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 D4.4

Work package **WP 4** / Task **4.3** / Lead: **APM**
Deadline **M16**

Interoperability of Collection Management Systems

Authors: Mathias Dillen[1], Quentin Groom[1], Alex Hardisty[2]

Contributors: Frédérique Bakker[3], Marian van der Meij[3], Sarah Phillips[4], Henry Engledow[1], Matt Woodburn[5], Urmas Kõljalg[6], Anniina Kuusijarvi[7], Zhengzhe (John) Wu[7], Simon Chagnoux[8]

1 - Meise Botanic Garden (Meise - Belgium)

2 - Cardiff University (Cardiff - UK)

3 - Naturalis Biodiversity Center (Leiden - the Netherlands)

4 - Royal Botanic Gardens, Kew (Kew - UK)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

5 - Natural History Museum (London - UK)

6 - Natural History Museum, University of Tartu (Tartu - Estonia)

7 - Finnish Museum of Natural History (Helsinki - Finland)

8 - National Museum of Natural History (Paris - France)

Summary

The collection management system (CMS) is a key tool for an institution. It provides numerous functions in cataloging and managing a collection, but is also the source of data that underpins research. It is a large investment for an institution and it must be supported over many years, and managed through its lifespan, until inevitable data must be migrated into a new system.

Currently, there are an enormous variety of systems in use across Europe. Some are as basic as spreadsheets, but others are complex, relational databases tailor made for the needs of a single institution. These systems differ in their underlying data model for specimens and in the vocabulary they use, which makes interoperability between systems difficult.

The planned DiSSCo infrastructure envisages two-way interoperability of data between a central datastore and collection management systems at numerous institutions. To achieve this goal will require considerable harmonization between these systems.

In this report we make recommendations to curators, managers and developers of collection management systems with the aim of improving interoperability. We also take a look at the suitability of today's data standards, in particular Darwin Core, to support the planned DiSSCo infrastructure. Our findings are based on surveys into collection management systems and their qualities, as well as interviews with system managers and analyses of published data from different systems. Ideally, observations of biodiversity could flow seamlessly between collection management systems, aggregators and researchers, but currently we are far from that ideal.

Key recommendations

1. **Institutions should not create their own collection management system. If they are inclined to do so, they should consider its long-term sustainability, the cost of maintenance and the short lifespan of any software.**

2. **DiSSCo should both encourage and facilitate data migrations into CMS's with better support, with the aim of reducing the overall number of different types of CMS in operation.**

3. **Institutions should implement in their CMS's links of their conceptual data, such as instances of people, habitats and species, to external resources through unique identifiers. If these identifiers are used across multiple CMS's, data from these systems can be more easily harmonized.**

4. **Versioning of data in CMS's should be addressed urgently in parallel with the storage of annotations. This will require the redesign or replacement of many collection management systems.**

5. **Ideally, it should be possible for every CMS to export data and re-import these data seamlessly and in a standardised manner. Also, CMS's should be designed**

in such a way that all data related to a certain specimen can be efficiently and effectively queried, as well as updated.

6. **A push is needed for the development and further adoption of controlled vocabularies for specimen data and the adoption of more standard formats specifically tailored to the needs of specimen data.**

7. **Globally unique stable identifiers should be adopted by all institutions. Institutions should recognise the importance of stability in identifiers and put procedures and policies in place that maintain stability. Care needs to be taken to distinguish specific aspects of identifiers, such as barcodes and resolvable URIs, and the difference between the physical specimen and its digital versions.**

8. **It should be best practice to use a value of “unknown” for the case of missing information where information is known not to be known, and a value of ‘empty’ for when information is known but not digitized or unclear whether it is one or the other.**

9. **Versioning, annotation and verbatim data should be considered more centrally in the development of data standards so that all the knowledge on a specimen can be captured, together with all the curational and interpretation steps.**

1. Introduction

The Distributed Systems of Scientific Collections (DiSSCo) envisages a pan-European digital infrastructure of biological collections held by natural history museums, herbaria and other institutions. These collections can be leveraged for scientific research and to help resolve some of the global problems, such as biodiversity loss and climate change adaptation. The European community of biological collections has agreed to work together to achieve this aim and bring their collections, in so as far as it is possible, to work together in a seamless infrastructure from the perspective of the users.

To achieve this goal, information technology will be used to bring together geographically, taxonomically and linguistically disparate collections. Therefore, interoperability of data is a key issue to achieve DiSSCo. Interoperability is a problem with many levels including issues related to data standards, software and regulations, but here we focus on the issues of semantic data interoperability, particularly at the level of the institution.

Most natural history museums and herbaria have some form of catalogue of their collection. Traditionally, these may have been maintained on paper. However, in most institutions such a catalogue is nowadays being maintained digitally in what is called a Collection Management System (CMS). A digital specimen catalogue may include data about the identity of the specimens, the person(s) who collected them, the date of collection and the location where they were collected (Carpinone 2010). In addition to providing an overview of what is present in a collection, these data may also allow specimens to be grouped together or searched for by a trait they have in common, even if they are stored in different physical locations or collections (Baird 2010). Data validating and enriching approaches may also be implemented on these systems, such as appending missing data informed by duplicates in other collections or automated georeferencing (Nelson et al. 2018). While most institutions have a digital catalogue, most do not have a complete catalogue. Even if every accession were recorded in such a database, it would be an exceptional institution where every detail of every specimen were entered in the database (Vollmar et al. 2010). Indeed, this is practically impossible as the data are dynamic, with new identifications, new citations and new specimens being added constantly.

With rapid technological advances in the storage, capture and transmission of digitized data over the past two decades, many collections have initiated large scale digitization and data publication programs. Specimens are imaged in (semi)-automated workflows and data associated with them are published to institutional portals and other repositories (Schindel & Cook 2018). This facilitates data aggregation and specimen findability for scientific research, but also requires the implementation of data standards to ensure optimal findability and accessibility of data and semantic interoperability between data from different sources (Beaman & Cillinese 2012). Biological data standards such as Darwin Core (DwC) and Access to Biological Collection Data (ABCD) have been created to address this need (ABCD task group 2007, Wieczorek et al. 2012). To ensure optimal access and (re-)usability of data, the FAIR data principles were drafted, specifying the importance of proper licensing, unique and persistent identification, clear provenance and standardized access protocols (Wilkinson et al. 2016).

Such standards are becoming commonplace now and becoming ever more widely adopted. This makes them an easy choice for the numerous data transactions and annotations that will be made in the context of DiSSCo. However, there will always be a trade-off between the flexibility of how data standards are implemented and how effective they are at ensuring findability and semantic interoperability (Scholes et al. 2017). Darwin Core, for example, was designed to "*create a loose federation of databases*", where "*the barriers to publishing data [...] were purposefully kept as low as possible*" (Wieczorek et al. 2012). DwC may fall short for following FAIR data principles, if additional measures are not followed. DwC does not fix any controlled vocabularies for its terms, though some are suggested and some, such as `dwc:establishmentMeans`, are validated when data are published to the Global Biodiversity Information Facility (GBIF). For most terms in DwC, a wide variety of controlled and free text vocabularies are used. This has led to different interpretations of the terminology and gaps in the established terminology. In this report, we will describe the obstacles to an optimal data flow from institutions to researchers and provide recommendations to the stakeholders in this process.

Even if optimal data harmonization between the various portals and repositories of biodiversity data such as GBIF can be achieved, it can be expected that institutional CMS's will remain a fundamental authority for data concerning specimens in their care. This will require workflows for data enriched and annotated in the DiSSCo infrastructure to interchange with the CMS. For optimal functioning of DiSSCo, it will also be needed that as much data concerning specimens as possible are published and are not restricted to CMS's themselves for any reason other than (institutional) policy. For these reasons, we will also examine in this report the CMS infrastructures and what problems they encounter for streamlined data import and export. In particular, we will investigate whether certain types of data do not get published and if modifications to or recommendations for data standards could address this. Finally, we provide some general recommendations for the designers and managers of CMS's for improvements in the context of the role they would play in DiSSCo. For all these assessments, we will take into consideration user stories compiled under ICEDIG Work Package 6, indicating relevant demands for interoperability of data at different levels. Our recommendations will also take into account some of the emerging principles of data management as discussed within ICEDIG's Work Package 6 and which will be expressed in the DiSSCo data management plan.

1.1. Functions of Collection Management Systems

Most collections are arranged taxonomically or, to varying degrees, geographically or historically. However, in some institutions, special collections of particularly significant collectors are kept apart from the general collection. Some specimens may be exhibited while others are in storage. Still, in most cases, if someone wants to find a specimen of a particular taxon or character, they can find it manually by following the ordering of the collection. For any search with greater complexity, a digital catalogue is essential. Examples of such queries are finding all the specimens collected by a collector, finding nomenclatural type specimens or finding the earliest specimens of a particular taxon (Carpinone 2010).

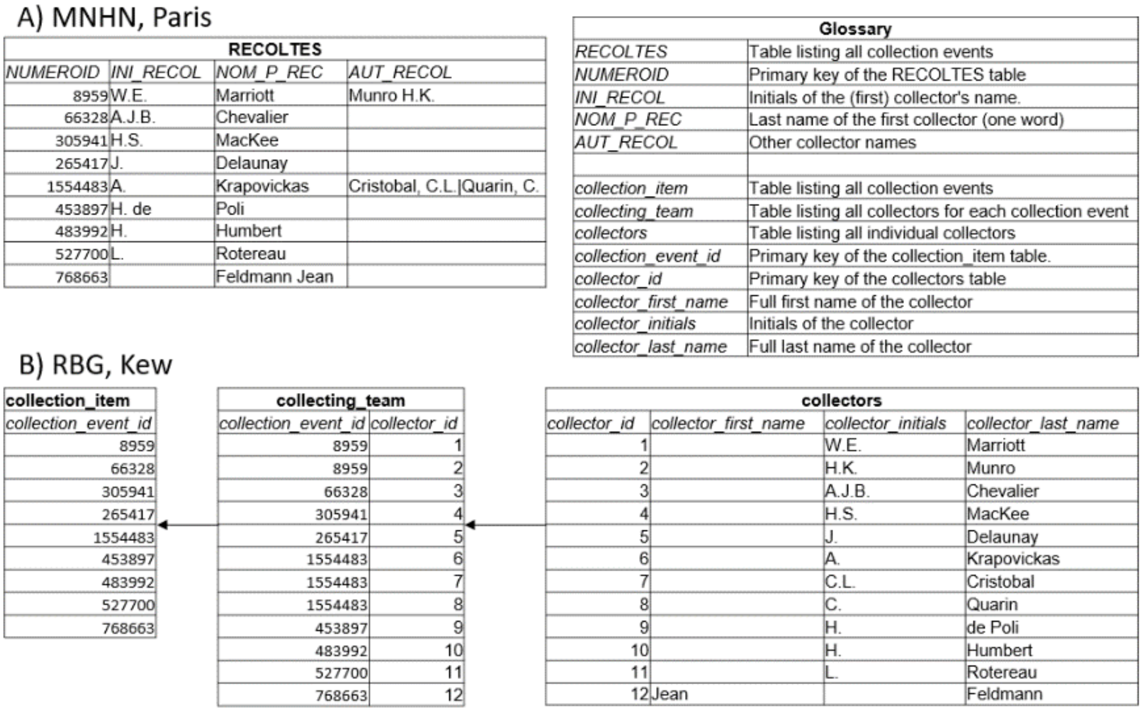
Such queries may be important for researchers who wish to use specimen data for research into conservation, taxonomy, biogeography and the history of science. Data managers are constantly enriching and correcting specimen data, in addition to making new types of data available. Morphological traits, for example, are important for identification and classification of taxa, but functional and chemical traits extracted from the specimens themselves can also give insights into historical pollution levels and atmospheric changes, such as the increase in atmospheric carbon dioxide during the past two centuries (Lavoie 2012).

Of course, a CMS also play an important role in daily management and planning. It provides curators and policymakers with an up-to-date dashboard of what is present and in what condition in their collection. In addition, loans and specimen exchanges have been part of taxonomic research before the time of Linnaeus. Even with the best quality images of a specimen, one cannot completely replace the need to examine specimens in person. Institutions must track loans that they send out and those that are returned. A CMS can keep track of these loans and the dates and people responsible.

1.2. Interoperability of data structures

The problem of semantic interoperability is "*the difficulty in integrating resources that were developed using different vocabularies and different perspectives on the data*" (Heflin and Hendler 2000). In general, semantic interoperability of specimen data can be broken down into three levels. Firstly, there is an issue of the data model used, secondly there are the formats and terms for individual concepts and thirdly there is incompatibility in the vocabularies for elements within concepts.

At the highest level, datasets stored using different data models require conversion algorithms to make queries across combined sources of data possible. Data may be stored in a completely denormalized ('flat') manner, such as CSV (Comma Separated Values) text files or Microsoft Excel spreadsheets, but also in markup language formats such as XML (Extensible Markup Language) and JSON (JavaScript Object Notation). Conversions between these formats tend to be fairly well-implemented and straightforward, assuming proper integrity of the data. Today, XML and JSON are used in so-called document-oriented databases, where a predefined data model is not required. But larger datasets tend to still be stored in relational databases, where data are stored in tables linked to each other with key fields (Fig. 1). Duplicate pieces of information are stored only once, which reduces the storage capacity needed, but most importantly improves search efficiency and redundancy. In Figure 1, in the data model of the Royal Botanic Gardens, Kew (RBGK), a collector name would be stored only once, whereas it will be repeated for each specimen collected by that person in the data model of the Muséum National d'Histoire Naturelle (MNHN). However, conversion between different relational models is often more difficult and the frequent usage of integer key fields, which are only unique to a specific table, may jeopardize combined queries (Parent & Spaccapietra 2000).



A) MNHN, Paris

RECOLTES			
NUMEROID	*INI_RECOL*	*NOM_P_REC*	*AUT_RECOL*
8959	W.E.	Marriott	Munro H.K.
66328	A.J.B.	Chevalier	
305941	H.S.	MacKee	
265417	J.	Delaunay	
1554483	A.	Krapovickas	Cristobal, C.L.|Quarin, C.
453897	H. de	Poli	
483992	H.	Humbert	
527700	L.	Rotereau	
768663		Feldmann Jean	

Glossary	
RECOLTES	Table listing all collection events
NUMEROID	Primary key of the RECOLTES table
INI_RECOL	Initials of the (first) collector's name.
NOM_P_REC	Last name of the first collector (one word)
AUT_RECOL	Other collector names
collection_item	Table listing all collection events
collecting_team	Table listing all collectors for each collection event
collectors	Table listing all individual collectors
collection_event_id	Primary key of the collection_item table.
collector_id	Primary key of the collectors table
collector_first_name	Full first name of the collector
collector_initials	Initials of the collector
collector_last_name	Full last name of the collector

B) RBG, Kew

collection_item
collection_event_id
8959
66328
305941
265417
1554483
453897
483992
527700
768663

collecting_team	
collection_event_id	*collector_id*
8959	1
8959	2
66328	3
305941	4
265417	5
1554483	6
1554483	7
1554483	8
453897	9
483992	10
527700	11
768663	12

collectors			
collector_id	*collector_first_name*	*collector_initials*	*collector_last_name*
1		W.E.	Marriott
2		H.K.	Munro
3		A.J.B.	Chevalier
4		H.S.	MacKee
5		J.	Delaunay
6		A.	Krapovickas
7		C.L.	Cristobal
8		C.	Quarin
9		H.	de Poli
10		H.	Humbert
11		L.	Rotereau
12	Jean		Feldmann

Figure 1: A comparison of the data model for herbarium specimen collector information in the CMS's of the MNHN in Paris and RBGK. Sample data were taken from the MNHN and for the sake of comparison mapped to how they would be imported into Kew's data model. The Kew collector_id's are examples for this figure only. Note the different methods of specifying multiple collectors for a single collection event and the different approach to storing surnames consisting of multiple words.

Data considered to be a single concept in one database may be stored under different fields in other databases. There are probably many reasons for this, but one of the causes is the use of different schemas and standards for data entry. Different data providers have different interpretations of what certain data fields mean. Generic terms such as 'location', 'date', 'name' and 'record number' may be interpreted and used differently by different people and in system documentation. Such mismatching semantics may be mapped to a single, harmonized standard, but this is not always straightforward if the frequency of mismatches is high. For instance, mismatched fields may also be conditional on other data values or require concatenation of multiple elements. And while concatenation is often a simple procedure, parsing text consistently poses greater difficulty. Neither is it conducive to error-free automated exchange of data. In Figure 1, concatenating all collector names for one specimen in RBGK's data schema is straightforward, but splitting up the AUT_RECOL field from MNHN to fit RBGK's schema would be difficult.

Finally, data may be stored both as the original text transcribed verbatim from the non-digital source and as interpreted data, with a limited number of possibilities. Some data models include fields for both, such as the numerous 'verbatim' fields in DwC. However, this may make it more difficult to map into other data models and cause information of *de facto* identical nature to be stored in different field as described previously. Data fields may follow a controlled vocabulary derived, or inspired by, international standard organizations, such as International

Organization for Standardization (ISO) or Biodiversity Information Standards (TDWG), or follow a local standard. CMS's usually enforce links to standard values in the database, but they often allow the creation of new additions to such tables. In some cases, this leads to the introduction of duplications of what ought to be identical values (e.g. "M", "Male" and "masculin" in the gender field). Ideally, data fields should be restricted to values that are possible for those fields, such as whole numbers, decimals, dates and Booleans.

For efficient and effective querying and processing, data harmonization and standards are preferred. This also avoids false positives and negatives. For example, it would be much easier to find all specimens collected by Charles Darwin if his name were formatted in an identical manner for each record. Modern search technology will have no problem finding specimens collected by "C. Darwin", "Darwin", "Darwin, C." and "Charles Robert Darwin, M.A." as well, but this approach may be jeopardized if collections by "E. Darwin" and "F. Darwin" are present too. A lack of standards for formatting names and controlled vocabularies also increases the risk of spelling, typographical errors and accidental character additions, such as spaces, tabs and line breaks.

An alternative to a controlled vocabulary or linking to internally curated reference lists is using external sources with persistent identifiers (Stork et al. 2018). For findability, it does not matter how Charles Darwin's name is processed in the data model if his collector record points to a biographic entry in a repository keeping trusted up-to-date biographic information available through a unique and persistent identifier. Such an approach also facilitates changes in this information, for instance adjustments of birth date or known itineraries of collectors - or name changes of localities or countries in the case of location data. A downside to using identifiers is that it is less evident to obtain an overview of the entire data structure. For example, determining what data are available and in which format. A solution to this problem is local indexing of the external information, but that has synchronization consequences and the amount of data which needs to be stored increases manifold.

1.3. Interoperability of Collection Management Systems

Under the planned DiSSCo infrastructure, the institutional CMS will feed data into a central architecture where it can be consulted for a multitude of purposes. However, to achieve this, there are also cases where systems will also exchange data between each other. As we will describe below, these situations exist were institutions have multiple systems, where data has to be migrated between systems and where we can improve data quality by connecting data.

If data interoperability issues exist within a single CMS, they may cause problems for the people working with this system and anyone who will use data that flows out of it. Certain sorts of queries cannot easily be performed, or they come with considerable errors of commission and omission. Data on specimens will require cleaning each time they are used (Wittenburg 2018). For many collections, these issues are historic in origin and their elimination is an extended work in progress. Additionally, ingrained working practices and unresolved data quality problems can demotivate staff who would play a role in resolving these issues. Data can be cleaned, but the problem will endure if the practice of how data are added is not changed.

Interoperability of CMS's with other systems is often more problematic. This is in no small part because, historically, such functionality has only rarely been needed. The recent rise in opening up collections to the public by publishing them to the internet is changing this. As exemplified in the FAIR Data Principles, a fundamental role of a scientific collection is ensuring that the specimens and their associated data are Findable, Accessible, Interoperable and Reusable (Wilkinson et al. 2016). Data aggregators such as GBIF have been created to mobilise the world's biodiversity data, allowing researchers to search data from different publishers. These aggregators also tend to enforce or at least recommend some form of data standardization such as the use of Darwin Core (Wieczorek et al. 2012), improving interoperability. This kind of encouragement is often limited to a few fields or not binding. This keeps the threshold for publishing data as low as possible. As a consequence, multiple interoperability issues still remain.

At the same time, before publication, data are often exported from the CMS and stored in a separate, intermediary database. This can cause synchronization issues, in particular if these published data are subsequently harvested by an aggregator, who then stores them in a third place. Published data are often converted during or after exportation from the CMS. Such conversions and caching increase the number of versions of data of a single specimen in circulation and not necessarily synchronised, and can lead also to information loss.

Aggregating data from different sources also creates opportunities. Specimens often have links to other types of information, including biographical details about their collector, genetic sequences, taxonomic names and geography. If these sorts of data are all sufficiently interoperable, then specimens can be enriched through these links. For example, if a specimen lacks information on the country of collection, but a collector's biography locates them in one particular country during that time, then this information can be added to the details of the specimen. Likewise, by linking the specimen's data to external sources of information, the details attributed to the specimen can be cross-referenced and fact checked (cf. user story 17, table 1). This might be as simple as checking whether the collection date and locality are consistent. A specimen can only be collected by a collector who is alive and a species is usually only collected if it is present in that country. In the latter case there are some exceptions, such as when specimens are created from living collections. However, even if this is the case, these exceptions should be explained and the explanation should be documented in the data on the specimen.

Table 1: Textual description of user stories compiled in ICEDIG WP6 (Unpublished Data). The story IDs are taken from a draft document and may be subject to change.

Story ID	**Story description**
17	A **curator wants to** cross-check data between specimens collected by the same collector on the same day, **so that** they can confirm that all specimens have similar geographical coordinates, or correct coordinates where necessary, **for which they need** to select all DiSSCo records by collector and date.
13	A **curator wants to** add annotated information from a Unified Curation and Annotation System (UCAS) to their Collection Management System (CMS), **so that** they can update information on their specimens in their CMS, **for which they need** interoperability between their CMS and UCAS.
14	A **curator wants to** curate a digital specimen (as it enters the DiSSCo data infrastructure), **so that** their CMS has curated specimens, **for which they need** direct access to their digital specimens from the DiSSCo infrastructure.

These enrichment and validation functions are to be implemented as part of the DiSSCo infrastructure. Yet, this also creates a problem. Data that are validated and/or enriched exist in a modified form at the repository where they were published, outside the CMS. Most CMS's lack a straightforward workflow to feed these modified data back into the system's own database. A streamlined import scheme is critical for these sort of data improving services (cf. user story 13). In ICEDIG Task 5.2, a data exchange standard has been proposed to harmonize data generated by transcription or annotation platforms (MS28, Le Bras, Chagnoux and Dillen 2019). However, this does not address the considerable diversity in import templates and methods that exist among the different CMS's in use. Some CMS's already address this issue by being web-based or by implementing an API (Application Programming Interface) to the database, which can render direct incoming and outgoing data flow requests to the database a possibility. Ideally, CMS's would keep their import and export templates as similar as possible. It is likely that DiSSCo will recommend a general data standard for such templates, greatly facilitating this sort of data turnover.

Other than the technical obstacles, collection managers will want to maintain control over what is imported into their CMS, particularly if no versioning is available (cf. user story 14). Multiple annotations of a single specimen may not be supported, or proposed annotations may conflict with one another. This necessitates some form of quality control and approval procedure on incoming data flows, which likely require human input and hence potentially a large overhead in time. Versioning, such as can be found in Wikidata, or annotation functionalities, such as those present in the Estonian cloud-based CMS PlutoF (Abarenkov et al. 2010), may mitigate this problem. Such functionalities will also increase the number of data versions for a single specimen, possibly complicating its scientific use. In particular, when using large datasets consisting of specimens aggregated from many different sources, proper identification of the specific version of every specimen used will be critical and may render data preparation more tedious and hamper replicability of the analysis performed. On the other hand, without proper versioning, reproducibility and open data will never be compatible.

We must also consider the issue of interoperability from the perspective of data migration. Given the short lifespan of software systems, institutions should expect to migrate their data on a regular basis and the lack of standards compliance is a major contributor to so called vendor lock-in. At all stages of the software lifecycle, data managers should balance local requirements with the need for broader interoperability at that time and in the future.

1.4. Data Standards

Biodiversity Information Standards, otherwise known as TDWG, is a non-profit association responsible for the development and maintenance of standards in the field of biodiversity informatics, such as Darwin Core (DwC) and Access to Biological Collection Data (ABCD). ABCD is a highly structured and comprehensive data standard designed for the exchange of primary biodiversity data (biological collection units, including living and preserved specimens, along with field observations that did not produce voucher specimens). It is intended to support the exchange and integration of detailed primary collection and observation data. Extensions of ABCD exist to allow the standard to cover other types of data as well, such as ABCDEFG for the geosciences and ABCDDNA for DNA samples. The standard is supported by GBIF and the BioCASe (Biological Collection Access Service for Europe) network.

Darwin Core is a data standard built upon Dublin Core, which is a standard for digital resources in general. DwC encompasses a glossary of different terms to describe certain properties and, while these terms are listed under certain categories, the standard is intended to be used with little to no underlying data structure (i.e. “flat”). Extensions exist to support certain one-to-many relations, such as multiple identifications or multiple images for a single observation. These allow so-called star schemas to be constructed, where a “core” table links through its primary (hence unique) keys to one or more other tables (Wieczorek et al., 2012). However, relational structures beyond this initial level are not supported. Darwin Core can also support collection event data (e.g. vegetation surveys) and taxonomic checklists (https://github.com/gbif/ipt/wiki/howToPublish). It is less comprehensive than ABCD, but its flexibility makes it easy to use as the basis for a wide variety of data types. Darwin Core is also supported by GBIF and BioCASe. Data retrieved from GBIF are delivered in the form of data structures known as Darwin Core Archives (DwC-A). DwC-A datasets include in addition to the raw data also metadata files, describing the terms used and general information concerning the dataset (e.g. description of the protocol used or the institution providing the data). Darwin Core suggests controlled vocabularies for certain terms, but these do not need to be followed. GBIF, for example only enforces a small number of rules on a limited set of terms.

2. Assessment of collections and their management

2.1. The DiSSCo Survey of CMS Institutional Collection Management System choice

In 2017, all partners of the DiSSCo Consortium were questioned concerning their use of a CMS. These data have not been published yet, but a categorized list of the CMS’s used by each institution can be found in Appendix A. Of 85 respondents, three do not hold any collections and ten either do not have a CMS or did not understand the question. Of those with a CMS, 17 had more than one system in use, presumably for different collections.

About a third of the institutions use in-house solutions, many of which are based on generic database and spreadsheet systems, such as Filemaker (FileMaker Inc.), Excel (Microsoft) and Access (Microsoft) (Table 2). Where respondents use several different CMS solutions, this seems to be for the needs of the different collections they are curating. For example, seed banks and herbaria are often separated, even though they must have data in common, such as taxonomy, collector and geographic data.

Table 2: Classification of CMS's noted in the survey. The various descriptions used for in-house developed systems were grouped together as "Other in-house". Systems which were not developed in-house but were only mentioned a single time were grouped together as "Other not in-house".

CMS	**n**	**CMS**	**n**	**CMS**	**n**
MS Access	10	Adlib	4	CB 3.0	2
Filemaker	8	JACQ	3	DaRWIN	2
SPECIFY	8	PlutoF	3	Jacim	2
MS Excel	7	unknown	3	SARV	2
none	7	ActiMuseo	2	Other not in-house	17
Kotka	5	BgBASE	2	Other in-house	15

Pooling all the in-house solutions into one category, this questionnaire revealed that there are 32 different systems in use among the European institutions surveyed. Of the purpose built systems there was not one clear preference. Although SPECIFY had the most users, national systems featured prominently, such as the Finnish Kotka, the Estonian PlutoF/SARV, the French JACIM interface, Austro-German JACQ and the Belgian DaRWIN.

In summary, institutions in Europe have not converged around a single or few CMS solutions and a large proportion use in-house developed systems. This means that a variety of data models and standards are in use and this is likely to be a barrier to direct interoperability between institutions. Furthermore, for the 17 institution with more than one CMS it is likely that within-institution interoperability is also an issue.

2.2. The Naturalis survey of CMS choices

For their zoological and geological collections, Naturalis Biodiversity Center in Leiden makes use of a customized version of Atlantis, which is developed by a Dutch software company. Because of general dissatisfaction with this system, they sent out a survey on CMS's in late 2017 (RMNH 2018). 108 individuals responded, from 82 different institutes worldwide, though 80% came from Europe or the USA. Eleven were from Naturalis itself. 75% were everyday users, the others were upper management or IT.

The majority of respondents to this questionnaire worked in botany and/or zoology. Half were satisfied with their CMS, a fifth were not. Furthermore, a quarter were planning a change of their CMS. Respondents had notably lower approval ratings for (in-house) custom-made systems. Commercial solutions or systems designed with natural history collections in mind scored quite well. These include Arctos, Axiell EMu, BRAHMS and SPECIFY. Naturalis' own system scored poorly, which was not unexpected as this issue was what sparked the survey in the first place. However, the average CMS satisfaction score for custom-made systems, such as those based on Excel, Access or Filemaker (ca. 26% of the respondents), was not much higher.

Respondents were also invited to describe their CMS' weaknesses and strengths. When Naturalis studied the results (89 and 95% of respondents respectively), they categorized the responses in five categories. Technical application management concerns issues related to the data model, the underlying code and the price of development or implementation. Collection management tools combines responses related to import/export as well as querying. The other three categories are user friendliness, (community) support and flexibility to customize the system.

More than half of the respondents noted technical application management as a strength, but a third also saw it as a weakness of their CMS, in particular users of Access. Collection management tools were often seen as a strength by users of CMS's specialized for natural heritage collections. Conversely, they were more frequently seen as a weakness for generic or custom CMS's, including Naturalis' own system. User friendliness was most often reported as a weakness, in particular for specialized and for customized CMS's. It can be considered the chief strength of generic systems, such as Filemaker and Access, and this trend was also seen in the survey responses. Community support was only mentioned by users of Arctos, EMu and SPECIFY. Generic and customized CMS's scored well on flexibility.

All of these categories are aggregations of potential issues with a CMS. Not every respondent's assessment of strength or weakness should be weighed equally either. For instance, respondents who described both the strengths and weaknesses of a certain category are likely to have a different appreciation of this category in their CMS than those who only report it as a weakness. Still, a general pattern emerges where generic systems are more user-friendly, both in everyday use and in ease of learning how to use it. These systems also offer more flexibility in setting up customized templates or interfaces. More specialized systems seem to provide better tools, better support and often come off better from a technical point of view. Furthermore, technical application management comes up most in the responses either mentioned as strength or weakness.

The results also showed that more than half of the respondents had been working with a system that had been in use at their institution for more than 10 years. More than a quarter of respondents indicated there were plans to switch to another system at their institution. For the generic systems, typically half of the respondents using them indicated this was the case. This was less the case for the customized systems (38%) and even less for the specialized CMS's (35% or less).

2.3. CMS assessment interviews

In-depth interviews were performed with data managers of the seven ICEDIG partners that hold natural history collections. Questions concerned internal standardization of data and field linking, as well as the procedures and problems for (batch) import, export and publication of data held by the system. A full list of the questions asked and the clarifications added can be found in Appendix B. We preferred to conduct these interviews in person or through videoconferencing, to achieve as much interaction and scope for clarification as possible for this complicated subject. However, this was not always feasible. For five out of the seven institutions, real-time interviews could be held. In the follow-up to the interviews, notes were restructured and clarified where appropriate, in some cases with the aid of additional documentation. The contact person for each institution can be found in Table 3. Summaries of interview responses for each institution can be found in Appendix C. Below, we discuss some of these responses, in particular issues raised by the personnel themselves or common to more than one of the institutions.

Table 3: Contact people for each institution. People with an asterisk (*) where interviewed in person or through videoconferencing. The institution IDs will be used further down the text.

Contact person	**Role**	**Institution**	**Institution ID**
Henry Engledow*	Scientific Manager Collection Databases	**Meise** Botanic Garden	APM
Frédérique Bakker	Collection Manager Hymenoptera	**Naturalis** Biodiversity Center	RMNH
Matt Woodburn*	Science Data Architect	Natural History Museum, **London**	NHM
Sarah Phillips	Research Leader, Digital Collections	Royal Botanic Gardens, **Kew**	RBGK
Urmas Kõljalg*	Director	Natural History Museum, University of **Tartu**	UT
Anniina Kuusijarvi* Zhengzhe (John) Wu*	IT Designer IT Specialist in Digitization	Finnish Museum of Natural History Luomus, University of **Helsinki**	LUOMUS
Simon Chagnoux*	Scientific Applications Manager	National Museum of Natural History, **Paris**	MNHN

2.3.1. CMS interoperability and standardization

Indicative numbers for the collection composition of the different institutions can be found in Table 4. Of the seven institutions, most used multiple databases. Library catalogues are generally independent systems for historical and organizational reasons. Beyond that, different types of specimens go into different systems: for example, the University of Tartu (UT), has developed a separate system for geological specimens. Other common separations are between zoology and botany, between living and preserved specimens, and between the method of specimen preservation, for example dried versus liquid preserved. Only at the Natural History Museum, London (NHM) are the majority of specimens, including books, kept in a single system: EMu, a generic commercial CMS. Even so, different databases may be accessible through a single interface, such as at the Muséum national d'Histoire naturelle (MNHN), which imposes a degree of interoperability.

Table 4. Collection info for each institution. Source: DiSSCo survey (cf. section 2.1.). Kew info from interview and website. Black means the institution does not hold such specimens.

ID	Botany	Zoology	Paleontology	Mineralogy	Mycology	Entomology	Microbiology	Tissue/ DNA	Living	Seed
LUOMUS	11-20%	1-10%	<1%	<1%	1-10%	61-70%	■	<1%	<1%	<1%
MNHN	11-20%	1-10%	11-20%	1-10%	1-10%	41-50%	<1%	<1%	<1%	<1%
UT	21-30%	11-20%	1-10%	<1%	1-10%	41-50%	<1%	<1%	1-10%	<1%
APM	>90%	■	■	■	1-10%	■	■	<1%	1-10%	<1%
NHM	11-20%	31-40%	11-20%	<1%	<1%	31-40%	1-10%	<1%	■	<1%
RMNH	11-20%	21-30%	21-30%	1-10%	<1%	21-30%	■	<1%	■	<1%
RBGK	71-80%	■	■	■	11-20%	■	■	<1%	1-10%	<1%

Another reason for using multiple systems is a historical separation. This can parallel a taxonomic separation, such as between zoology and botany as at Naturalis Biodiversity Center (RMNH). In some cases, collections may have consolidated under a single institution, but still be managed independently in different CMS's until the problem of their digital integration has been resolved. Further examples of this are the mycological collections at the Royal Botanic Garden, Kew (RBGK) or certain older databases at the Finnish Museum of Natural History (LUOMUS), whose integration into the Kotka CMS has not yet been achieved. It should also be noted that integrating all data into a single CMS is not necessarily the optimal solution to interoperability problems. For instance, concerning Kotka, the data manager stated that different institutions making use of the CMS cause some interoperability problems when they "*develop a habit of submitting their own data in a slightly different manner*". On the other hand, making links between different databases is often not straightforward and may entail considerable manual work.

The software used for media and data management differ considerably between institutions. While, in principle, they could still be interoperable, in practice interoperability does not appear to have been a priority in their design. Some management systems are chiefly developed at the institutions themselves, others are commercially available software. Upon a closer look at data management, there are also differences in underlying data models of the CMS's, with various degrees of standardization and quality of such standardization, but also capacity for verbatim information and complementary information. The potential number of data fields linked to a specimen may range from the hundreds (e.g. MNHN or LUOMUS) to the thousands (e.g. APM). It is also common for data model incompatibility to have been addressed through *ad hoc* solutions, complicating understanding of the effective current data structure.

Most systems include a hierarchical taxonomy to link specimens to taxonomic names (i.e. a taxonomic backbone). A specimen is often linked to multiple scientific names through multiple determinations of a single specimen. These taxonomies are rarely curated and kept up-to-date however. For the locations and collectors, specimens are also link to tables of gazetteers and person names, but duplicates and unlinked specimens frequently occur.

2.3.2. CMS import

A common issue with imports is that the new data should be linked to other data in the system, such as IDs for taxonomic names, collectors and localities. As these internal entries are of variable origin and quality, there is no easy, general method to facilitate creating these links. The availability of web services providing such IDs and proper integration of them into the system would greatly improve importation routines – in particular of newly gathered data or data processed from other CMS's. Some web services have been catalogued at www.biodiversitycatalogue.org, but greater adoption is also needed of existing services.

Various templates for import are in use for batch import of data, making use of proprietary software such as Microsoft Access templates or Excel spreadsheets, and workflows may involve multiple conversion steps. Open formats such as XML or CSV are also often used. Markup language has a critical advantage when dealing with multi-value fields, but is more complex to understand and use by the average user. However, most systems also offer an interface for adding new records. Linked or validated fields may make suggestions for values (e.g. scientific or collector names) or enforce them.

2.3.3. CMS export and publication

Workflows for export and publication range from those that are completely automated and efficient, to those that are cumbersome and require human intervention. There is also quite some variation in the number of intermediary states the data can have between its initial format within the CMS and its final format as stored and indexed by the public repository. If these workflows take a long time, a lack of versioning may cause problems. Updates of public data portals happen at a wide variety of intervals. All seven institutions we studied publish to GBIF. All institutions also either have an online portal of their own or have this functionality built into the CMS itself. Synchronization between GBIF, the institutional portal and CMS varies considerably.

2.4. Recommendations for CMS management

2.4.1. CMS harmonization

It seems that many organizations have chosen to create their own CMS, either building upon generic database software such as Filemaker or adapting existing CMS's to their needs. In the short term, this has many advantages for the organization. Systems can be created on an *ad hoc* basis initially, fit to the current users' needs most directly and easy to use. Furthermore, local curatorial practices can be supported by creating bespoke platforms, rather than confronting and changing ingrained practises. However, such systems are generally built by a small team or a single individual upon which the whole integrity of the CMS relies. This makes the system highly vulnerable to a change of staff or a lack of suitably skilled people. As indicated by the Naturalis survey and the interview responses from NHM and RMNH, even CMS's developed by larger organizations and private companies are vulnerable to these weaknesses, as they often offer tailor-made modifications to their core system for different collections. As the core system gets developed further, keeping those modified versions up-to-date becomes increasingly strenuous, buggy and expensive.

Another frequently seen problem is that IT is a fast changing field, with operating systems, programming languages and programming frameworks ageing rapidly. An individual cannot be expected to keep up-to-date with all the potential technological changes and different skill sets needed for the development of software as compared to software maintenance. The lifespan of commercial software is generally estimated to be between 6 to 8 years (https://mitosystems.com/software-evolution/), while most users indicated their institution had been working with their system for longer than that. Compounding the problem, interoperability is often not a priority until it is too late. Small issues can snowball out of control over the course of a few years and data migrations become stressful and inefficient if they have not been anticipated in the design. This includes maintaining documentation of changes and data encoding, but also keeping up to date with software developments so as to avoid relying on unsupported software (or hardware) applications.

The establishment of CMS consortia, be it international ones such as Arctos or SPECIFY, or regional ones like JACQ, PlutoF or Kotka, goes some way towards mitigating these problems. They allow a relatively small group of skilled developers to maintain a system used by multiple institutions. When issues arise, local data managers may request feedback from other partners or rely on previous experiences of other institutions - including for data migrations. This also allows multiple institutions to share the considerable cost of paying a dedicated and sufficiently skilled development and management team.

We recommend that, before institutions embark on creating their own CMS, they should consider if the advantages outweigh the costs. They should also consider the long-term sustainability of such a project, provided that such a system may need replacement in a decade. A sufficient number of different systems with different development models are already available for institutions to choose from. The complexity of information technology and the challenges of integration with the web services and infrastructures already push collection managers towards shared systems. Ultimately, DiSSCo will require interoperability of CMS's with the DiSSCo core infrastructure, which will be easier for a CMS developed and supported by a consortium. However, the problems pertaining to bespoke in-house developed CMS's are often historical in nature. **Hence, we also recommend that DiSSCo both encourages and facilitates data migrations into CMS's with better support, with the aim of reducing the overall number of different types of CMS in operation.**

2.4.2. Linking to taxonomic backbones and other resources

Controlled vocabularies are used to some extent in most CMS's. Good examples are country codes (ISO 3166), Index Herbariorum codes and DwC terms. Vocabularies do have the downside that not everyone may agree with them or they can be inflexible. For instance, there is no agreed syntax for a person's name, and different languages and cultures have different traditions on this. A strict syntax for such types of data can also clash with certain data models, such as when the data are stored in different places or have a different effect on other data fields, depending on how they are formatted. An alternative is to use web resources, where through persistent links certain instances of information can be simultaneously provided to different CMS's in a uniform manner. The GBIF Backbone Taxonomy is an example, providing unique identifiers for taxonomic names. Any CMS linking its specimens to such taxon name identifiers will in principle render its data interoperable with the data in other systems which have gone through the same procedure. A similar approach can be taken for the names of people (collectors, identifiers, authors…), which can be found in multiple online databases (e.g. Harvard University Index of Botanists, Virtual International Authority File, ORCID, Wikidata). For some entities, the interpretation of what is meant by an entity is fairly clear, such as people. However, we need to be mindful that for instance taxonomic names, habitats and morphological traits can be interpreted differently.

Another problem with web resource links is the issue of keeping the information contained in them up-to-date, but also how to tackle lower information quality than the locally available data, confidential information or, in the case of open systems, vandalism. Resources such as the Harvard Index of Botanists are rarely updated and will hence be systematically out of date. On the other side of the spectrum, an open resource such as Wikidata can immediately be updated by any contributor, but this risks vandalism and errors of inexperience. The GBIF backbone is updated monthly, but individual taxonomic experts/institutions do not always agree with some of its classifications and may continue to maintain their own system. Furthermore, none of these aggregate systems contain all published scientific names and their variants. A solution would be to use a hybrid approach, whereby a variety of identifiers are used, but this complicates data management and versioning considerably.

Concerning taxonomy, in 2015 the global biodiversity information initiatives Biodiversity Heritage Library (BHL), Barcode of Life Data systems (BoLD), Catalogue of Life (CoL), Encyclopedia of Life (EoL), and the GBIF Secretariat took the first step to work on the idea of building a single shared authoritative taxonomic backbone that can be used to order and connect biodiversity data across various domains. Each of these initiatives focuses on the delivery of a consistent, normalised view of available data for a particular class of biodiversity information (GBIF - occurrence records, CoL - species concepts and names, EoL - species-level information and species traits resources, BHL - biodiversity publications, BoLD - barcode sequence records).

As a fundamental axis for organising their data, these global biodiversity information initiatives depend on the use of scientific names and the associated species concepts. Following an initial meeting in 2015, the Catalogue of Life Plus initiative works towards a shared, extended catalogue and completing the reviewed name coverage. Creating an open, shared and sustainable consensus taxonomy to serve the proper linking of data in the global biodiversity information initiatives is the ultimate goal the initiative aims for. The Catalogue of Life and Species2000 governance have adopted and endorsed the initiative. Also, GBIF has formally embraced and incorporated it in their strategic and implementation plan for the period 2017-2021.

A similar approach may be taken for person names, habitat categories or place names, but it's likely that some differences in interpretations of these concepts may never be resolved, in which case different external resources will continue to exist side by side. But at the very least the problem of harmonization will be narrowed down to the level of those resources rather than every individual CMS.

Hence, we recommend that institutions implement in their CMS's links of their conceptual data, such as instances of people, habitats and species, to external resources through unique identifiers. If these identifiers are used across multiple CMS's, data from these systems can be more easily harmonized. Various services can be found at https://www.biodiversitycatalogue.org/services.

2.4.3. Support for annotations and versioning of specimen records

All systems include a timestamp of when the record was last edited. This does not always provide information on what has been changed. A few systems have support for different annotations of the same (specimen) record or a version history of it. This can provide a degree of data provenance, but we often encounter multiple, potentially conflicting, versions of specimen data that are hard to reconcile. Most commonly for biological specimens, multiple identifications of the species name are possible for a single specimen. Another prominent case example is the attachment of verbatim transcription and standardized data to a record. This may include older record entries from previous database systems or field notes, but also automatically generated transcriptions from OCR (Optical Character Recognition). Such automatically generated data will be generated on a wider and more frequent basis as machine learning technologies find wider application on digital specimen images and other data. Increasingly, these automated methods are being used to generate data from specimens, for example see ICEDIG deliverable 4.1 (Owen et al. 2019). In the time frame of the DiSSCo infrastructure establishment, high volume image and data flows are anticipated. Hence, the function of a CMS to store different annotations or versions of the same data concept will become even more important, ideally with a method to associate such data with the metadata of their origin as well. There is also a need to define more carefully what is meant by an annotation. Annotations could be made by humans and machines, they can refer to the whole specimen record or part of it, and in some cases it might be more appropriate to create a new version of the record, rather than an annotation.

A version history facilitates reproducibility and repeatability of data analyses. It becomes possible to identify which data were used and how they differ from the most recent version. In general, most CMS's do not store version histories. The additional complexity and volume of data have been prohibitive. Most institutions consider their CMS to hold the authoritative version of the data on a specimen. However, the increasing number of sources of data are beginning to make this position look increasingly untenable. Different types and versions of data are coming from different transcription systems, digital literature and automated analysis. Each has its own record of metadata, quality control and provenance. In the past this was resolved by merging datasets, often manually, and often at the loss of provenance. However, the sheer volume of specimens and data is making this approach unsustainable.

We therefore recommend that the versioning of data in CMS's is addressed urgently in parallel with the storage of annotations. This will require the redesign or replacement of many collection management systems.

2.4.4. Import and export routines

Ideally it should be possible for every CMS to export data and re-import these data seamlessly and in a standardised manner. Otherwise, CMS's without such capabilities will become bottlenecks, with the possibility that institutions using them will be prevented from fully contributing and benefiting in a multi-collection digital specimen infrastructure.

In section 3.3. of this report, the occurrence of unpublished data in CMS's was investigated. These data will not be available to be utilized in the DiSSCo infrastructure, despite that some instances may be quite interesting for some use cases (e.g. elaborate descriptions of the preparations done to a certain specimen or links to references in the library database). Often, the problem is related to the data standards used for the data export template, but sometimes the difficulty in constructing the right queries to comprehensively obtain all data and the time it takes to process these queries is a factor as well. **Ideally, CMS's should be designed in such a way that all data related to a certain specimen can be efficiently and effectively queried, as well as updated.**

2.4.5. Conclusion

All of the above recommendations should improve the current state of the art in collections management with Collection Management Systems. Consolidation of CMS's in use into a set of well-supported and continuously developed systems can avoid a lot of double-work and reduce the compatibility efforts needed. Linking to external sources reduces the mapping between different vocabularies needed, as this can all be done at the interaction of the different services themselves rather than on a case by case basis at each CMS. Support for annotations, versioning and provenance preempt some of the looming problems that progress in the biodiversity informatics field will be accompanied by. They are critical for proper scientific use of the data published by collection holders.

3. Assessment of interoperability of published data

3.1. Benchmark dataset

To obtain an overview of the diversity, quality and interoperability of collection data, multiple European institutions holding botanical collections were approached to provide a sample of 200 of their digitally imaged herbarium sheet specimens. Our focus was on herbarium specimens as these are the sort of specimens all collection-holding ICEDIG partners are curating. In addition, herbarium collections have often made most progress in terms of digitization, compared to other kinds of collections. In addition to evaluating interoperability of the data on these specimens, they have also been utilized for other ICEDIG ventures, such as assessment of automated or crowdsourced data transcription (Tasks 4.1 and 4.2), comparison of different citizen science platforms (T5.2) and evaluation of cloud infrastructure for specimen publication and long-term archiving (T6.3.3). A more extensive description of the establishment, processing, publication and applications of this dataset can be found in Dillen et al. (2019) (open access), as well as the ICEDIG deliverables associated with the ICEDIG tasks listed above. In this report, we focus on the aspects of this dataset relevant for the interoperability question. This will be a summary of information that can be found in the data paper mentioned above. We also perform a few analyses of interoperability and attempt to address the relationship of the published data with the CMS's themselves.

In addition to the seven ICEDIG partners already interviewed in section 2 of this report, the Royal Botanic Garden Edinburgh (RBGE) and the Botanical Garden and Botanical Museum, Berlin (BGBM) also provided specimens to this dataset. We proposed a few guidelines for selecting specimens, to ensure an adequate degree of coverage of the institution's collection and acceptable representativity of the dataset in both time, space, taxonomy and language of the label information. Table 5 lists these guidelines and Table 6 lists all contributors and how they interpreted them.

Table 5. The guidelines given to herbaria to select specimens for the test dataset. The goal was not to have a representative sample of all specimens, but to have comparable subsets, which will have labels written in different languages; will be printed or handwritten; will cover a wide range of dates; will be both type specimens and general collections and will provide specimens from different families and different parts of the world.

Number of specimens	**Type status**	**Date of collection**	**Geography**
25	Type	≤ 1970	Any country
25	Type	> 1970	Any country
25	non-Type	≤ 1970	From the country where the herbarium is located
25	non-Type	> 1970	From the country where the herbarium is located

100	non-Type	Any	non-Type specimens from one other country or region of which the herbarium possesses a substantial number of specimens

Table 6: Contributions of 9 different institutes to the dataset. Availability of JPEG and TIFF images is indicated, as well as the source of label data. Most institutes were able to follow the template in table 5. The regions picked for the 100 non-Type specimens are indicated in the last column, as are deviations from the template in Table 5. The DOI of the collections is listed if GBIF was used as a data source. FinBIF is the Finnish Biodiversity Information Facility available at www.species.fi (Schulman et al. 2018). JACQ is a joint specimen data management system of over 30 European and Asian herbaria available at https://herbarium.univie.ac.at/database/ (Rainer & Vitek 2009).

Institution ID	**Data Source**	**Composition**	**Lossy JPEG retrieval method**
APM	10.15468/wrthhx	As Table 3; 100 from AU, CA, NZ, US	Transferred through Local Area Network (LAN).
RBGK	10.15468/ly60bx	As Table 3; 100 from BR	Transferred through Google Drive.
NHM	10.5519/0002965	As Table 3; 100 from AU, CA, NZ, US	Transferred through www.box.com.
BGBM	JACQ	As Table 3; 100 from AU, BR, CN, ID, TZ, US	Transferred through BGBM ownCloud.
RBGE	10.15468/ypoair	As Table 3; 100 from CN	Transferred through Google Drive.
MNHN	10.15468/nc6rxy	50 type, 50 non-Type FR, 100 non-Type not FR	Direct download through provided URIs.
UT	10.15156/bio/587444	100 ≤ 1970, 100 > 1970	Direct download through provided URIs.
RMNH	10.15468/ib5ypt	As Table 3; 100 from ID; no selection on date	Direct download through constructed URIs.
LUOMUS	FinBIF	As Table 3; 14 FI, 36 ET instead of 50 FI; 100 from AU, BR, CN, ID, US	Transferred through digitarium.fi.

From each institution, lossy JPEG images of the specimens were collected. Various methods of image transfer were used, as can be seen in Table 4. Some of these methods were used because non-lossy TIFFs were included as well, which constitute much greater file volumes. As these TIFFs are not relevant for this report, we will not discuss them further.

Published data on these specimens were harvested from GBIF through its API, implemented in the R programming language using *rgbif* (R Core Team 2017, Chamberlain 2017). As not all data were published to GBIF for the BGBM and LUOMUS specimens, we obtained their data in different ways. The BGBM data were provided as a direct DwC-A export from their CMS (JACQ). The LUOMUS data could be requested through the FinBIF API. However, the functionality of exporting FinBIF data into Darwin Core was still limited at the time. Hence, the specimen data were requested in a JSON format corresponding to the FinBIF data model. These data were then mapped to Darwin Core using the R programming language (*jsonlite* package, Ooms 2014). The methodology used for this transformation can be found in a script supplemented to the data paper (Dillen et al. 2019), yet this was in some respects an *ad hoc* operation and may hence not be scalable for transforming data on other specimens.

In addition to the published data, exports representing the internal CMS model as good as possible were requested from the seven ICEDIG partners. This was done in part to obtain a better understanding of the inner workings of these systems and the consequences of differences in this respect, but also to identify information which does not make it to publication, possibly due to shortcomings in the publishing data standards. MNHN, RBGK, LUOMUS and NHM were able to provide such samples, which made clear how diverse CMS data models can be. The limited relationality of the MNHN botanical database (called Sonnerat) rendered an export in multiple spreadsheet tabs fairly straightforward, although related data on links with literature and ethnobotanical information were not exported. The raw exported data provided by NHM only constituted elements in the core catalogue and elements from other modules if they were indexed by the system in the core catalogue already. In addition, nested tables were flattened into multi-value fields. By contrast, LUOMUS had little issue providing a relatively flat export file, where multi-value data elements were parsed as additional columns. The CMS of RBGK is quite relational in nature, but a fairly flat file could be generated just as well by manually joining the results of multiple queries. Here, multi-value data elements were parsed as additional rows.

The difficulties encountered in this process highlights a common problem with many CMS's, where the notion of “all data” concerning the specimen or its collecting event quickly snowballs into an unwieldy and very time-consuming query, both in properly constructing the query and running it. For instance, the collector information can be stored in a separate table, which itself links to various biographical elements as well as other specimens or data types associated with that person.

3.2. Interoperability analysis of the dataset

All collected data were joined into a single DwC table. An assessment was performed concerning the semantic interoperability of this aggregated dataset. Certain issues already came to light while data were prepared for rendering the graphs reflecting the dataset’s coverage or for conversion into single JSON-LD files.

An issue already presented itself when the data were harvested from GBIF. If we harvest data from GBIF using the specimen's persistent identifier, there is no single field in GBIF data to find them for every institution. The `dwc:occurrenceID` values on GBIF had different meanings depending on the institution. These values were mostly the persistent identifiers (URIs) for the physical specimens (i.e. the occurrences), as agreed within CETAF by all contributors (except UT, but they employ a similar approach). For example, at APM, the fixed part of the persistent identifier is "www.botanicalcollections.be/specimen/" and the specifying part, the specimen's barcode is appended to it. As can be seen in table 7, UT only listed the barcode and not the whole persistent identifier, which was stored under `dc:references`. The NHM did not include their CETAF persistent identifier anywhere and listed the GUIDs (Globally Unique Identifiers) they use to uniquely identify their specimens, which are indeed the specifying part of their persistent identifier as opposed to the barcode. Unlike barcodes, GUIDs are not physically present on the specimen and no registries is needed to ensure their uniqueness (Nelson et al. 2018). They consist of a sequence of 32 characters, which can be virtually guaranteed to be unique as the probability of generating a duplicate is very close to zero.

Table 7: Examples of occurrenceIDs in use by each of the partners and their actual meaning.

	occurrenceID example	**interpretation**
BGBM	http://herbarium.bgbm.org/object/B100000389	Persistent URI
APM	http://www.botanicalcollections.be/specimen/BR0000005701353	Persistent URI
RBGE	http://data.rbge.org.uk/herb/E00064130	Persistent URI
RBGK	http://specimens.kew.org/herbarium/K000311023	Persistent URI
MNHN	http://coldb.mnhn.fr/catalognumber/mnhn/p/p00543937	Persistent URI
NHM	333757c3-081a-478e-803d-6b2674a42754	GUID
UT	TU256926	Barcode
RMNH	http://data.biodiversitydata.nl/naturalis/specimen/L%20%200044849	Persistent URI
LUOMUS	http://id.luomus.fi/EIG.2383	Persistent URI

In DwC, `dwc:occurrenceID` is described as "*An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique* *identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique*". Any of the provided IDs could be considered to fit this definition, although the barcodes used by TU are possibly not as unique as the others. Yet, a problem occurs when a system of identifiers is agreed upon, such as the CETAF identifiers, but then not consistently used. If someone were to use a list of barcodes or persistent URIs to extract specimen data from GBIF, they would have to perform *ad hoc* fixes to find all specimens unless a common meaning is attached to a certain DwC term.

A further complication in this context were mismatches between the barcode as can be found on the physical specimen, the `dwc:catalogNumber` value and the barcode number as used to identify the images in form of the filename. In Table 6, a few examples can be noted, such as replacing underscores (_) with spaces, removing spaces, and a change in the case of letters. There was also an inconsistent use of file extensions (.jpeg vs .jpg). The use of "EXTU" by UT was done to indicate multiple images for a single specimen and is not a problem in itself. However, there is currently no standard approach to indicate such instances. At APM, RBGK and NHM, for instance, the convention is to append _a, _b and so on to any image file beyond the initial scan. But encoding versioning into filenames is not an ideal solution anyway, as this versioning is information about the digital image, which may be subject to change.

Table 8: Barcode templates for the different collections as they are used in the data (`dwc:catalogNumber`) and with the images (as filename or specifying part of the persistent identifier). Numbers in brackets reflect the number of numeric digits the barcode contains. Inconsistencies are marked in red. Different letters (such as for RMNH) reflect different specimen origin and are not a problem in itself.

BGBM	APM	RBGE	RBGK	MNHN	NHM	UT	RMNH	LUOMUS
Data								
B [2] [7]	BR[13]	E[8]	K[9]	P[8]	BM[9]	TU[6]	AMD.[6]	EIG.[3]
	BR[13]V			PC[7]			L [7]	EIG.[4]
							L.[7]	H.[7]
							U [7]	HA.H.[7]
							U.[7]	
							WAG.[7]	
							WAG[7]	
Images								
B_[2]_[7].jpg	BR[13].jpg	E[8].jpg	K[9].jpg	p[8].jpg	BM[9].jpg	TU[6].jpg	AMD.[6].jpg	EIG.[3].jpg
	BR[13]V.jpg			pc[7].jpg	BM[9].jpeg	EXTU[6].jpg	L [7].jpg	EIG.[4].jpg
							L[7].jpg	H.[7].jpg
							L.[7].jpg	HA.H.[7].jpg
							U [7].jpg	
							U[7].jpg	

							U.[7].jpg	
							WAG.[7].jpg	
							WAG[7].jpg	

Not listed in the table and often not immediately apparent, there may be differences between the barcode on the physical specimen and how it is used digitally as an identifier. For example, at APM, initially specimens were barcoded with codes containing only numbers and not any other character. A barcode scanner or OCR algorithm would see only this numeric code, rather than the identifier used as part of the persistent URI or the image file name, which includes an alphabetic prefix and in some cases a prefix. As a result, a digital specimen might need three different data fields to depict (1) the actual barcode printed on a label on the specimen, (2) the catalog number used as part of the persistent identifier and (3) the catalog number used to identify the image in the media management system.

Some other, less critical issues could be identified as well. The DwC terms of `dwc:institutionCode` and `dwc:collectionCode` were not used in the same manner by all institutions (Table 7). This is a consequence of different levels of organization between institutes, as institutions with different types of collections (e.g. botanical, zoological, geological, ethnobotanical…) are more likely to use specific names or identifiers for their botanical collection in `dwc:collectionCode`. But also a lack of a suggested or controlled vocabulary for these terms contributes to the inconsistencies. While this may seem a minor problem, it complicates consistent identification of an institution/collection when aggregating data from different sources. For example, to analyze and process the whole dataset, a new variable had to be created to consistently identify each institutional contributor. Also, as a potential complication, for botanical collections the Index Herbariorum code is commonly used for these DwC terms. If institutions choose to use a different code, such as an institutional abbreviation, instead, this could cause conflicts if these codes correspond to Index Herbariorum codes from other institutions. In this case, this problem could arise for MNHN (IH code: P, PCU, PAT or PC), as the abbreviation of their institution's name (MNHN) is also the Index Herbariorum code for the Museo Nacional de Historia Natural in Havana, Cuba.

Table 9: Side by side comparison of `dwc:institutionCode` and `dwc:collectionCode` as used by the different contributors when publishing to GBIF (or as exported from JACQ). LUOMUS is not included in this table as we did the mapping to DwC for those data ourselves.

	BGBM	APM	RBGE	RBGK	MNHN	NHM	UT	Naturalis
dwc:institutionCode	B	BR	E	K	MNHN	NHMUK	UTE	
dwc:collectionCode	Herbarium Berolinense		E	Herbarium	P|PC	BOT	TU	Botany

Values for `dwc:eventDate` are validated by GBIF and always follow the ISO 8601 standard. As the BGBM specimens were not processed this way, some deviations could occur here. Some of the BGBM specimens had partial dates (e.g. 1890-05 for May 1890) or dates with additional zeros (e.g. 1890-00-00). The former is not supported by GBIF, where 01 gets padded onto the partial date, and the latter is inconsistent with the former as well as not supported by the ISO standard. Some BGBM specimens also lacked padded zeros for specific values, e.g. 1890-5-1 instead of 1890-05-01. For the GBIF data, the added 01 values (e.g. 1890-01-01 for a collection date for which only the year 1890 is known) misrepresent the data, but this could be captured by looking at the `dwc:day`, `dwc:month` and `dwc:year` terms. Nevertheless, this approach still keeps the system from parsing specific date ranges such as 1890-05-03/1890-05-07, which is a syntax supported by the ISO standard. Another technical complication at the GBIF level is that the time is added onto the `dwc:eventDate`, as T00:00:00 or T01:00:00, both without apparent meaning.

Missing or unknown country codes were either provided as empty fields, “none” or “ZZ”. For `dwc:recordedBy` (i.e. the collector name), multiple indications for “collector is not known” were present: s.c., S.C., unknown, unreadable, collector unknown, etc. `dwc:recordNumber` had the same problem.

Finally, the fundamental nature of the occurrence was not always as refined. `dwc:basisOfRecord` was always indicated and the vocabulary followed, as this is mandatory for GBIF publication. The exact syntax differed for BGBM’s JACQ exported data, but this was because GBIF does not follow the TDWG recommended vocabulary (PRESERVED_SPECIMEN rather than PreservedSpecimen). This may be due to technical reasons, as IPT validation does require this vocabulary to be followed. But beyond this, only 3 collections specified that these records were herbarium sheets (all in a different manner: “HerbariumSheet”, “hb” and “herbarium specimen of unspecified type”) in `dwc:preparations` and only 3 specified that the `dc:type` was a PhysicalObject.

3.3. Comparing CMS data with published data

Two specimens from RBGK were compared as to their data which were exported from their CMS, and data which were retrieved from GBIF. While most of the important data are published, certain Darwin Core fields are filled with information not present in the CMS - at least not literally. This includes data such as the `dwc:occurrenceID`, the `dwc:basisOfRecord` and `dc:language`. In addition, a few pieces of information could not be found in the published records. This included a verbatim country field, description of the specimen’s nature (e.g. item = “Sheet” and plant_part = “Leaves, Flowers/Infl.”), whether the specimen was a cultivated one and whether the specimen’s identification was present on the actual sheet. The country data for the examined specimen failed to get published, because an ISO country code was missing and the name used was an old one (Zaire). Also missing was the location data following the World Geographical Scheme for Recording Plant Distributions (WGSRPD), a TDWG standard. Of course, data relevant for the structure of the CMS (like internal keys or user rights settings) and important for daily management, like the physical sublocation within the collection, is not relevant for data publication in a global aggregator. These were not considered here.

While assessing the flattened export prepared by NHM, it was immediately clear a lot of published data are already prepared in the CMS itself as introduced Darwin Core fields (receiving the prefix “Dar”), which retrieve the information they contain from other fields. Again, the bulk of the data make it to publication this way. Some enrichment or validation by GBIF itself complicates the matter, such as matching certain taxonomic fields to the backbone. For one specimen, a different `dwc:genus` was noted than the genus part of the `dwc:scientificName`. The `dwc:countryCode` value was also interpreted from the textual DarCountry field. A lot of curatory data do not make it to publication. These can be taken as not relevant enough or even sensitive (e.g. names of responsible personnel or specific information on the specimen's current location), but informative and categorizing data on the specimen's nature, curation and preparation do not fit those criteria. Also, verbatim (partial) transcriptions are pieces of information which rarely make it to publication.

Attaching extra data to specimens when publishing is a standard procedure, but it does create potential problems. It complicates an eventual re-import of these data into the source system, if needed. The added information will then most likely be dropped. This could cause problems if some of the data were modified after validation or annotation. It may also cause synchronization issues when data are put onto different publishing platforms.

There are various possible reasons as to why these data might not get published. Their effective use may be very rare and hence not noteworthy enough. There may also be doubts about their accuracy or tidiness. They may have been overlooked by those implementing the publishing workflows. Or there may have been problems mapping them to the publishing standard. This is often the case for verbatim data (e.g. country for RBGK, but also verbatim label transcriptions generated through OCR), uncommon yet standardized data (e.g. the WGSRPD code for RBGK) or specific aspects of the sheet itself (e.g. "name not on sheet").

3.4. Recommendations for the development and adoption of standards

3.4.1. Darwin core templates for subclasses of specimens

Darwin Core was developed to be a flexible standard. This makes it easier to describe various kinds of taxon observations, but this flexibility hampers interoperability with similar data. This could be noted for multiple terms in the benchmark dataset. To address this, DwC templates for certain subclasses of observations could be proposed. One specific example can be found in ICEDIG Milestone 28 (Le Bras, Chagnoux and Dillen 2019), where a data exchange standard using the DwC-A model has been formulated specifically for data transcribed from digitized specimen images. This proposed standard includes more strict controlled vocabularies and prescribes a syntax to use (for instance) to format names of specimen collectors and determiners. It also lists which terms to use and which not to use, and what to use them for.

Such a standard constrains the flexibility of DwC to reduce the intermediate conversion steps required between export and import operations. However, one of the key issues identified in this analysis is the poor identifiability of what kind of specimen a certain record represents. Right now, there is no standard way to identify a specimen as being a herbarium sheet, a pinned insect or preserved in a certain liquid. Collections can be divided in various ways, including taxonomically, by their history, by their preservation method and by their storage location. With their response to the interview, the NHM described a list of the sorts of specimens in their system. Excluding the library entries, this list has 110 entries with records distributed among two departments (Geology and Life Sciences) and three descriptive terms: collection item type, item category and preservation method. This list includes separate records for every possible unique combination of those three terms, so vocabularies for individual terms would be more constrained. However, it would still be a challenge to fit all information in just the existing `dc:type`, `dwc:basisOfRecord` and `dwc:preparations` fields.

For instance, a herbarium sheet in DwC could now be identified with “`dc:type:PhysicalObject`”, “`dwc:basisOfRecord:PreservedSpecimen`” and “`dwc:preparations:HerbariumSheet`”, but this would imply other information pertaining to the sheet’s preparations (e.g. partial or whole, any chemical treatments) would have to be put in another field. This could be one of the DwC “Remarks” fields, but there would be confusion as to which of these to use and they would lack a controlled vocabulary. A proper template for herbarium specimens would prescribe what field to use for this information and would attach a controlled vocabulary to at least one other term than just type and `dwc:basisOfRecord`. Similar templates could be devised for other prominent categories of collection specimens.

We recommend the development and further adoption of controlled vocabularies for specimen data and the adoption of more standard formats specifically tailored to the needs of specimen data.

3.4.2. The importance of identifiers

As of April 2019, fourteen European institutions have adopted the CETAF Stable Identifier Framework (Güntsch et al. 2017). This framework provides a URI-based unique stable identifier for digital information representing a physical specimen and a machine readable version of the associated data. Critical to this approach is that institutions take responsibility for keeping identifiers resolvable. Ideally, no information pertaining to the specimen itself is part of the identifier. For example, problems might occur if institutional names or physical locations are encoded in the identifier. Barcodes are often used as the link between the physical specimen, the digital images made of it and the data on them. They are physically attached to the specimen, viewable on the image and encoded in data fields such as `dwc:catalogNumber`. They are a locally unique identifier and therefore also feature prominently in many stable identifiers, such as those part of the CETAF framework (e.g. http://www.botanicalcollections.be/specimen/BR0000005117321). As described during the interoperability analysis (section 3.2), this may cause various sorts of inconsistency problems if barcodes and identifiers derived from them are not consistently formatted.

Institutions should consistently format their local identifiers for all usages. The only instance where this may not be feasible, is the physical code stuck to the specimen. Old barcodes or codes stemming from different collections may turn out not to be unique anymore or physical barcodes might not have certain characters, such as letters, encoded in them, even when these letters are considered part of the local identifier (e.g. the BR in BR0000005117321).
Physically replacing barcodes is a laborious task, which also comes with considerable risks of inadvertently breaking links or creating duplicates. Currently, there is no field in DwC specifically for a verbatim catalogue number (i.e. the exact code as present on the physical

specimen). `dwc:otherCatalogNumbers` can be used, but it is used for other purposes as well (such as database keys and older catalogue numbers).

Another category of identifiers are those for entities such as people, locations and taxa. By using unique identifiers, we avoid format incompatibilities of textual controlled vocabularies. Most Darwin Core fields have an equivalent in the `dwciri` namespace (https://dwc.tdwg.org/rdf/). This allows identification of external resources that identify an entity uniquely. However, this is a relatively new initiative and has yet to be widely adopted in CMS's.

Adoption of globally unique stable identifiers has many advantages for data interoperability. **We recommend that institutions recognise the importance of stability in identifiers and put procedures and policies in place that maintain stability. Identifiers do not persist without constant maintenance, usage and guardianship by the community that uses them. Care needs to be taken to distinguish specific aspects of identifiers, such as barcodes and resolvable URIs, and the difference between the physical specimen and its digital versions.**

3.4.3. Unavailable data should be differentiated from undigitized data

There is a difference between information that is not known, information that is known to be unknown, and information that is known but not yet entered in digital form (digitized). Sometimes such cases can be determined by the use of values such as “unknown”, “S.C.” (???), “none”, etc. but often these three cases cannot be distinguished from one another because relevant fields are left empty or are absent altogether.

We recommend that best practice is to use a standardized value for unknown data for the case where information is known not to be known, and a value of ‘empty’ for the other two cases. Fields without any value should be interpreted as ‘empty’.

3.4.4. Verbatim information and annotations

Specimen data are generally not born digitally. Either they are collected in the field in notebooks or digitized from labels created during a pre-digital era. The different properties that are part of these handwritten or typed texts often have to be interpreted to some extent, e.g. differentiating a scientific name, a person's name or a locality description. However, interpretation can lead to errors, so many data managers prefer that both a verbatim transcription of the data and an interpretation are stored. Darwin Core has several fields with verbatim alternatives, such as `dwc:eventDate` and `dwc:verbatimEventDate`.

Nevertheless, even though these data are registered in CMS's, they are rarely published to aggregators such as GBIF. Yet, they do provide an important source of information that users of aggregated data might find useful. They provide a context to the interpretation and an indication of how much interpretation was required. It has also been argued in impromptu discussions on Darwin Core that verbatim fields should be completely removed and that each instance of digitized specimen data should be treated as an individual version of those data. It can be argued that even a supposed verbatim transcription entails some interpretation of the transcriber. Each version of a specimen's digitized data would then have to be described with metadata explaining to the way the data have been digitized, the assumptions made and methods used. The user would then have to select the version of the data most suited to their user requirements. In the future, one might expect more versions of a specimen's data, because automated systems are being developed to read and interpret specimen data without human intervention.

Interpretation and versioning could be at the individual field level, particularly with data elements related to location and georeferencing, where considerable interpretation is required. Nevertheless, it is not clear how data should be best structured to make them useful. Annotation of whole records or individual fields could be used to capture these data, and currently there is no clear guideline for when annotation and versioning are most appropriate.

Verbatim information, versioning and annotations have long remained an unresolved issue in biodiversity data management. This is a potentially difficult issue for interoperability when data are aggregated. This problem is not only an issue for specimens, we see the same issue in citizen science platforms, such as iNaturalist, where there are numerous identifications, interpretations and annotations of the same photograph.

We recommend that versioning, annotation and verbatim data are considered more centrally in the development of data standards so that all the knowledge on a specimen can be captured, together with all the curational and interpretation steps.

3.4.5. Conclusion

Data standards such as DwC already provide substantial improvements in data interoperability. However, one of the weaknesses of such standards is the trade-off between data interoperability and flexibility to accommodate different data types. We recommend the development of controlled vocabularies and templates for terms to be used for specific types of collection data, such as herbarium specimens. Hence, data interoperability can be improved with minimal impact on the standard's flexibility. For the specific case of data that is unknown, we recommend an unambiguous way to harmonize the absence of data and values of unknown in terms of their actual implications. Finally, it is clear that there is still some work to be done in standardizing methods to account for versioning, annotation and persistent identification of specimens as well as the provenance of the data available on them.

4. References

ABCD task group, 2007. Access to Biological Collection Data (ABCD), Version 2.06. Biodiversity Information Standards (TDWG). Available at: http://www.tdwg.org/standards/115.

Baird, R.C., 2010. Leveraging the fullest potential of scientific collections through digitisation. Biodiversity Informatics, 7(2), pp.130–136. Available at: https://journals.ku.edu/index.php/jbi/article/view/3987.

Beaman, R.S. & Cellinese, N., 2012. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys, 209, pp.7–17. Available at: https://doi.org/10.3897/zookeys.209.3313.

Le Bras, G., Chagnoux, S. & Dillen, M., 2019. Specification of data exchange format for transcription platforms (Version v1). Zenodo. Available at: http://doi.org/10.5281/zenodo.2598413.

Carpinone, E.C., 2010. Museum Collections Management Systems: One Size Does Not Fit All. Seton Hall University Dissertations and Theses (ETDs). 2366. Available at: https://scholarship.shu.edu/dissertations/2366.

Chamberlain, S., 2017. rgbif: Interface to the Global "Biodiversity" Information Facility API. 0.9.9. CRAN. Available at: https://cran.r-project.org/package=rgbif.

Dillen, M. et al., 2019. A benchmark dataset of herbarium specimen images with label data. Biodiversity Data Journal, 7. Available at: https://doi.org/10.3897/bdj.7.e31817.

Güntsch, A. et al., 2017. Actionable, long-Term stable and semantic web compatible identifiers for access to biological collection objects. Database, 2017(1), pp.1–9. Available at: https://doi.org/10.1093/database/bax003.

Heflin, J. & Hendler, J., 2000. Semantic Interoperability on the Web. Available at: https://www.researchgate.net/publication/242233182_Semantic_Interoperability_on_the_Web

Lavoie, C., 2013. Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. Perspectives in Plant Ecology, Evolution and Systematics, 15(1), pp.68–76. Available at: http://dx.doi.org/10.1016/j.ppees.2012.10.002.

Marshall, C.C., 1998. Making Metadata : a study of metadata creation for a mixed physical-digital collection. The International Journal of Digital Curation, pp.162–171. Available at: https://doi.org/10.1145/276675.276693.

Nelson, G., Sweeney, P. & Gilbert, E., 2018. Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens: Applications in Plant Sciences, 6(2), pp.1–7. Available at: https://doi.org/10.1002/aps3.1027.

Ooms, J., 2014. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. Available at: https://arxiv.org/abs/1403.2805.

Owen, D. et al., 2019. Methods for Automated Text Digitisation. ICEDIG D4.1. Available at: https://icedig.eu/sites/default/files/deliverable_d4.1_icedig_methods_for_automated_text_digitisation.pdf.

Parent, C. & Spaccapietra, S., 2000. Database integration : the key to data interoperability. Advances in Object-Oriented Data Modeling, p.31. Available at: https://infoscience.epfl.ch/record/99123/files/OObook.pdf.

R Core Team, 2017. R: A language and environment for statistical computing. Available at: https://www.r-project.org/.

Rainer, H. & Vitek, E., 2009. Virtual herbaria - an open platform to join. In: Stevanović V (Ed.) Book of Abstracts. 5th Balkan Botanical Congress, 2009.

RMNH, 2018. Natural History Collection Management Systems: an international survey on user experience. Naturalis Biodiversity Center, Leiden, the Netherlands. Unpublished data.

Schindel, D.E. & Cook, J.A., 2018. The next generation of natural history collections. PLoS Biology, 16(7), pp.1–8. Available at: https://doi.org/10.1371/journal.pbio.2006125.

Scholes, R.J. et al., 2012. Building a global observing system for biodiversity. Current Opinion in Environmental Sustainability, 4(1), pp.139–146. Available at: https://doi.org/10.1016/j.cosust.2011.12.005.

Schulman, L., Juslén, A. & Lahti, K., 2018. The Finnish Biodiversity Information Facility FinBIF – an integrated open data infrastructure supporting research and decision-making in conservation. Oral presentation by the first author and poster at the 5th European Congress of Conservation Biology, Jyväskylä, Finland, June 12–15, 2018. Abstract available. Available at: https://peerageofscience.org/conference/eccb2018/108028/.

Stork, L. et al., 2018. Semantic annotation of natural history collections. Journal of Web Semantics. Available at: https://doi.org/10.1016/j.websem.2018.06.002.

Vollmar, A., Macklin, J.A. & Ford, L., 2010. Natural History Specimen Digitization: Challenges and Concerns. Biodiversity Informatics, 7(2), pp.93–112. Available at: https://journals.ku.edu/index.php/jbi/article/view/3992.

Wieczorek, J. et al., 2012. Darwin core: An evolving community-developed biodiversity data standard. PLoS ONE, 7(1). Available at: https://doi.org/10.1371/journal.pone.0029715.

Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, pp.1–9. Available at: https://doi.org/10.1038/sdata.2016.18.

Wittenburg, P., 2019. From Persistent Identifiers to Digital Objects to Make Data Science More Efficient. Data Intelligence, 1(1), pp.6–21. Available at: https://doi.org/10.1162/dint_a_00004.

Appendix A: CMS survey classification

Table A1: CMS's, publishing system ("Pub") and library management system ("Lib") if specified in the survey responses. "NONE" is used if no system was described, whereas "UNKNOWN" is used if a CMS is suspected, but its nature was not clear from the response.

Institution Name	CMS1	CMS2	CMS3	Pub	LIB
Stichting De Bastei (Natuurmuseum Nijmegen)	Access	Adlib		NLBIF	
Kuopio Natural History Museum	Kotka			FinBIF	
Royal Museum for Central Africa, Tervuren	DaRWIN			In-house	
Finnish Museum of Natural History, University of Helsinki	Kotka			FinBIF	
Stichting TwentseWelle	Adlib				
Institutionen för ekologi, miljö och geovetenskap, Umeå universitet (Department of Ecology and Environmental Science, Umeä University)	Swedish Virtual Herbarium			Swedish Virtual Herbarium	
Universidade de Porto (MHNC – UP)	NONE				
The Science Museums, Aarhus University	Filemaker			Aubot.dk	
Museo di Storia Naturale dell'Università degli Studi di Firenze	NONE				
Institutionen för biologi och miljövetenskap, Göteborgs universitet (Department of Biological and Environmental Sciences, University of Gothenburg)	Filemaker				
Naturhistorisk Museum Aarhus	SPECIFY				
NIOZ Royal Netherlands Institute for Sea Research	UNKNOWN			IPT	
Natural History Museum, University of Oslo	Musit			MUSIT	

Open Science Centre / Museum, University of Jyväskylä	Kotka			FinBIF	
National and Kapodistrian University of Athens	SPECIFY			Scratchpads	
Université de Toulouse III-Paul Sabatier	Excel	SN-Base			
Westerdijk Fungal Biodiversity Institute	BioloMICS			BioloMICS	
Natural History Museum of Denmark, University of Copenhagen	SPECIFY				
University of Patras	Excel	Access			
Senckenberg Gesellscharft für Naturforschung	In-house			In-house	
Aristotle University of Thessaloniki	NONE				
Västarvet, Göteborgs Naturhistoriska Museum (Västarvet, The Gothernburg Museum of Natural History)	SPECIFY	Access			
Biodiversity Unit, University of Oulu	Kotka			FinBIF	
Conservatoire botanique national Alpin	Jacim				
Université Clermont Auvergne	In-house				
The Royal Botanic Garden Edinburgh	BGBASE	In-house		In-house	ATOM
University of Namur	NONE				
Univerzita Pavla Jozefa Šafárika v Košiciach	In-house				
Muséum d’histoire naturelle Philadelphe-Thomas de Gaillac	ActiMuseo				
Bergianska stiftelsen (Bergius Foundation)	In-house				
Muséum d’histoire naturelle de La Rochelle	Alienorweb				

National Museum of Natural History, Sofia – Bulgarian Academy of Sciences (NMNHS-BAS)				Scratchpads	
Université Claude-Bernard Lyon 1	Filemaker	Excel		In-house	
Stichting Museon	The Museum System				
La Société National des Sciences Naturelles et Mathématiques de Cherbourg	NONE				
Göteborgs botaniska trädgård (Gothenburg Botanical Garden)	IrisBG				
L’Université Pierre et Marie Curie	Filemaker				
Institute of Biodiversity and Ecosystem Research – Bulgarian Academy of Sciences (IBER-BAS)	NONE				
Université Libre de Bruxelles	Access				
Le Jardin Botanique de la Ville de Lyon	4D				
Université de Strasbourg	Dewey				
Natural History Museum Rotterdam	CB 3.0				
Hungarian Natural History Museum	Excel	Access			
Universidade de Lisboa (MUHNAC – Ulisboa)	SPECIFY	Filemaker	Access		
Biodiversity Unit, University of Turku	Kotka			FinBIF	
Uppsala universitet, Evolutionsmuseet (Uppsala University, Museum of Evolution)	Filemaker				
The Agencia Estatal Consejo Superior de Investigaciones Cientificas (CSIC)	In-house			GBIF	

Institut Recherche et Développement	Access				
Université Lille 1 – Sciences et technologies	ActiMuseo			Webmuseo	
Utrecht University Museum	Adlib				
Université de Rennes 1	Excel				
Royal Belgian Institute of Natural Sciences	DaRWIN			NaturalHeritage	Plone
Université de Bourgogne	In-house				
Naturhistoriska riksmuseet (Swedish Museum of Natural History)	SPECIFY	Filemaker		In-house	
Muséum national d’histoire naturelle	Jacim				
The University of Warsaw	In-house	Access			
Museum and Institute of Zoology, Polish Academy of Sciences	In-house				
Natural History Museum of Crete, University of Crete	In-house				
Université de Montpellier	In-house			In-house	
Biologiska institution, Lunds universitet (Department of Biology, Lund University)	Filemaker			Swedish Virtual Herbarium	
Estonian Museum of Natural History	PlutoF	SARV		PlutoF	
Centre de coopération internationale en recherche agronomique pour le développement	UNKNOWN			Plantnet	
Teylers Museum, Haarlem	Adlib				
Centrum für Naturkunde, Universität Hamburg	SPECIFY	Excel	Access		

University of Tartu	PlutoF				
Universidade de Coimbra (MUC – UC)	InPatrimonium	InNatura		Inwebonline	
Instituut voor Natuur – en Bosonderzoek (INBO)	UNKNOWN				
Nature Conservation Agency of the Czech Republic	NONE				
Agentschap Plantentuin Meise	BGBASE	In-house			
The Universidad de Navarra (UNAV)	In-house				
Natural History Museum London	Emu				
Naturalis Biodiversity Center	Atlantis	BRAHMS			
Tallinn University of Technology	SARV			SARV	
Naturhistorisches Museum Wien	JACQ			Biocase	
Musée national d’histoire naturelle (MnhnL)	Recorder				
Natuurmuseum Maastricht	CB 3.0				
Botanischer Garten und Botanisches Museum Berlin – Freie Universität Berlin	In-house	JACQ		JACQ	
Centrum biológie rastlín a biodiversity, Botanický ústav Slovenskej akadémie vied	In-house			In-house	
Royal Zoological Society of Antwerp	ZIMS				
Estonian University of Life Sciences	PlutoF			eBiodiversity	
Universidade de Coimbra (MUC – UC)	SPECIFY			In-house	
Institute of Vertebrate Biology, The Czech Academy of Sciences, Czech Republic	Excel				

Charles University, Faculty of Science	JACQ	Access		JACQ	

Appendix B: Interview form

This form was sent around to all institutions and constituted the structure for the interview.

Description of the deliverable:

D4.4 Interoperability with institutional collection management systems

A specification report for software engineers involved in the development of institutional collection management systems to ensure minimum standards of compatibility between systems. Linking to external services in EOSC will also be covered.

Approach

Interview the database managers of ICEDIG partner institutes, initially those partner in WP4, structured around the questions listed below. The goal is to identify common issues or shortcomings of collection management systems in use.

Questions

1a) What types of specimens do you have in your collection?

- Types of specimens : e.g. *Animals, plants, fungi, DNA, fossil, wood, living or dead, mixed samples, derivatives (drawings, pictures, casts), books…*

1b) Which collection management systems (CMS) do you use to keep track of these specimens?

- Can be a specific, dedicated system such as BG-BASE, Specify, Arctos, BRAHMS, or an in-house development based on an available software environment or database system, such as Filemaker or Access, or even a simple spreadsheet in Excel.

1c) What data on these specimens do you keep in your CMS in general?

- Not just the possibilities of your CMS, but data which is often available in it.

1d) Do you standardize some of these data (e.g. date, collector, taxonomy, identifier)?

- Are there restrictions on what can be put in certain fields and/or periodic reports on (potential) errors ? e.g. dates and coordinates have to be valid, coordinates have to match country, date has to fit realistically within the collector's lifespan…
- Are some fields linked to checklists (e.g. collector lists, habitat classes), gazetteers, (internal) taxonomic backbones, international standard codes (e.g. country codes) ?

2a) Do you have multiple collection management systems? Why?

2b) If so, can you combine data from one with others if necessary?

- e.g. You want to link different kinds of specimen with the same origin (wood, herbarium sheet, DNA sample, picture and microscopic slide of the same original plant). Or you want to find everything you have from a certain collector or species.
- Do you have unique identifiers to readily link data from different CMS's ? e.g. a collector ID which matches the ID for that collector in another CMS for different types of specimens. Or an ID of a living specimen which allows you to track preserved specimens derived from it.

3a) How do you import data to your management system? What type of standard do you require for data to be imported?
- As a case study, how would you import transcribed data from crowdsourcing transcription platforms like DoeDat or DIGIVOL (ca. 200 specimens)? What about data for 10.000 specimens ?
- How much time/work does this take?

3b) How do you think this could be improved?

4a) How do you export/publish data from your management system?
- Do you need any conversion afterwards and, if so, how do you do it?
- How much time does it usually take for typical export operations? How long to do a conversion?

4d) How do you think this could be improved?

5) What data (properties or specimens/observations) do you have but not publish and why?
- Data could be under embargo or sensitive
- Data could also not be interpretable: e.g. lack of proper standard in the database, inconsistent use of database fields.
- Not convertible: e.g. lack of corresponding fields in publishing standard, relational nature (such as host-parasite)
- Not practical: e.g. not enough time and/or money available, operation scope and/or data volumes unwieldy

6) Any other compatibility problems you have encountered with your CMS?

Appendix C: Summary of interview responses

1a) What types of specimens do you have in your collection?

APM:
Primarily biological specimens, living and preserved. A substantial diatom collection, which often constitute mixed samples. Also DNA samples and a few botanical curiosities.

LUOMUS:
The collections comprise animal, plant, fungal, mineral and fossil samples. There's living and preserved specimens and soon to be a DNA databank. The library management is kept separate.

MNHN:
Geological, biological and cultural specimens of all sorts. Also animal and environmental sounds. There's also an anthropological and ethnological collection.

NHM:
Geological, biological and cultural specimens of all sorts, except living specimens.

RBGK:
Plants: Herbarium (c. 7,000,000). Collection Spirit (76,000)
Fungi: 1,250,000 specimens
Economic botany: 100,000
Seeds: 85,800
DNA and Tissue bank: 58,000
Microscope slides: 150,000 (includes 40,000 palynology slides, 10,500 fungi)
In Vitro Collection: 6,000 (includes orchids, mycorrhizal and non-mycorrhizal fungi)
Library: 300,000 printed volumes, 5,000 journal title and 20,000 maps
Illustrations: 200,000 prints and drawings
Archive: 4,600 collections – correspondence, notebooks and photograph albums, records of plants received and sent out from Kew, maps and plans of Kew
Living: 178,000

RMNH:
The types of specimens in the Naturalis collection are zoological, botanical, geological and paleontological specimens. There is also a DNA collection, casts and a collection of 2D material (books, drawings etc.).

UT:
Geological, botanical, mycological, zoological and environmental or DNA samples.

1b) Which collection management systems (CMS) do you use to keep track of these specimens?

APM:

At APM, BG-BASE is used for non-living botanical collections and LivCol for the living collections. LivCol is a Progress-based in-house database. Its living collection data are planned to be transferred into BG-BASE in 2019. BG-BASE is a proprietary system based on Revelation Software architecture and contains more than 6.000 potential data fields. Details of DNA samples are currently still kept in Excel sheets outside the database. Media management is done separately. There is no automated publishing.

LUOMUS:

At LUOMUS, a system called Kotka is used. This system is developed at LUOMUS and used throughout Finland. There are still separate databases for vertebrates, such as Selma, for which data migration into Kotka is underway. For vascular plants of Fennoscandia there is Kastikka (5M specimens, migration into Kotka is planned) and there are also a few other subcollections in MS-Excel and MS-Access. Kotka itself does not cover publication, but data are automatically ingested from Kotka into the FinBIF data warehouse, where they are processed and subsequently published to its portal and other data publishers.

Kotka covers multiple sorts of specimens and manages data from different institutions. Sometimes, different institutions want different functions or vocabularies. This can cause some interoperability problems when organizations develop the habit of submitting their own data in a slightly different manner.

MNHN:

At the MNHN, an in-house built Java front-end called JACIM is used, which is also in use by some institutes in France. The front-end connects to a cluster of 15 Oracle databases, with each database covering a different type of specimen (living vs preserved plants, arthropods, minerals…). Media and loans management is done in a separate system.

The system is quite old and proprietary. It is primarily a balance between different needs at a very large scale. A more relational database would be preferred.

NHM:

At the NHM, a modified version of Axiell EMu is used. A small part of botany uses BRAHMS, but this is to be incorporated into EMu as well. Axiell CALM deals with the management of archives. OpenText takes care of the digital asset management and there is a separate system called Freezer Pro for the frozen collections. A Data Portal platform was developed in-house using CKAN for open publication of collections data

RBGK:

At RBGK, different systems are used for different collections. A Sybase system built in-house called HerbCat covers the botanical collections (sheets, spirits). The fungarium is covered by the Herbtrack and Herb IMI databases, the latter originating from another institute, CABI, but the merged collection now the curatory responsibility of Kew. Loans and transactions are tracked separately through a system called CRIS. There are also separate, all in-house built, databases for the Economic Botany collection, the DNA and Tissue bank, the microscope slides and the living collections. The seed collection uses BRAHMS.

RMNH:

At RMNH, the Oxford University-developed system BRAHMS is used for botanical collections and BioXL, an Atlantis-based system developed by Dutch company DeventIT, for zoological and geological collections. BRAHMS is MS SQL based and incorporates image management.

UT:
At UT, the cloud-based system PlutoF is used for most database needs. Only for geological collections SARV is used. Both systems have been developed by a consortium in Estonia, the Natural history archives and information network (NATARC), and are now branching out towards other countries. Development of PlutoF is based at UT. PlutoF is based on a MySQL database, which can be consulted through a web interface. The system incorporates some analysis services, validation checklists, annotation capabilities and automated publishing of data (if open). Multimedia management is also covered by PlutoF.

1c) What data on these specimens do you keep in your CMS in general?

APM: ca. 4M specimens
Available data includes specimen nature, current location, cultivated or not, donor, filing name, collector, collection number, collection date, country, locality, coordinates, habitat, description, elevation, associated material, type status, type name, barcode and much more. If a data field is not readily available, it can be easily added.

LUOMUS: ca. 13M specimens
Documentation on the Kotka data fields can be found here. Data includes information on specimen ownership, the collection event, location, type status, taxonomic identifications, identifiers, current location, preparations and other notes.

MNHN: ca. 65M specimens
Documentation on the different databases can be found here. A lot of available data fields, but for recent digitizations only the basic info is available.

NHM: ca. 80M specimens
A graphic depiction of the data model can be found here. The completeness and quality of data varies quite widely across the collection dataset.

- Specimen data includes:
  - Geographic location (different levels of verbatim, atomised data and georeferenced data)
  - Specimen types and preservation methods
  - Derivatives (parents/parts, preparations)
  - Collectors and collection details
  - Stratigraphy (chrono-, litho-, bio-)
  - Taxonomy and determinations
  - Storage locations
  - Bibliographic records
  - A small amount of analysis data
  - Multimedia (mainly specimen, label and register images)
  - Registration and index lots (multi-specimen records)

- Process-related data includes:
  - Registration
  - Acquisitions and donations
  - Condition reporting
  - Valuations
  - Exhibitions
  - Loans
  - Disposals

RBGK: ca. 9M specimens
Answers For the Herbarium Catalogue only.

Barcode (or other ID); Preparation Code (i.e. sheet or spirit); Type status; Type Qualifier; Determinations with family, genus, species, author, infraSpec Rank, InfraSpec Name, Infra Spec Author, determiner, determiner date and determination notes. There are tick boxes for the name the species was filed under, of which it is a type and whether the name is on the sheet or not for each determination.

Collector; Collection number; Collection date (range); Date from label; Country; TDWG code; Country from label; Herbarium region; Max and min altitude; Locality description (First division, Second division, Place Name – fields available but used less often); Latitude; Longitude; Source; Reference for source; Accuracy of coordinates. Note that GID (Group Identifier) fields could do with improvement.

Habitat; Plant description; Private notes; Restrictions (free text); Uses; General Comments; Duplicates.

Donation details. Donation date and Donor (these fields used in spirit collection more than herbarium).

Cultivated tick box. Cultivated date and cultivator.

RMNH: ca. 41M specimens
The data contains fields for identification, names of collector(s) and/or donator(s), dates (collection date, donation date), gathering site, sex, phase or stage, general info about the specimen as a unique registration number, basis of records, preserved part, mount and storage location. The data model of BioXL is ABCD-EFG based. Furthermore, for a lot of specimens a scan/photo is available. These are not stored in the collection management system but in a separate media library.

UT: ca. 1M specimens
The data model can be seen here. Mandatory fields if you want to upload include scientific name, collection date, locality, collector and determiner. This can be skipped if you have the rights to do it, for old specimens. Mandatory fields not needed if you only want to keep these data local, but if they're to be published somewhere officially or moved into an official collection, they're required.

Specimen datasets connect to other modules, like DNA module, trait module, ecological data. There's an interaction module for interactions between specimens. A checklist exists for the sort of interaction.

Ecological data are also supported. Functionality is present for nested location tables for vegetation surveys (plots and subplots).

1d) Do you standardize some of these data ?

APM:
Often both linked (to another table) and verbatim fields are available.
- Collector: Linked field links through ID to collector table, but there can be multiple ways to specify a collector's name. There exists a standard how to put in names, but it is not always followed. That is to say: for the same collector ID, a collector's name can be syntaxed in multiple ways for different specimens.
- Collection date: Has to be full date, but resolution (day, month, year) can be specified in separate column. Allows for a second date if a range, but no differences in resolution between the two. If not a full date, the default is half (i.e. 15th of month or 30 JUN / 01 JUL).
- Link to gazetteers are possible. ISO and TDWG locality codes can be added in specific fields. Coordinates need to be possible.
- Taxonomy: Species tables are part of the database and constitute a backbone to which specimens can be linked. Links need to be made manually or during import. The backbone in use is the result of a few data dumps from RBGE and IPNI and manual edits.

LUOMUS:
There are two types of validations: errors (can't save) and warnings. For instance, collector name must be entered as last, then first name. A warning occurs if this protocol is not followed (there is no collector checklist). Some fields are mandatory and some fields have limited optional values, also during batch import. Coordinates need to fit the location. There is a taxonomic backbone, focused on Finnish species, which suggests names, but does not impose restrictions.

MNHN:
The only thing common to all databases is country code. There are some shared checklists, e.g. molluscs and fish. It really depends on the database in question, but no link to a backbone. The philosophy is: digitize first, taxonomy later.

NHM:
There is a certain amount of basic validation on some fields. However, much of the basic field-level data validation was removed during the original NHM implementation of EMu and not re-applied to the data entry interfaces, which has caused major legacy and ongoing data quality issues. The CMS Data Management team have invested a large amount of resources in the last few years to work through these issues and prevent their reoccurrence, but it still represents a significant challenge. More sophisticated validation methods (e.g. matching dates to collector lifespans) have been considered for some time, but aren't feasible to run within EMu, and if run in bulk outside of EMu, face a technical and resourcing challenge to get the corrected data back in.

Many fields are linked to internal checklists such as taxonomies, collectors, sites and stratigraphy. However, in many cases these haven't been curated in a coordinated way over the years, and there are ongoing efforts by the Data Management team to clean up and de-duplicate those parts of the dataset. There are some imported reference lists (e.g. stratigraphy, specific taxonomies for parts of the collection, ISO-3166 for country codes etc) but in many places external standards are not currently used.

There is some automated mapping of internal fields to Darwin Core terms within EMu, which is mainly to streamline publication of data via the Data Portal.

RBGK:
Country, TDWG code and herbarium region must match drop down list value. Selecting a value in one field restricts options in other fields. It is possible to select another value that does not match. A warning is given, but you can save anyway. Dates must be a valid date. Coordinates must also be valid. Maximum altitude must be greater than minimum altitude. Collector list is a messy look-up list: names get added to it whenever someone enters a new collector.

For taxonomic names you can look up IPNI or the World Checklist of Selected Plant Families (WCSP) to automatically populate fields. You can also search Tropicos, but not populate fields with it. You can also just add new names.

There are look-up lists for: Plant parts, Preparation Code, Type Status, Type Qualifier, County, TDWG code, Herbarium Region and Source.

Coordinates do not have to match country and there's no validation of dates to other pieces of information (e.g. biographical data on the collector).

RMNH:
Some fields are standardized. For BioXL, there are mandatory fields like the registration number, where the prefix (institution and collection code) must be chosen from a predefined list. This is the same for fields like collection name, preserved part, mount, property of, taxon rank and basis of record. For fields like identification, gathering site, sex, etc. the user can enter a verbatim value and a relation to a thesaurus concept can be made, manually or automatic. A geographic gazetteer (Getty) is available, as well as the international Stratigraphic list and for some collections a taxonomic list is available. Furthermore there are lists for sex, phase or stage, taxon rank, type status, additional number type, collecting method, etc. Dates are recorded verbatim and automatically transformed in 'real' dates.

In Brahms taxonomic, geographical and collectors list are available.

UT:
Most of the time, there are validated and verbatim fields.
- Taxonomy: Validation to GBIF backbone, Genebank and a few other online resources, such as Index Fungorum and Fauna Europaea because these are updated faster than the Catalogue of Life itself. Their own backbone is based on these resources. If all fails, custom entry can be done, but needs to be validated by taxonomic expert. There are also DNA-based species and taxa, identified through DOIs.

- Collector: Linked to a checklist. If not found, you have to add a new entry to this list. Problem of duplicates due to spelling, formatting differences.
- Collection date: Can be partial date, can be a range.

Every occurrence can have unlimited number of identifications. Only new ones can be added, not old ones removed. History can be tracked this way. Identifications can be at different taxon ranks: e.g. a fungus, later identified by a DOI. This sort of versioned annotations system is not present for all types of data, e.g. not for coordinates.

2a) Do you have multiple collection management systems? Why?

APM:
Yes, but to be phased out and merged. VUBIS is used by the library and will be kept separate.

LUOMUS:
Yes. Older systems to be merged into Kotka.

NHM:
Disregarding the library and archives systems, we now have one core CMS, Axiell EMu (BRAHMS still exists for legacy/cultural reasons, but still has EMu as the master data source). However, we also have Freezer Pro for frozen sample management, as it’s much better designed for that work than EMu. OpenText for handling multimedia was integrated some years ago as a back end for storing collections images and serving them through the EMu interface, which was mainly a strategic move to consolidate digital asset management across the different parts of the Museum. The Data Portal was developed because EMu did not offer a suitable web interface for the publication of collections data, and also to provide an open repository for NHM research datasets.

RBGK:
Yes. The systems are not joined up, in part because they were developed at different times with different teams involved. Although it is not sure if one system could do everything, e.g. the system requirements for the living collection and DNA bank could be different.

RMNH:
Yes, as described above. The reason is historical: as of 2010, the National Museum of Natural History (Naturalis) combined with the Zoological Museum Amsterdam (ZMA), and the Dutch National Herbaria to form the Nederlands Centrum voor Biodiversiteit (NCB Naturalis). The Zoological and geological data were merged together in BioXL and the botanical data in Brahms. At the moment, a survey for a new collection management system for all specimens is ongoing.

2b) If so, can you combine data from one with others if necessary?

APM:
A link between living and preserved specimens is possible, but not straightforward and not always readily implemented. A link between the different identifiers in both systems needs to have been made somewhere. No information on link between specimens and DNA data.

LUOMUS:

It is a lot of manual work doing the transfer from old systems into Kotka, including validations during the procedure. Often new Kotka features are needed.

MNHN:

Everything is synthesized on the portal. However, one would want interoperable taxonomy, geography, collector names… In particular collector names.

NHM:

Internal Record Numbers (IRN) link everything together within EMu. Across all systems, GUIDs are used as appropriate record identifiers. There are also a range of different legacy specimen identifier series, with many different formats, which are used in registers, on labels and barcodes, and in citations. These are unique within specific contexts but not across the collection. A common specimen identifier series was introduced a few years ago, mainly for barcoding in mass digitisation projects, but there have been cultural challenges in getting it adopted more widely across the collection.

A lack of standards and connectors mean that compatibility with other systems is poor in general. EMu is only available as a Windows desktop client, which restricts compatibility both with other OS's but also with web services. There is no ability currently to easily create references to and extract data from external standards and checklists.

RBGK:

Not without difficulty. There is no one place to search everything or where it is linked up. It has been an ambition to upgrade systems and link for the past 5 years or more. Some information through all systems is made available through POWO (Plants Of the World Online), but still much info to add. There are no universal standards or IDs to easily match together specimens: one would need to match by collector name and collection number, where there could be different formats. Some linking is facilitated by adding, for example, the barcode of the herbarium specimen voucher in the DNA and tissue bank voucher, but this is not always happening and hence mostly no linking across is possible.

RMNH:

No immediate links between systems, but rarely necessary given the different nature of the collections. Everything can be searched through the Naturalis BioPortal.

3ab) Import

APM:

The vocabulary needs to be checked, then the data are converted to the Revelation format, and subsequently imported. Very easy to add new data fields, if necessary for an import. No conditional overwrites are possible: a scripted edit of a record overwrites all information of that record, even if that implies deleting it. Scripted batch changes can be made to multiple records. Linked fields need to be linked manually or semi-manually by putting in the proper internal ID's. For multi-value fields, the proper delimiter needs to be added.

LUOMUS:

Import is done with Excel sheets: one tab for data, one tab with validation of the data fields, listing the values allowed (to avoid errors) or suggested (to avoid warnings). Multi-value fields need to be put in as extra columns, for example MYGathering[0][MYUnit][0][MYIdentification][0][MYTaxon] and MYGathering[0][MYUnit][0][MYIdentification][1][MYTaxon] for two taxon names in two identifications). If no multi-value fields occur, about 177 data fields are supported right now. The terminology used is loosely based on DwC. Issues encountered: validations could be more strict and Excel can cause data corruptions.

MNHN:

There are few linked fields and the process is iterative, throwing up warnings/errors for what did not fit in the database.

NHM:

Import is done using an online interface or from Excel spreadsheet/Access database. Data needs to be atomised, normalised and cleaned before import. Links need to be made with existing collector records or with newly added collector records.

Linking with existing records would be easier if these records were in direct sync with web services ensuring their quality (e.g. taxonomic backbones, collector PIDs). Same for automated validation tools (e.g. coordinates and locality). Standards for crowdsourced data quality would facilitate the workflow of these sort of imports. Unstructured data (e.g. from OCR, suggested associated records, crowdsourcing) can be expected to become more prominent in the future, so tools to store and query these optimally would be helpful.

RBGK:

Need to convert data into the correct format for import. The final table needs to be in Access and have correct field names or be mapped to the fields in the import routine. The import is completed through an admin function in the client. A table in MS Access is uploaded and converted into XML as the first step and then this is uploaded into the database. There is a test system available to check any import first. There is some validation and report on import including checking to see if records are already in the system. You can make changes in the XML in the intermediate step.

Some fields cannot be imported into through this import routine, e.g. TDWG standard codes. This happens because some fields were added later to the database, but not to the import routine. You would need to run a separate update query through Access. Also, you cannot add multiple determinations through this routine; this should also be completed through an update query. It is not possible to overwrite records or add additional information to records, e.g. coordinates, through the import routine.

The process could be improved by enabling import to all fields, updates to records and by being consistent with the naming of fields (name in database different from name displayed different from name needed for import routine). Also, by allowing import in different formats such as DwCA.

RMNH:

Brahms uses RDE (Rapid Data Entry) files for imports. for BioXL XML based templates are used. In both cases the import files represent the data model.

UT:
Premade CSV template files depending on what sort of data. The system automaps them, to other tables as well. Anything that couldn’t be linked is given back for evaluation to the importer. Batch scripting possible by IT experts and recommended for large datasets. It’s particularly important if matching to a web service (like GBIF backbone) is required.

There’s a policy that, on loan, you have to transcribe it into the system. New labels are printed by the system, so all data are present there (and often more).
A template with the original info as it exists in the system can be downloaded, then new info added to it, followed by an updating upload. Edits can also be done manually in PlutoF if you have the authorization as a user.. You can’t add extra coordinate interpretations. Annotation system available for many other fields, but not this one.
There’s a major problem of getting data out of (old) black box systems with poor and/or undocumented standards, and convert it for an import. Also not a lot of specimens in Estonia: 2 million in total across all fields. Problems might arise when scaling up more. Collaboration with supercomputer centers for optimal data aggregation.

4) Export

APM:
Data can be exported locally as PDF or CSV files. Exports can take very long if many linked fields are present.

For export to the web, unrestricted data are exported from the database by manually launching an export script. The resulting flat export files (tab-separated value files) are mapped to DwC by semi-automated scripts in the R programming language. These scripts also filter out certain data problems, including problems with data integrity (delimiter corruption) and specimen identifiers. The scripts produce a CSV file, which is then manually published to GBIF through the own IPT server.

The same export is also uploaded to the APM virtual herbarium and then imported using ETL scripts. Minimal validation is performed before the upload in R, to clear out corrupted file delimitation and broken identifiers. First data publication was done in March 2018. More frequent, monthly updates are expected in 2019 after problems resulting from database update complications are resolved.

LUOMUS:
Exports can be made into Excel sheets by users. This allows batch edits through export-import. These exports take little time (max. a few minutes), but are limited to 10.000 records.

Automated export occurs from Kotka into the FinBIF data warehouse, from which it is indexed into the FinBIF portal. The rate is about 1.000 specimens per minute. DwC mapping also happens at the FinBIF side. Publishing to GBIF happens from FinBIF.

MNHN:

Data are exported from Oracle databases into a data warehouse, where they are merged with exports from the media management system (the “mediathèque”) concerning data on images. These data are mapped onto DwC or the schema of the MNHN portal and then published there. The process happens weekly and is fully automated.

NHM:
Data are exported five days a week in incremental dumps from the CMS into a proprietary format by an automated script written in Perl. The export is then converted, mapped and validated by Python scripts into a PostgreSQL database. These scripts also perform a number of data integrity validations, filter out records inappropriate for publication and build required links between data and multimedia assets. The data portal’s CKAN Solr service then updates its indexes using a query against the PostgreSQL database. Updates are incremental and take about six hours, mostly because the timing of the cron job is set very generously. Full exports might be required occasionally and are much slower, both in the data export step and the data load and reindexing steps

A closer integration with the CMS would be the key to improving this process. If there were a suitable API, then it would be possible to extract the data directly from the CMS database rather than relying on a regular dump. Failing that, it would be easier if the CMS dump were in a standard format e.g. JSON rather than a proprietary format, allowing the use of standard import tools and libraries. Ultimately, there should be no need for the separate PostgreSQL database for the data portal. The portal should just act as a public interface to the CMS database via the API, with transformations, filters and all being handled by security and business rules within the CMS. This tight integration would allow the Data Portal to provide real-time public access to the collections dataset.

RBGK:
Automatically the data are published online with records appearing the next day on the online Catalogue. However searching and download from the site is limited. Query results are limited to 20,000 records and the maximum you can download is 1,000 records. Not all data fields are returned in the summary that can be downloaded. To download more data-rich records, you would have to get them one by one. Hence why we get lots of requests for downloads or actually send others to GBIF to download data from there. Data can be downloaded by us through MS Access though queries, but this can be very slow.

It depends on the complexity of the data, but it can take around 2 hours to do a query and export as it is really slow. You also need to think about the logistics of how to query to get the data you need, so it will be even longer if you need to repeat. It can take a day to get the export you want or if it crashes even more than that.

We also send data to GBIF, Europeana, iDigbio and Reflora virtual Herbarium as DwC-A/ABCD using Biocase. A new archive should be produced weekly, but this has often broken due to the time needed to create the archive on the infrastructure and we have had issues getting the mapping correct. Issues with mapping are still ongoing. Better documentation for mapping fields in BioCase with examples would help.

RMNH:

Data from the two CMS's is exported, converted and mapped, then uploaded to the Netherlands Biodiversity API – which also publishes to GBIF. This process happens every three months and is slow and labor-intensive, because it is not a delta procedure.

UT:
If data is open, it can be queried through an API. It will also automatically be published to GBIF and other repositories, as well as on the own portal. This portal can only be searched by those who have an account? For data to be open, some data fields are required.

5) What data do you have but not publish and why?

APM:
Never published "manuscript names" are withheld, as are any specimens flagged by an embargo checkbox. The filter scripts before publishing ensure validity of the specific identifier (i.e. the barcode) and remove specimens with faulty values which conflict with the ETL scripts, such as the presence of tabs in data fields or double quotes conflicting with separated value file string delimitation..

LUOMUS:
For sensitive data, such as threatened species, information on locality is made less precise. This is done automatically based on a Finnish list of species with sensitive data compiled by an expert group. Certain data issues may cause the specimens to be hidden from searches by default on the FinBIF portal.

MNHN:
Sensitive specimens. Data which can't be mapped to Darwin Core or is not important enough.

NHM:
Exceptions are mostly for political and legal reasons (confidential, sensitive, third party rights…).

Records are also omitted if they have insufficient data to be published or for relational reasons, such as child or preparation records which would result in duplicates upon publication.

RBGK:
Lots of data are transcribed, but collated in other systems, e.g. MS-Access database for minimal folder level data. Specimen label data are then captured through crowdsourcing. Individual researchers have their own databases (mostly in MS-Access or Brahms). Much data are awaiting import due to lack of resources for performing the operation. Sometimes projects do not release their data for import until the project is finished. Additionally, some databases will be managed continuously, meaning that if they're imported into the institutional database, the data can be modified in two places causing divergence.

Some specimens are restricted due to collecting agreements with countries restricting their reuse, e.g. some Millennium Seed Bank partnerships. There are some issues with georeferenced fields: there is a lack of fields for georeferencing notes and accuracy information, so this is not imported.

There are also some issues with hidden characters and diacritics.

RMNH:
Sensitive data and research data of which the manuscript is not published yet (i.e. under embargo). There is also some information in the CMS's which has less value and is impractical to publish.

TU:
PlutoF attempts to avoid such inconsistencies by doing data management from the very beginning. Most limitations are restrictions at the publisher side.

Now material samples are possible and DNA results from high throughput sequencing. But there is the issue that not all data are searchable on GBIF.

Verbatim text fields suffer from inconsistency. Annotations may move data to correct fields.

Fear of open data, not under creator's control. Proper citation functions, like GBIF's, are therefore very important.