[https://www.idigbio.org/content/shining-new-light-world’s-collections](https://www.idigbio.org/content/shining-new-light-world%E2%80%99s-collections).

# Appendices

## Appendix 1. Details Round Table

### Summary

The Round Table is framed under Task 2.3 and was held on the 11th of June 2018 during the first ICEDIG All-Hands meeting in Leiden, the Netherlands. Twenty-one people attended, consisting of a mix of ICEDIG participants and external experts. A general introduction was given by Luc Willemse (Naturalis) on the scope of the Collection Digitisation Dashboard, which is to be designed within ICEDIG Task 2.3. The focus is initially on a dashboard showing collection level information to identify which collections has been digitised already and which collections still need to be digitised. Elspeth Haston (RBGE) then explained what is happening regarding internal dashboarding at RBGE. Wouter Addink (Naturalis) explained how different dashboards will come together within DiSSCo. Finally, Simon Chagnoux (MNHN) spoke about dashboard metrics related to citizen science projects. After the general introduction, there was a break-out in two groups: the first group focused on the end users, parameters and criteria and the second group focused on the technical aspects and unifying data.

In the first group, end users and their user stories were identified and listed. These were supplemented with what data elements (parameters) would be necessary to be displayed in a dashboard for each user story. A next step is to further identify which data elements are associated with each user story and whether user stories can be grouped based on the data elements. Together this will provide the basis for different kind of visualisations, including a dashboard, as indicated by the conceptual model on collection digitisation visualisations. In the second group, some technical aspects of the dashboard and how to bring together the data were discussed. Discussions were started from the data side, instead of the user side. The main conclusion is that it is essential to have a standard for the description of the collection, as to date this only exists for specimen level data. This is a requirement so all data can be unified and presented in the dashboard in a harmonized manner. Also, collection level data is already gathered in several ways, including the annual reports of institutes, so it would be good to combine these efforts to feed into the dashboard.

When regrouping again after the break-out, the chairs of each of the subgroups gave a summary of the outcomes. In the general discussion, it became clear that there are several initiatives that are related to collection description standards (e.g. the group of TDWG tackling description standardization as Natural Collections Description- NCD) and collection digitisation dashboards (e.g. a task group on CDD recently started by Naturalis;) It will be good to keep in contact and have an open communication to make sure we combine efforts and no duplication takes place.

### Participant list

General

* Ana Casino (CETAF)
* Agnes Wijers (Picturae)
* Myriam van Walsum (Picturae)
* Jeroen Bloothoofd (Picturae)
* Luc Willemse (Naturalis) – overall chair
* Emily van Egmond (Naturalis) – taking minutes
* Olaf Banki (Naturalis) – chair subgroup 1
* Wouter Addink (Naturalis) – chair subgroup 1
* Letty Stupers (Naturalis) – taking minutes

Subgroup 1- End users and parameters

* Niels Raes (Naturalis)
* Gwenaël Le Bras (NMNH)
* Jeremy Miller (Naturalis)
* Deborah Paul (IDigBio)
* Pierre-Yves gagnier
* Jaume Piera (chair working group ECSA)
* Jeroen Bloothoofd (Picturae)
* Luc Willemse (Naturalis)

Subgroup 2- Technical aspects and unifying data

* Hannu Saarenmaa (University of Helsinki)
* Andrea Hahn (GBIF)
* Elspeth Haston (RBGE)
* Dominik Röpert (BGBM)
* Robert Tiessen (Picturae)
* Simon Chagnoux (NMNH)
* Matt Woodburn (NHM)

### User stories

**Table A1.** List of all collected user stories during the Round Table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **User groups** | **As a** | **I want to** | **So that** | **For this I need (data elements)** | **Level of digitization** | **Digitized/non-digitized** |
| Media | Journalist | Link to primary source data (scientific literature, museum collections databases etc.) | My readers can learn more about the topic of an article | Collections database records | Specimen | Digitized |
| Governmental | Policy maker | Information on the distribution of species under the nature directives | Assess conservation status and distribution range | Detailed distribution data | Specimen | Digitized |
| Collection | Collection Manager | Check in which institutions certain collection categories are kept so that I can forward a collection on offer to an institute that is interested | I can forward this information to a collection holder | Details about taxonomic/geographic specialism and possibly wish lists for certain specimens | Storage/species | Digitized |
| Institution | Director | Hire a curator with knowledge of specific groups | I can be sure they have a background that includes knowledge of the main collection | Collection types, importance of collection gauged by size, scope, and time period | Collection | Both |
| Citizen science | Citizen scientist | Know where was a certain collector on a certain day | To help transcribe a specimen | Existing transcription of specimens collected around the same time by the same collector | Specimen | Digitised |
| Industry | Solution provider | Build and provide solutions and related services | The keepers and scientists can work better and easier with their collections for less cost | Volumes, locations and physical sizes plus an insight on what is digitally represented and what not. Even better would be if there is an institutions priority as to what needs to be digital first | Collection and partly storage level | Both |
| Research | Scientist | Model South East Asian biodiversity patterns | To gain an answer to a scientific question | Detailed taxonomic and geographic information | Specimens | Which institutes hold the largest non-digitized collection |
| Non-governmental | Association | To gather information to have overall figures representative of partners' state-of-the-art | We can showcase the relevance of collections to policy makers and attract funds | High-level figures that feature the collections as a whole | Collections | Both, digitized and non-digitized information are valuable (to indicate the progress and the support needed, respectively) |
| Research | Scientist | Query when and where one or more species have been recorded, and their characteristics, and the institutions that archive specimens | I can collect more specimens, or borrow collections | Taxonomic fields, geographic coordinates, date of collection | Specimen | Digitized |
| IT | Software developer | Create new usages with the data and ways to add to the data, through apps or web interaction | Data is more accessible to the masses and different collections can be, for instance, cross-referenced. At the same time additional data can be added and fed back into the core databases. Geographic location will be involved as every man has GPS access today. The vantage point to access these 'big data' sources could be educational, entertaining, medical, historical and natural sciences | Scope: Collection level, details: Specimen level | Specimen | Digitized |
| Citizen science | Citizen scientist | Help with transcribing | I can enjoy this voluntary work | Images without transcription | Specimen | Partly digitized |
| Governmental | Policy maker | Know the use of the collections by other domains as a key indicator of its impact | I can distribute resources and allocate them in alignment to the strategic priorities of the government that I represent | Access to the collections, virtually and physically, from different types of users | Collection | Both, digitized (publicly available) and non-digitized (to understand the need to bridge the gap) |
| Education | Curious person | Learn about the species that might be in my environment | I can improve my bioliteracy | Taxonomic fields, common names, geographic coordinates, species characteristics, images | Specimen | Digitized |
| Citizen science | Citizen scientist | Be recognized as contributor | I can apply for funding to digitize my own collections | Contribution indicators | Could be at all levels | Digitized |
| Institution | Director/administrator | Know what makes our collections unique | I can effectively advertise/highlight the collections to improve usage | Collection types, with size, locality scope, time, taxonomic scope, important collectors | All levels | Both |
| Collection | Collection Manager | Start a digitizing project | I like to digitize a certain group of my collection, I like to do this internationally because of funding | Know where else there are collections of this group | All levels | Digitized |
| Citizen science | Citizen scientist | Be recognized as contributor | I can identify my contribution on validating data from external sources | Contribution indicators (as validator) | Specimen | Digitized |
| IT | Solution provider | Tap into the vast market of digital storage solutions for digital natural collections | I can sell my services and consult | Predictable numbers on collection type, volume and progress in digitization | Collection | Both |
| Collection | Collection manager | Redirect a researcher to colleagues | They can examine more collections | I need to know which institute holds specific kinds of collection | Species | Digitized |
| Institution | Collection manager, Director, Administrator | Know the situation with collection sizes | I can plan for new space/storage needs | I need to know existing sizes of collections, and the number of new material coming in. Also, need to know status/condition (e.g. wet, dry) of existing material. Also collection health information. | Collection, species | Both |
| IT | Automatic identification systems developer | Which collections are available to use as a reference (training data set) | I can training my algorithms for automatic identification | Collections of target species (validated) | Collection, species | Digitized |
| Citizen science | CS site manager | Select a load of images | To build a CS project | Basic elements on the images | Specimens | Partly digitised (images + OCR results, other projects result) |

## Appendix 2. Details Dutch collection overview – NWO-ALW

### Background

In the Netherlands, the NWO-ALW project was funded to connect stakeholders (collection managers, researchers and IT) on a national level and prepare them for RI developments such as DiSSCo. This project will focus on four themes: Data (Dutch collections and registration policy), Technics (infrastructural tooling and national infrastructure requirements), Usage (user perspectives) and Added value (what conceptual, semantic and communicative changes lead to added value of data). The meetings regarding the Dutch collection overview of the natural history collections are part of the Data theme.

The current state of information regarding the Dutch natural history collections is limited. Information of previous investigations is often disconnected, spread out over multiple sources and at times, lost. Thus, composing a quantitative overview of the holdings within the Netherlands with additional information on institute specialties and visions is needed. A current and coherent overview will show national institutions as well as other stakeholders what is present in the Netherlands and is pivotal to form a common national collection policy. The Dutch collection overview will be presented as a dashboard.

### Participant list

* Museon (Den Haag)
* NHM Rotterdam
* NIOZ
* Stichting de Bastei (Nijmegen)
* Groningen University collection
* Natura Docet Wonderryck Twente
* Natuurmuseum Brabant (Tilburg)
* Naturalis Biodiversity Centre
* NLBIF
* Natuurmuseum Fryslan
* Teylers museum
* Universiteitsmuseum Utrecht
* Wageningen WUR/NWVA

### User stories

**Table A2.** List of all collected user stories during the Dutch collection overview workshop from the NWO-ALW project.

|  |  |  |
| --- | --- | --- |
| **As a**  **(who)** | **I want to**  **(what)** | **So that**  **(why)** |
|
| Exhibition designer | Know which musea have which taxonomic groups in their collection. | Obtain knowledge and objects for my exhibition. |
| Director of an institution | Know which taxa are already present in the ‘Dutch collection’. | Better evaluate offered donations to the collection. |
| Policy maker | Know which institution has knowledge of nature and ecology in my domain. | Develop policy for nature conservation. |
| Nature enthousiast | Know where I can find certain animals/insects/plants. | Identify my own observations. |
| Employee of insect knowledge centre | Distribution information of specific insect species. | Make a reliable distribution map. |
| Researcher in Japan | Information about land snails of the Philippines before 1900. | Make a revision or overview. |
| Child with self-excavated object from the garden | Find pictures of bones. | Discover that it is the bone of a cow. |
| Potential borrower/artist | Find pictures of mounted/stuffed animals | Find an animal in a specific pose. |
| Local governance body | Gain insight in what is already known about a natural science object. | Use and/or promote interesting or unique information of the local nature within activities. |
| Scientific researcher | Find a specimen, species group, time period and/or locality. | Wish that the collection is a source and not merely cultural heritage. |
| Collection manager | Know where specific taxa and/or types objects can be found and related expertise can be found. | Improve my collection by re-positioning objects. |
| Collection manager | Know what the modern name is of a fossil (current name is over 30 years old). | Keep the knowledge within my collection up to date and improve the discoverability of objects for researchers. |
| Exhibition designer | Find artefacts that are related to specific persons, institutions or a time period. | Give substance to a historical story in an exhibition. |
| Researcher | Discover if samples/specimens that are unknown to me could be used for my purposes and what additional material is available internationally and under what conditions. | Not relevant, unless the user can only be helped when this is known. |
| Author of a collection policy plan | Determine the position of my own collection within the ‘Dutch collection’. | Account for this when new policy is being formulated and written. |
| Exhibition designer | Know which objects are relevant to my exhibition, in which institutions these can be found and what the quality of the objects is. | Determine what to include in my exhibition and where to obtain these objects. |
| Policy maker | Gain insight in the state of digitisation and the use of Dutch collections. | Determine whether funds are being well-used. |
| Curator/Collection manager | Get in contact with colleagues elsewhere. | Exchange knowledge/information. |
| Collection manager | Know what else is present within the Netherlands. | Know what to collect or de-collect. |
| Researcher | Know what mounted/stuffed birds are available. | Perform color research. |
| Curator/Researcher | Compare my collection of a certain species/taxa with similar collections elsewhere. | Expand my research on a certain species/taxa and (potentially) work on a publication. |
| Citizen scientist/Nature enthousiast | View the ‘Dutch collection’. | Compare and/or interpret my finding/observation. |
| Collection manager/Director | Know what the distribution is of geographic collection locations within the Dutch institutions. | Analyse and perhaps sharpen our geographic focus. |
| Curator/Collection manager | Gain insight in focus areas (geographic and taxonomic) of other institutions. | Strengthen and refine the niche of our institution. |
| Specialist | Know which specimens of my interest/species group are available. | Compare, measure and sample specimens for study. |
| Curator/Researcher | Know which objects can be found where objects, but in particular on a detailed level such as collector or origin. | Perform scientific research. |
| Palaeontologist studying fossil vertebrae of monitor lizards | Know which institutions have fossil vertebrae of monitor lizards. | Test whether my presumptions are correct. |

## Appendix 3. Details Task Group CDD

**Table A3.** List of current members of the Task Group CDD (March 2019).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Name** | **First name** | **Organisation** | **Email** |
| 1 | Addink | Wouter | Naturalis Biodiversity Center | wouter.addink@naturalis.nl |
| 2 | Casino | Ana | CETAF | ana.casino@cetaf.org |
| 3 | Cocks | Naomi | Natural History Museum | n.cocks@nhm.ac.uk |
| 4 | Gödderz | Karsten | CETAF | karsten.goedderz@cetaf.org |
| 5 | Haston | Elspeth | Royal Botanic Garden Edinburgh | EHaston@rbge.org.uk |
| 6 | Koivunen | Anne | LUOMOS | anne.koivunen@helsinki.fi |
| 7 | Lahti | Kari | LUOMOS | kari.lahti@helsinki.fi |
| 8 | Love | Kevin | Florida Museum | klove@flmnh.ufl.edu |
| 9 | Motz | Gary | Indiana University | garymotz@indiana.edu |
| 10 | Paul | Deborah | iDigBio | dpaul@fsu.edu |
| 11 | Petersen | Mareike | Museum fūr Naturkunde | Mareike.Petersen@mfn.berlin |
| 12 | Raes | Niels | Naturalis Biodiversity Center | niels.raes@naturalis.nl |
| 13 | Smith | Vincent | Natural History Museum | vince@vsmith.info |
| 14 | Trizna | Mike | Virginia Polytechnic Institute and State University | TRIZNAM@si.edu |
| 15 | van Egmond | Emily | Naturalis Biodiversity Center | emily.vanegmond@naturalis.nl |
| 16 | Woodburn | Matt | Natural History Museum | m.woodburn@nhm.ac.uk |