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 D6.5

Work package WP6 / Task T6.1 / Lead: University of Helsinki
Deadline M21

DOI: 10.5281/zenodo.3465285

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Open access implementation guidelines for DiSSCo

DELIVERABLE D6.5

Authors:

Hannu Saarenmaa, Donat Agosti, Mathias Dillen, Willi Egloff, Pierre-Yves Gagnier, Quentin Groom, Alex Hardisty, Niels Raes

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

Abstract

The paper investigates how to implement open access to data in collection institutions and in the DiSSCo research infrastructure. Large-scale digitisation projects generate lots of images, but data transcription often remains backlogged for years. The paper discusses minimum information standards (MIDS) for digital specimens, and tentatively defines 4 hierarchical MIDS levels. Even partially available data can be useful for some purposes and it is recommended that data and media be made openly accessible after minimal delay. The paper then discusses the FAIR concepts and obstacles and restrictions to making data openly accessible. Legal restrictions for making data openly available are discussed from the perspective "as open as possible, as closed as (legally) necessary". Specific legislation applying to DiSSCo data are listed and guidelines on how to deal with each are made. The information architecture of DiSSCo is briefly described from the perspective of facilitating open access to data, and the essential features of DiSSCo and GBIF are compared against each other, noting that both have their distinctive roles but need close cooperation. Data policies and data management plans (DMP) of six ICEDIG collection institutions are reviewed. Typically, only one or the other exists, depending on how digitisation and related data management is organised. Data policies are found at institution level whereas DMPs seem to belong to project level. The paper comes to the following conclusions: 1) Digital Specimen Objects (DSO) must be findable and accessible already at the lowest (MIDS-0) level. Data should be deposited in a trusted public research data repository. 2) As far as possible, projects must enable third parties to access, mine, exploit, reproduce and disseminate this data by using a copyright waiver such as CC0 or an open access licence such as CC-BY. 3) Exceptions to openness policy must be stated clearly and strictly limited to reasons of national security, legal or regulatory compliance, sensitivity of collection information, and third party rights.

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

Contents

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

Introduction

To set the scene for this document, we need to refer briefly to the general objectives of ICEDIG WP6 on Data infrastructure:

> *"ICEDIG needs to design the data flows and storage infrastructure for DiSSCo. Exact user requirements will be determined at project start and in cooperation with other work packages, but it is assumed that up to 40 million specimens may be digitised each year, and images and data may weigh up to 100 MB for each specimen. These data will be generated in distributed digitisation facilities across Europe, but need to be accessed and analysed in one pool. All data will be freely and openly accessible globally, with the exceptions that have legal basis."*

Task 6.1 Open access and data management plan is then described as follows:

> *"Open access to biodiversity data is given in declarations of Bouchout, GBIF, GEOSS, RDA, UN CBD, etc., which Task 7.3 will assess, but how that can best be put in operation in the context or the DiSSCo needs to be designed for. Institutional policies and their implementation will be considered, as well as the restrictions for sensitive data. The task will also take all actions necessary for participating in the Open Research Data pilot, and prepare the DMP for the ICEDIG project."*

This deliverable report was preceded by a manuscript which was the Milestone MS35 document at Month 13. It was reviewed widely within the ICEDIG project, but also distributed externally. The received comments have been taken into account in expanding the manuscript into this final deliverable report.

This Deliverable D6.5 Open access implementation guidelines for DiSSCo, due at Month 21, has been described as follows:

> *"A concise report on how to implement the principles of Open Data / Open Access in DiSSCo with the basic premise 'as open as possible, as closed as (legally) necessary'. The report will be based, e.g., on the experiences of the Finnish Biodiversity Information Facility FinBIF, and other major data sharing initiatives."*

The emphasis here is in the words: *"How to implement"*. Open access has been the norm in biodiversity science since the establishment of the Global Biodiversity Information Facility (GBIF) in 2001. Many if not most collection institutions now have data policies which reflect this aim. However, there still are obstacles which cause that the data may not findable, accessible, interoperable, and re-usable (FAIR), as recommended by EU policies on research data (European Commission 2017).

Legal barriers, intellectual property, and lack and slowness of digitisation are major factors in hampering implementation of open access. Simple guidelines how to deal with each of these need to be put forward so that collection owners have clarity on these major factors and do not need to delay publishing their data.

Minimum information for digital specimens (MIDS)

Digitisation requires many steps and is never complete. The ICEDIG proposal puts it like this: *"Digitisation is annotation"*. After imaging, and extraction or entry of rudimentary metadata, years can pass until the full data about the collecting event and identification of the species is available. The full data also may require quality checking before they are released as Open Data. The major question is can we somehow speed up this process? Whether and when to make incomplete and even erroneous data available for scientific discovery, augmentation, and annotation? How do collection institutions handle this process, and at what point does data become Open Data? Answers to these questions are needed, because when DiSSCo is operational, it will be producing images of tens of millions of specimens every year, and it will not be feasible to leave them off-line until data capture is fully complete. Pioneering institutions, many of which are ICEDIG partners, already have produced millions of images of specimens in their collections. It will be important to learn from their experiences in making these data available.

Defining the levels of digitisation is an important starting point for this study. GBIF (2015) established a task force on accelerating the discovery of bio-collections data which has started its work by defining the essential information needed about various types of collections. These 'metadata' will describe the contents of each collection and help data holders assess and prioritize their digitization activities. There now is an ongoing discussion on the subject. In the following text we will define the Minimum Information for Digital Specimens (MIDS), which will be a foundation for the discussion later in this document. We understand that the below definitions are tentative and the work at GBIF, CETAF Digitisation Working Group, and ICEDIG/DiSSCo will be consolidated at some point. However, for the purposes of this document we need to have working definitions now.

A Digital Specimen Object (DSO) is a special type of Digital Object (e.g., Kahn and Wilensky 2006) that serves as a digital representation of the physical object held in a scientific collection (ICEDIG MS37 report[1], Hardisty et al. 2019). Other kinds of objects, such as Digital Collection Objects, also are possible. A DSO can be characterised as a container that can be identified through one or more globally unique identifiers, such as DOI, Handle, LSID, URI, UUID, etc. There is a recommendation by the CETAF Information Science and Technology Committee (ISTC) of the form of a persistent unique identifier for physical specimens[2], but that is not sufficient for all the purposes of DiSSCo (Hardisty et al. 2019). A DSO packages metadata, data, media files, and annotations. A DSO may also offer "methods", i.e., executable code that would perform operations on the DSO. A DSO and its contents can be versioned. DSOs can be implemented in many ways. In its simplest form, a DSO is just a folder in a file system, containing XML and JPG files, and which is addressable through a web server. An advanced implementation could involve object-oriented database in a cloud service. When DiSSCo gets into its construction phase, choices how to implement DSOs need to be made.

We also need to define metadata, data, and annotations. Data (a.k.a., "full data", "real data", "content data") tells something about the physical specimen, such as scientific name, location where collected, date collected, collector name, etc. Data also includes images and other media. Data typically originates from the labels attached to the specimen, its containing box or drawer, and in

[1] Hardisty A, et al. ICEDIG MS37 - Requirements and specifications of the DiSSCo infrastructure. Slide document version 0.4, 2019-01-29 https://dissco.teamwork.com/index.cfm#files/6219996

[2] https://cetaf.org/cetaf-stable-identifiers

some cases the related rows in a field notebook. Metadata are technical, relating to the process of digitisation, such as the date of creation of the DSO and for each document it contains, the creator, equipment used, quality checks made, media file characteristics (e.g., EXIF). Metadata relates also to the management and use of digitised data. So, we have multiple kinds of metadata for different purposes. Interpretations (or determinations) are authorised modifications of the original data made by a competent curator or transcriber. Examples of such include conversion of verbatim date format to ISO standard and assigning a new identification to the species. Annotations are assertions and comments made on the DSO by other users. Annotations may become data when processed and accepted by a responsible curator. An automatic OCR result of the labels may be considered as a machine-made annotation until verified and structured. Annotations are different from interpretations. In other words, annotations are staging posts to data, i.e., suggestions made by external actors which have not yet been incorporated to data by the responsible curator and may cover also other information not present in the label data.

For the sake of the discussion further below, we tentatively suggest the following MIDS levels. The purpose of the MIDS approach is to offer clarity to collections about minimum quantity and quality of information they should be publishing to make Digital Specimens useful for multiple purposes of teaching and learning, research, etc. MIDS is divided into hierarchical levels because not all digitisation programmes can publish the 'gold standard' MIDS-2 level information right from the beginning; thus, we accept that lesser quantities of information at an earlier stage can still be useful for many purposes.

MIDS-0 -- Catalogue: Modern mass-digitisation techniques are based on imaging all specimens and their labels and creating a basic DSO by interpreting any identifiers (such as barcodes) among the labels. At this level the DSO only contains metadata and zero or more media files. This level includes the following Darwin Core[3] (DwC) elements that are related to the process of digitisation and collection management rather than the specimen.

- `DwC:occurrenceID` – a NSId
- `DwC:institutionCode` – from e.g., *Index Herbariorum* and other catalogues
- `DwC:catalogNumber` – automatically readable from the specimen label; must be attached to the specimen prior to imaging
- `DwC:dynamicProperties` – a list to store any technical metadata such as the quality checks of the digitisation result that have been made

MIDS-1 – Basic: Includes MIDS-0 but adds basic data elements that can be entered in a bulk operation by a human operator for a number of DSOs. In other words, MIDS-1 does not include any data that must be read from the label(s) attached to the specimen(s) but instead are inherited from collection level labels. Most scientific collections include this information in their boxes and folders (plants), or drawers and unit trays (insects). These elements typically are:

- `DwC:scientificName` – at some taxonomic level
- `DwC:higherGeography` – at some accuracy such as “Palearctic”, or “Europe”

[3] http://rs.tdwg.org/dwc/terms/

- `DwC:collectionCode` – within the institution. If a donated collection is being digitised as a whole this could be also the name of the collector, i.e., the same value as in `DwC:recordedBy`

MIDS-2 -- Regular: Includes MIDS-1 and the most important data elements describing the specimen and the collecting event, and which have been transcribed and interpreted from the specimen labels. These data include location, date, collector name, and scientific name, and involve using many more DwC elements, so we do not list them here. Optionally, the `DwC:dynamicProperties` element includes metadata of the quality control of the transcription result, following the ISO 2859 standard (e.g., Mononen et al. 2014) or some other protocol.

MIDS-3 -- Extended: Includes MIDS-2 but adds interpretations made using external information sources (beyond what can be determined from specimen and collection labels). Example of this is finding the geographic coordinates of the collecting locality through research on gazetteers or field notebooks. Also an interpretation is asserting a taxonomic concept to the specimen (`DwC:taxonID`)and the currently valid scientific name (MIDS-1 and MIDS-2 level scientific names are not necessarily the valid ones). Mapping text in the verbatim DwC elements from MIDS-2 into the corresponding well-structured DwC elements and updating their values to the current situation also counts as interpretation.

Additional data: Including images and additional media such as sounds, chemical extracts, DNA-barcodes, 3D models, and OCR results does not fit in the above classification. Such media can be added at any MIDS level and should be flagged up separately in metadata. Another form of data that should be considered are links between specimens and external data sources.

The rationale for making data openly accessible and FAIR concepts

Citing the Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 (European Commission 2017),

> *“modern research builds on extensive scientific dialogue and advances by improving earlier work. The Europe 2020 strategy for a smart, sustainable and inclusive economy underlines the central role of knowledge and innovation in generating growth. Broader access to scientific publications and data therefore helps to:*
>
> - *build on previous research results (improved quality of results)*
> - *encourage collaboration and avoid duplication of effort (greater efficiency)*
> - *speed up innovation (faster progress to market means faster growth)*
> - *involve citizens and society (improved transparency of the scientific process)*
>
> *This is why the EU wants to improve access to scientific information and to boost the benefits of public investment in research funded under Horizon 2020.”*

At what level should a DSO then be made findable as Open Data? Based on the above rationale, the answer is simple and clear: at MIDS-0! However, the MIDS level should always be documented.

This recommendation may be surprising. What is the utility of MIDS-0 level data? The answer is that we cannot imagine all uses of the data in advance. Sometimes data will be used for science, sometimes for innovations, even bioprospecting. For example, scientists developing machine learning methods for interpreting data from additional images can benefit from a large pool of “raw” MIDS-0 level DSOs for training their algorithms. Taxonomists may benefit from MIDS-1 level data of unidentified specimens when searching specimens for a taxonomic revision. Curators will benefit from MIDS-2 level data when managing their collection. Modellers will appreciate MIDS-3 data so that they do not need to spend much time in data cleaning.

Making data openly accessible follows the steps of making it Findable, Accessible, Interoperable, and Reusable (FAIR). The requirements of the Open Research Data Pilot for H2020 projects require only the two first steps (Findable, Accessible):

1. Preferably, projects “*must deposit the research data in a public research data repository*”, such as EUDAT, Zenodo, etc.
2. “*As far as possible, projects must take measures to enable third parties to access, mine, exploit, reproduce and disseminate this research data*”, which would require using an open access license, such as Creative Commons CC-BY or by public domain dedication, such as Creative Commons Zero (CC0), and to provide minimal machine-readability.

We should note that these requirements are for H2020 projects such as ICEDIG, and not exactly for ESFRI research infrastructures, such as DiSSCo. For the latter, the full range of FAIR services can be required, and as permanent research infrastructures they can be expected to build their own repositories and access mechanisms[4].

Making data interoperable means using standard vocabularies, taxonomies, and ontologies, and standard exchange formats such as XML, JSON, RDF, etc. as the means of moving data from one place to another. The biodiversity community has this well covered through its data standards Darwin Core (DwC), ABCD, and Ecological Metadata Language (EML); and the taxonomies of PESI and Catalogue of Life. In order to support other kinds of scientific collections such as rocks, minerals, fossils, core samples, etc., this list of standards must be widened (e.g. ABCD EFG extension for geology). Ontologies are being used sparingly, but the Biological Collections Ontology (BCO) (Walls et al. 2014) is an example and a base for further work. The Finnish Biodiversity Information Facility (FinBIF) has been built on the FinBIF ontology[5], where all its classes and their properties are described according to the specifications of the Resource Description Framework (RDF). Backend of the operational FinBIF system is a triple store on Oracle.

Making data available for re-use is the last aspect of FAIR that must be considered. This was discussed in the ICEDIG DMP (D6.1) as follows. The H2020 DMP template places this question: *“When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.”* The draft DiSSCo DMP (MS37) puts it like this: *"as a rule, digital specimen data shall be made publicly available as soon after digitisation, verification and curation as is practically possible”.* DiSSCo data should be as open as possible and as closed as legally required.

[4] See, for instance, http://nsidr.org/
[5] http://schema.laji.fi/ and https://laji.fi/en/about/796

The exceptions for data re-use include, for example sensitive species, (a short) embargo for ongoing research, and personal information.

Legal restrictions

Our motto is "*as open as possible, as closed as (legally) necessary"*. So, what is legally necessary? Specific legislation applying to DiSSCo data are the following:

- International conventions related to natural science e.g., CBD, CITES, CMS, etc.
- Nagoya protocol on access and benefit sharing
- GDPR (Regulation (EU) 2016/679)
- PSI (Directive 2013/37/EU) and the Open Data Directive (Directive 2019/1024)
- INSPIRE Directive (2007/2/EEC)
- Habitats Directive (92/43/EEC), Birds Directive (amended, 2009/147/EEC) and protected areas (Natura 2000);
- Relevant national legislation

We present below a short summary of the implications of each with regard to open access.

From international conventions there are only a few requirements for data accessibility. These are related to preservation of traditional knowledge and use of genetic resources. Otherwise there is general aim to promote open access and lower the barriers to data sharing, as the document UNEP/CBD/COP/11/INF/8 *"A review of barriers to the sharing of biodiversity data and information, with recommendations for eliminating them*" well states.

The Nagoya Protocol[6] on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization (ABS) to the Convention on Biological Diversity is a supplementary agreement to the Convention on Biological Diversity. It provides a transparent legal framework for the effective implementation of one of the three objectives of the CBD: the fair and equitable sharing of benefits arising out of the utilization of genetic resources.

The provisions in the Nagoya Protocol pertaining to the users' obligations are transposed by a regulation of the European Union (Regulation of the European Parliament and the Council (EU), which is binding for all EU member countries. The EU regulation implementing the Protocol within the Union came into force in October 2015. The instructions and guidelines related to the EU Decree (scope of application and sector-specific instructions) are still under way.

According to this EU regulation, taxonomic and systematics research is not "utilization". Even DNA barcoding is not utilization, but wider chemical analysis is. So, publishing distribution maps, etc., is not regulated, as it is concerned with mere description of background material.

The Nagoya protocol does not restrict museums from accepting materials. However, their utilization must be restricted, if the country of origin so demands. This requires that each collection must be documented properly, so that users can be informed of what it can be used for. This may be an overwhelming task if the collection contains material which have been collected from many places in many gathering events.

[6] https://www.cbd.int/abs/about/

When accepting materials to a public collection, 1) collecting permit, 2) export permit, and 3) where applicable, an import permit should be furnished. Furthermore, when foreign genetic resources are used for research and/or development activities, 4) a Prior Informed Consent (PIC) must be acquired for their use, and, 5) if the country of origin so requires, Mutually Agreed Terms (MAT) must be established between the provider and the user. The PIC must always be acquired if traditional knowledge of indigenous people (or local communities) is included in the use of the genetic resources.

In principle this implies that collection institutions must maintain their repository of these five documents and link the documents to the pertaining parts of collections and eventually down to specimen level. We have not yet seen this implemented anywhere.

The EU General Data Protection Regulation[7] (GDPR) should be considered for collector names. That is, names of living persons, not deceased as is the case of collectors of many older specimens. When combined with location and time which also can be found in specimen labels, the movement of the collector could be tracked. On the other hand, knowing the name of the collector is an important factor in determining data quality and re-use in transcription of new data. Therefore, it is the practice in natural science to also release the collector name and make it searchable, as well. What is held and processed is actually 'professional information' and not usually 'personal information'. This is comparable to the fact that there is a public interest in holding and processing data about parliamentary and government representatives, and, e.g., names of members of an orchestra. Furthermore, deposition of a specimen in a public collection may be considered equivalent to publication, although the act of deposition may not always come directly from the collector. Nevertheless, it is not fully clear how far the GDPR rules apply in the case of publicly held scientific collections, and more work to clarify issues is probably needed. This will be touched below, when we look at the privacy statements of some ICEDIG collection institutions.

How to deal with information about persons who transcribe and annotate the data has not yet been discussed. When these persons are employees, their information can be dealt through institutional policies. In the case of external users of systems, pseudonyms are often used, and the identity of the person is known only by the institution that hosts the information system.

The new Open Data Directive (Directive (EU) 2019/1024) has taken over the rules introduced by the PSI Directive; in addition, it addresses the remaining and emerging barriers to a wide re-use of publicly funded information across the Union and brings the legislative framework up to date with the advances in digital technologies. Minimum harmonisation of national rules and practices on the re-use of publicly funded information should contribute to the smooth functioning of the internal market and the proper development of the information society in the Union.

Under the new rules[8]:

> *All public sector content that can be accessed under national access to documents rules is in principle freely available for re-use. With this Directive, Public sector bodies are not be able to charge more than the marginal cost for the reuse of their data, except in very limited cases.*

[7] https://ec.europa.eu/info/law/law-topic/data-protection_en

[8] Citation from https://ec.europa.eu/digital-single-market/en/public-sector-information-psi-directive-open-data-directive

> *This will allow more SMEs and start-ups to enter new markets in providing data-based products and services.*
>
> *A particular focus is placed on high-value datasets such as statistics or geospatial data. These datasets have a high commercial potential and can speed up the emergence of a wide variety of value-added information products and services. The Commission is now working together with the Member States to define the list of specific high-value datasets that can be made available for free and easily re-usable across the entire European Union.*
>
> *Public undertakings in the transport and utilities sector generate valuable data when providing services in the general interest that will enter into the scope of the Open Data and Public Sector Information Directive: once the public undertakings make such data available, they will have to comply with the principles of transparency, non-discrimination and non-exclusivity set out in the Directive and ensure the use of appropriate data formats and dissemination methods. They will still be able to set reasonable charges to recover the costs of producing the data and of making it available for re-use.*
>
> *Some public bodies strike complex data deals with private companies, which can potentially lead to public sector information being 'locked in'. Safeguards are therefore put in place to reinforce transparency and to limit the conclusion of agreements which could lead to exclusive re-use of public sector data by private partners.*
>
> *More real-time data, available via Application Programming Interfaces (APIs), can allow companies, especially start-ups, to develop innovative products and services, e.g. mobility apps.*
>
> *Publicly-funded research data is also being brought into the scope of the directive: Member States are required to develop policies for open access to publicly funded research data while harmonised rules on re-use will be applied to all publicly-funded research data which is made accessible via repositories.*

Research data has been covered by Article 10 of the Directive:

1. *Member States shall support the availability of research data by adopting national policies and relevant actions aiming at making publicly funded research data openly available ('open access policies'), following the principle of 'open by default' and compatible with the FAIR principles. In that context, concerns relating to intellectual property rights, personal data protection and confidentiality, security and legitimate commercial interests, shall be taken into account in accordance with the principle of 'as open as possible, as closed as necessary'. Those open access policies shall be addressed to research performing organisations and research funding organisations.*
2. *… research data shall be re-usable for commercial or non-commercial purposes … insofar as they are publicly funded and researchers, research performing organisations or research funding organisations have already made them publicly available through an institutional or subject-based repository. In that context, legitimate commercial interests, knowledge transfer activities and pre-existing intellectual property rights shall be taken into account.*

Digitisation of cultural heritage has been covered by Article 12 of the Directive. It is focussing only on giving private actors, which are carrying out digitisation, exclusive rights to the public distribution of their results for a limited time that can be maximum 10 years.

The Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)[9], in its Article 13 lists eight cases where Member States may limit public access to spatial data sets and services where such access would adversely affect international relations, public security or national defence. Several of these cases, such as IPR issues and personal information may be relevant for scientific collections. At least, the last case is *"the protection of the environment to which such information relates, such as the location of rare species".*

The INSPIRE Directive is special in not only allowing some restrictions, but also because it obligates Member States to actively provide on-line services to access some data. Its Annex III includes the following:

> *18. Habitats and biotopes*
>
> *Geographical areas characterised by specific ecological conditions, processes, structure, and (life support) functions that physically support the organisms that live there. Includes terrestrial and aquatic areas distinguished by geographical, abiotic and biotic features, whether entirely natural or semi-natural.*
>
> *19. Species distribution*
>
> *Geographical distribution of occurrence of animal and plant species aggregated by grid, region, administrative unit or other analytical unit.*

The Habitats Directive (92/43/EEC), Birds Directive (amended, 2009/147/EEC) and the regulation on protected areas (Natura 2000) do not directly forbid publishing data on the listed species. However, maintaining their populations at proper levels is required, which may require restricting releasing such data, but that is not clearly stated.

Following through all national legislation may become a challenge in a European initiative such as DiSSCo. What is sensitive is a complicated question to answer, though in general includes rare species that are vulnerable to exploitation or persecution. Data providers in each country usually know what is sensitive in their domain and in their country, but do not necessarily know that for other domains. Yet scientific collections routinely publish data of specimens collected in other than their own country, which has sometimes led to protests from the country of origin. There is neither a European, nor global clearing-house of these issues, which is something DiSSCo probably should tackle.

Nevertheless, restrictions can only be put in place on a justified basis according to objective criteria in the applicable legislation.

[9] https://inspire.ec.europa.eu/documents/directive-20072ec-european-parliament-and-council-14-march-2007-establishing

DiSSCo information architecture and open access

It is assumed that DSOs have a life-cycle where they migrate from MIDS-0 towards higher MIDS levels. This can be visualised in the high-level engineering architecture view (Fig. 1).

The top level data flow diagram of DiSSCo is shown in Fig 2. This diagram is being detailed with several level 2 diagrams such as: 1) pre-digitisation curation, 2) imaging station set-up, 3) imaging and image processing, 4) image archiving, 5) data capture, 6) retrieving digital specimen metadata and content. This is not yet an exhaustive list. We discuss below a new level 2 diagram (Fig. 3) for data publishing.

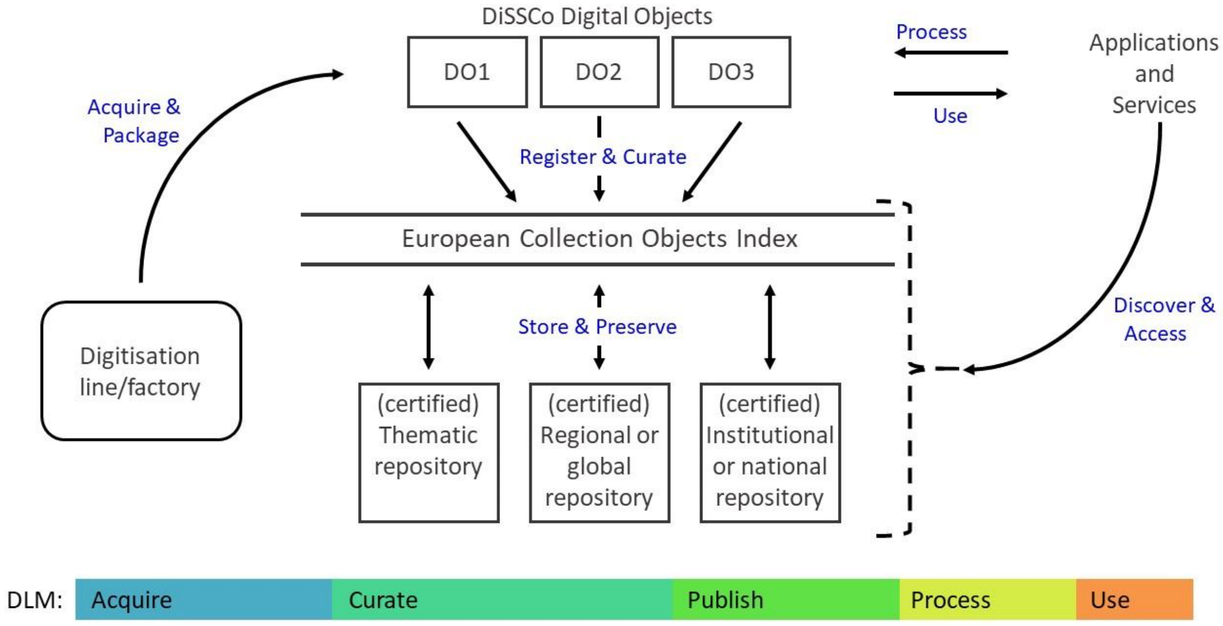
DiSSCo Digital Objects
DO1
DO2
DO3
Process
Use
Applications and Services
Acquire & Package
Register & Curate
European Collection Objects Index
Store & Preserve
Discover & Access
Digitisation line/factory
(certified) Thematic repository
(certified) Regional or global repository
(certified) Institutional or national repository
DLM:
Acquire
Curate
Publish
Process
Use

Figure 1. High-level engineering architecture (API) view of DiSSCo (from ICEDIG MS37 report[1]). Digital specimens (DO 1…3) typically evolve through different MIDS levels as additional information is added to them.

Data publishing begins with an institutional open access policy. DiSSCo also has an open access policy, which is described in the DiSSCo DMP Section 6.3.1 (Hardisty et al. 2019). It states that *"the DiSSCo open access policy must be implemented at DiSSCo Facility level, using institutional open access policies in conjunction with the open access recommendations of DiSSCo*[10] *to adopt MIDS and MICS"*. The policy should preferably be machine-readable and be executed automatically (periodically such as daily, weekly, monthly) by the data publishing process. The publishing process should really be automatic, so that the data do not remain unnecessary inaccessible. This automation might be best done in cooperation between institutional data managers and a forthcoming DiSSCo technical helpdesk.

[10] The bold text refers to this document ICEDIG D6.5.

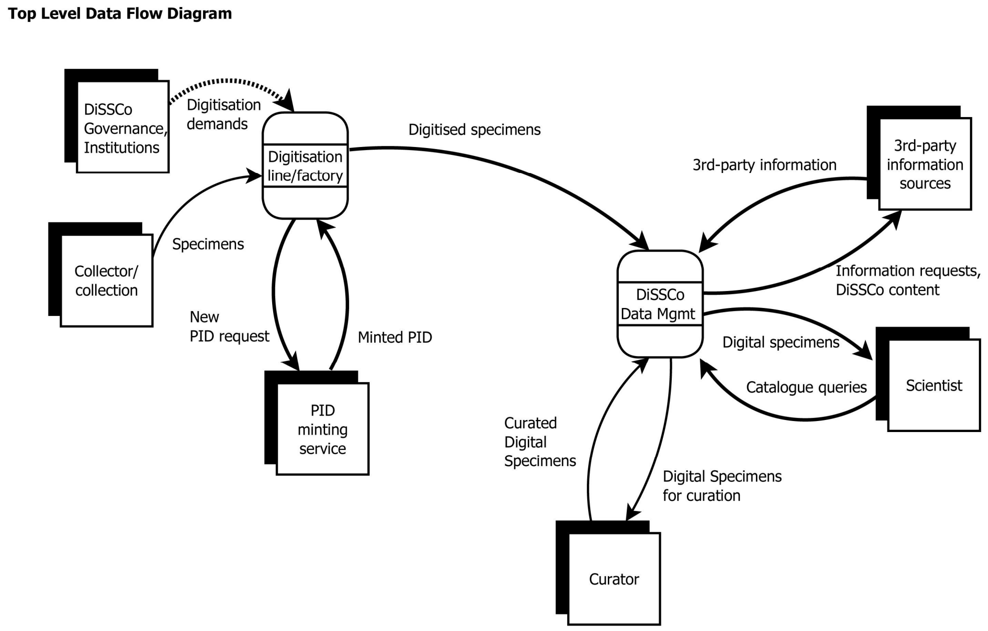
Top Level Data Flow Diagram
DiSSCo Governance, Institutions
Digitisation demands
Digitisation line/factory
Digitised specimens
Collector/ collection
Specimens
New PID request
Minted PID
PID minting service
3rd-party information
3rd-party information sources
DiSSCo Data Mgmt
Information requests, DiSSCo content
Digital specimens
Catalogue queries
Scientist
Curated Digital Specimens
Digital Specimens for curation
Curator

Figure 2. The top-level data flow diagram in DiSSCo digital specimen architecture (from ICEDIG MS37 report[1]).

The publishing process needs to adhere to the quality control process of data, and to MIDS-levels. Data would be exported from an institutional collection management system to a publishing queue such as the GBIF IPT (Robertson et al. 2014). The queue would run automatically and produce a publishing log.

The data would land in one or more open access repositories. Some of these would offer annotation services to external users. The annotations would be fed back to the quality control process and close the loop for migrating data through MIDS levels.

Answering these questions also requires defining the relationship between institutional collection management systems and the DiSSCo data infrastructure. This is not within the scope of this document, but is being covered by the DiSSCo DMP (Hardisty et al. 2019), and by the Deliverable D4.4 on Interoperability with institutional collection management systems (Dillen et al. 2019).

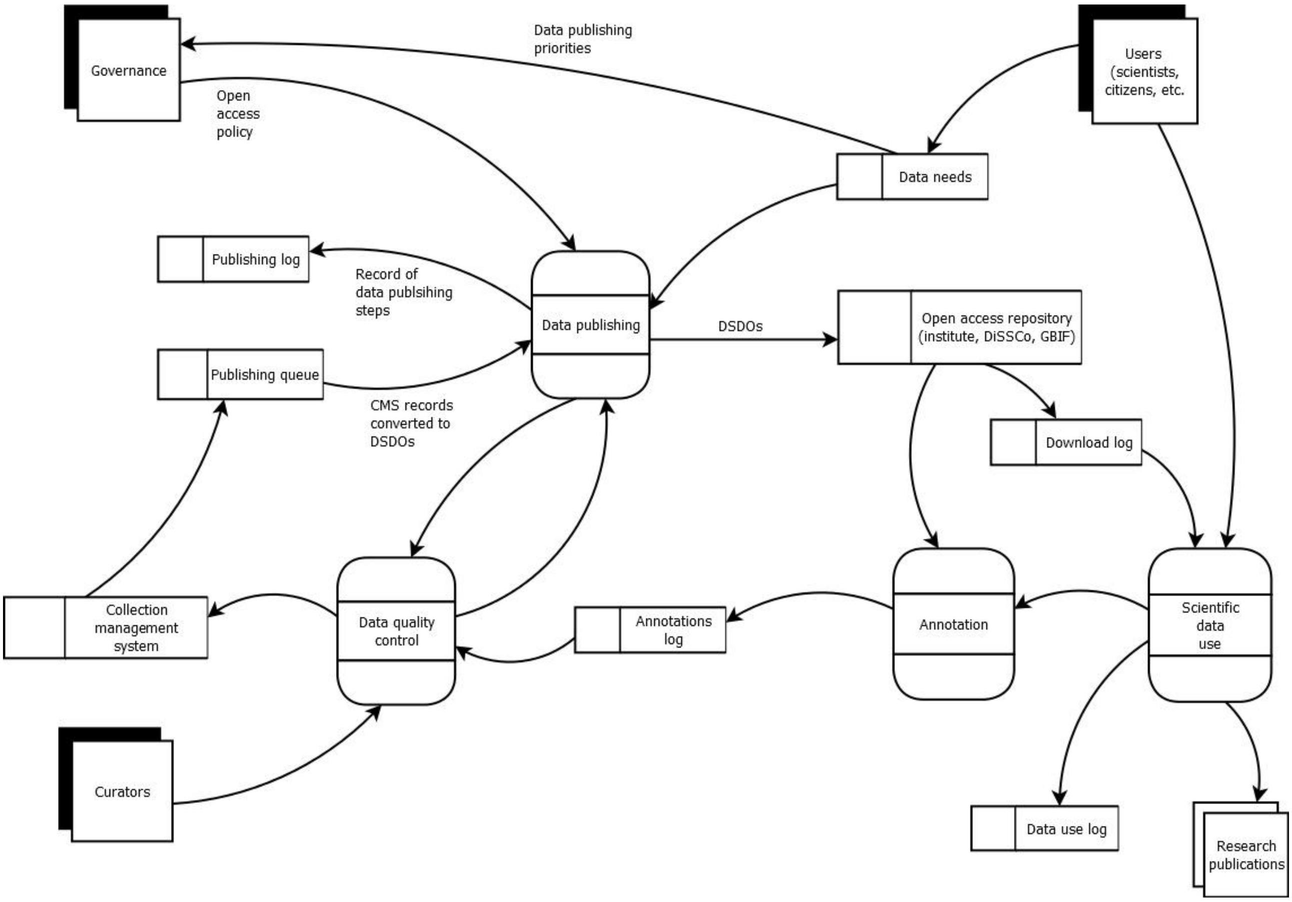

Figure 3. Level 2 data flow diagram for data publishing, the related quality control, and data use. This diagram does not cover data curation fully. That would be a different DFD, which will have implications to this diagram. (Version 2; available as slide document from https://dissco.teamwork.com/index.cfm#milestones/514475)

What is the relationship between DiSSCo and GBIF, and how do their respective missions and scopes compare? This is a common question. The main difference is that DiSSCo will not build a portal that would be a static copy of secondary data, exported from some primary data source where data is born and will be curated. Instead, the DiSSCo research infrastructure will be used to curate data and receive, in near real time, fresh data from various digitisation factories and facilities. There will be one virtual collection, which is a live system. Table 1 compares DiSSCo and GBIF from several angles. This comparison is still tentative as many features of DiSSCo are still being defined, but it seems that there are separate roles for both and a need for close cooperation.

At this writing, GBIF boasts 1,357 million occurrences. About 172 million of them represent preserved specimens (12.7%). This is a surprisingly small proportion, given the fact that GBIF was largely launched by the museum community. DiSSCo therefore has an important role to help fill this gap in GBIF data.

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

Table 1. Tentative DiSSCo to GBIF comparison.

	DiSSCo	GBIF
Content scope	Geo- and biodiversity	Biodiversity
Geographic scope	European collections, global data	Global
Basis of record	Specimens (”hard evidence”)	Any occurrences
Data structure	DOA, DOIP	DwC star schema
MIDS level	Any	MIDS-2 and up
Portal data flow	Bi-directional community curation system	Uni-directional data flow
Vocabulary	DiSSCo vocabulary (ontology?)	DwC vocabulary
Curation	Yes	No
Annotations	Unified Curation and Annotation System, with provenance	Available but disconnected
Purpose	One virtual collection	Sharing of any biodiversity data
Access	As open as possible, as closed as required	Free and open only
Data source	Keeps an authentic up-to-date copy of data of members	Receives selected data from institutions
Data production	Digitisation facilities	No
Identifiers[11]	Stable PID	Volatile for occurrences; DOIs for datasets

[11] *“...this is perhaps the most significant reason for tight collaboration between DiSSCo, GBIF and other aggregators to ensure that all specimens have a single PID in use everywhere.”* – D.Hobern

Review of Data Policies of ICEDIG collection institutions

This section presents facts and figures of the databases and data policies of ICEDIG collection institutions. Text in *italics* has been selectively copied "as is" from the document in question. At the end a short conclusion is presented on how the open access policy has been implemented at each institution. The sources can be found in the list of References at the end of this document (pages 32-33).

Botanic Garden Meise Herbarium

Partner abbreviation:	APM
Size of collections:	Approximately 4 million
Records on own site:	1,731,265 (39% of specimens)
GBIF records:	1,719,681 (100% of own records)
Licenses:	CC-BY 4.0 on data, CC-BY-SA on images, CC0 for barcode, scientific name and country name.
Data policy:	Refer to the DMP

> *The DMP will help us improve data integrity and security. It will also clarify what access management is required and who is responsible for this. The data of the Garden should be used to improve opportunities for research collaboration and funding. It will enhance the research reputation of the Garden, particularly if data are cited properly.*
>
> *Data created by the Garden are assets of the Garden, they are valuable from a scientific, management, historical and cultural perspective. They are also public assets that have current and future value. As a public research institution we need to ensure that we comply with Flemish and Belgian regulations on access to and storage of data. We also need to ensure that we are aligned with the policies of other Flemish and Belgian institutions responsible for biodiversity data, such as natural history museums, universities and other institutes. For the Flemish Government an Open Data policy is considered the norm for public data. The administration has created a useful guide to Open Data*[12]*. Movement towards Open Data in the Flemish community is not currently mandatory for research institutions and we do not have a clear position of how the administration sees data from museums and herbaria (cf. European legislation on reuse of public sector information)*[13]*. Long term storage will be conducted at VIAA, which is the Flemish Institute for Archiving that was founded to provide access to multimedia from the cultural, heritage and media sectors.*
>
> *Finally, it is recognised that science as a whole has a poor record on research reproducibility (Baker 2016). In order to improve this record it is essential that data used in research are properly documented and identifiable. Many of the data we administer require validation and can change as more information becomes available. This presents challenges for research*

[12] http://opendataforum.info/images/Open_Data_Handboek_20141119.pdf

[13] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

reproducibility from the perspective of tracking and data integrity. However, proper processes will ensure that this can be achieved cheaply and reliably.

DMP: *The scope of this DMP[14] are the outputs of herbarium digitization.*

Although this DMP was created as a result of the DOE! project, it covers all images and data of herbarium specimens (including related collections such as alcohol material, wood samples, DNA samples, etc.) and as such is an institution-wide policy paper.

Process of publishing:

The DMP has a section on Roles and responsibilities where 4 different roles have been identified. Decisions on data sharing have been assigned to the Curator role. Details of the publishing process have not been described. The DMP states

The sharing of images and data with people outside the garden has proven to be a controversial subject. Prior to the completion of this data management plan a consultation was held with scientists in the garden to decide on a strategy for data access and sharing. It is this strategy that will be given below. However, this should not be considered to be the final word on data access and sharing, because it is yet to be seen if this is compatible with the statutory obligations of the garden and additional internal and external stakeholders may need to be consulted and considered.

Exceptions to openness policy:

Users of the online data portal will have freedom to browse and search the digital data of the herbarium without the need to login. They will be able to view data and view high quality images as they do currently.

Basic data for download are available from the website. Basic data consists of the following data fields: Taxon, Family, Collection date, Collection country, Accession number (barcode), Collector, Collection number, Type status, Permanent URI, Suggested citation, Licensing information, Data providence, Contact email address.

Except for the exceptions listed below, full data will be available only to approved users who have the permission of the curator. The curator should respond to requests for data as soon as possible, with a maximum of one month from the date of the request. Reasons for rejecting a request should be given. The curator is expected to come to their decision after consultation with the Garden's scientists and considering ongoing projects in the garden.

Such downloads will contain the basic fields plus these additional fields if requested: location, coordinates, georeferencing error range, habitat, determiner, determination date, vernacular name, ecological data, ethnobotanical uses, Information on collection permits, associated species, associated specimens, macroscopical and or microscopical characteristics, details of former herbaria where the specimen was deposited.

[14] The Botanic Garden Meise Herbarium Data Management Plan, version 6.1, 11 May 2018, UDO_2018_06_12_bijlage_3_DataManagementPlanDOE v6.1 (1).docx

Exceptions

- *Any data made openly available in publications or public databases, such as on the Global Biodiversity Information Facility (GBIF) or as supplementary data in publications.*
- *Data from the whole Belgian vascular plant herbarium.*
- *Any data transcribed by volunteers on the DoeDat website, including the Gall herbarium and Crépin's Rose Herbarium, will be made available in full.*

Embargoing data

It is recognized that scientists expend considerable effort collecting and identifying their own specimens and digitizing the data on those specimens. They should have a grace period of exclusive use of those data, before they are available to the community as a whole. Should a scientist wish to do so, they can embargo use of the specimens they work on and their associated data. This embargo will last for four years from the date of digitization. This embargo only applies to data collected by a scientist or digitized on request of the scientist. The scientist has to argue the why they are blocking the data. When the project has concluded, the scientist should inform the database manager so that the embargo can be lifted. No specimen should be blocked that is specifically referred to in a publication. If a specimen is published to the virtual herbarium already it cannot be removed. When the embargo period has expired the embargo will be lifted after consultation with the scientist, who has the possibility to extend the embargo. While data are embargoed it will be invisible to users of the data portal of the Garden.

Note that it is important that specimens collected on expeditions are digitized and mounted as soon as possible upon return from the mission. This will protect the garden by ensuring that documentation procedures are followed and that details related to collecting and export permits are digitized immediately.

Sensitive data

There may be occasions where we are requested to restrict access to data. The reason may be to protect the sites of rare species, such as those listed under CITES, but there may be other reasons, for example protecting the biographical data of living collectors who may not want these shared.

The data on plants and their localities will not be considered sensitive by default. Should we be asked to obscure information on the grounds of sensitivity we will review it on a case-by-case basis, considering issues of whether the data are available elsewhere and whether the benefits of secrecy outweigh the potential risks. Currently, we do not restrict access to any specimens on grounds of sensitivity and have never been asked to restrict access for this reason.

We will only make full biographical information available for dead collectors. Living collectors will be identified only by their name.

Annotation, feedback: Records can be publicly commented through the Annosys system[15]*.*

[15] https://annosys.bgbm.fu-berlin.de

Terms of use:

Conclusion:

Digitisation has progressed well and currently 39% of all specimens have been digitised and published. Interestingly, nearly all data on the institutional portal also is available on GBIF. The process of publishing has not been described in detail in the DMP but is available in Dillen et al. (2019), but roles are clear. Hiding of certain details such as coordinates is an interesting decision, which we cover in the overall Discussion below.

Muséum national d'histoire naturelle à Paris

Partner abbreviation: MNHN

Size of collections: Approximately 68 million specimens.

The Museum has about 25 different databases, which are being consolidated in the JACIM collection management system developed in-house. In addition, the Museum is the bearer of the Récolnat national infrastructure.

Records on own site, MNHN specimens : 7,755,000 (11% of the estimated specimens)

Records on own site, Récolnat specimens : 10,000,000 (10% of the estimated specimens including MNHN)

GBIF recordings: 7,275,501 (94% of own records)

Licenses:

CC-BY-NC-ND ; CC-BY-NC ; CC-BY 4.0 In-house work is underway to make our licenses compliant with open science national regulations and limited to CC-BY 4.0.

Data Policy: Ongoing redaction

The data policy of the collection information system of the National Museum of Natural History in Paris is governed by an internal regulation respecting the laws and obligations of the Musée de France.

In its data policy, the MNHN's compliance with policies, strategies and programs related to free access to digital data decided to national level and the EU, including the policy of the French government open science.

The public nature of the Museum's activities is governed by the laws of the Musée de France, universities and the law on the transparency of government activities, which have a direct impact on the data policy of the Museum.

Our research data management policy is being studied to assess storage and management to more clearly examine current and future use of this data. This policy should also reflect the requirements established by the funding agencies (governmental or not), while adhering to the FAIR guidelines for the management and management of scientific data. The policy of the MNHN must also comply with our institutional mission and the National Plan for Open Science.

In addition to its collections, the MNHN provides the scientific responsibility inventories conducted under the Environmental Code which establishes for the entire land territory, river and marine national inventory of natural heritage, defined as " inventory of ecological, faunal, floristic, geological, mineralogical and paleontological resources ."

One of its components, the INPN (National Inventory of Natural Heritage), set up in order to ensure the standardized way the provision of synthesis data necessary for expertise, the development of conservation strategies and the dissemination of national and international information and reports on natural heritage French (plant and animal species, natural environments and geological heritage).

DMP: Being drafted based on adaptation to French legislation.

Process of publishing:

The mission of the Museum's technical teams is to record specimens of collections in the database. Since 2019, priority has been given to computerization and digitization of incoming and outgoing material. It is therefore, in addition to monitoring collections, to structure the digitization policy on demand.

As part of the research programs, the data should be published after the completion of a research project, either on an institutional repository or on an external data repository in accordance with the regulations and the requirements of this policy. This policy should be extended to all areas of research in the coming years.

Researchers will have to respect the FAIR principles for their data. The MNHN management will put in place the tools for their dissemination and conservation. Likewise, for training, support and advice related to practices and regulations in data management, as well as to provide methods and procedures for controlled access to research data.

Annotation comments:

The records cannot be commented publicly by users. Development is in discussion. Annotations and comments could only be submitted by a Colhelper registered user[16], the web-based tools for requesting physical access to collections) or to the webmaster. Direct access via Recolnat's explore interface[17] is under study for registered users and comments are already arriving in this way.

Exceptions to the opening policy:

The MNHN has an officer to put in place the regulations for the protection of personal data, GDPR. The provision of data should not violate the protection of individuals.

Open data is available to everyone. Open science legislation allows for the restriction of access to certain sensitive data.

The interpretation is open to discussion, and institution internal policy is not yet established. Some of the restricted data do not reach consensus, such as those related to the specific location of certain sensitive species.

Due to their nature, in the anthropology collections, the Web access to some specimens is restricted.

[16] http://colhelper.mnhn.fr/requests

[17] https://www.recolnat.org/en/

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

Current research data may be subject to an embargo. This embargo is of variable duration. For imaging, a two-year period may be proposed, renewable once. In case of access requests, the demand is relayed to the researchers for the follow-up. At the end of the embargo period, the data is automatically published.

In the research framework-related by an industrial agreement, the terms of the contract may request a ban on the data publication.

The data production is stored directly in a directory of the institution that manages the publication as the eventual embargo.

Term on the use of data: In preparation

Conclusion:

The MNHN Paris has a data policy in the process of being defined. Its implementation is planned for the years 2019–2020, and to cover the whole range of data. All digitization and scans are directly visible on the institutional database for an estimate of approximate 11% of the total number specimens. The data management plan is being drafted and will be presented as a separate appendix in the institution's digital strategic plan.
The digitization efforts are now focused on incoming material, and massive digitization will be on slides. The rest of the collection will be linked to success in seeking funding. The publication of our data to the GBIF is done directly by IPT.

Natural History Museum, London

Partner abbreviation: NHM

Size of collections: About 80 million specimens

Records on own site: 4,203,000 (5% of specimens)

GBIF records: 3,804,000 (91% of records on own site)

Licences:

Several, depending on dataset. CC-BY-4.0 (21), CC-BY-SA-4.0 (16), CC-BY-4.0 (15), ODC-BY-1.0 (11), CC0-1.0 (4), License not specified (30), other licenses (15)

Data policy: Consists of 5 different policies:

1. Data Protection Policy
2. Staff and Volunteer Data Protection Policy
3. Digital Collection Programme: Digital Licensing and Citation Policy
4. Managing Science Data Embargoes Policy
5. Open Data Exceptions Policy

Digital Collection Programme: Digital Licensing and Citation Policy states the following (in the summary)

The Museum must commit to Open Data and Open Access principles, to be consistent with UK government and European Commission objectives, peer institutions and the wider scientific *community. This strategy would also significantly enhance our collaborative research potential and position as a world-leading resource of natural history and biodiversity data. The argument for this approach has been laid out in detail in the 'Open by Default' Natural History Museum digital information issues paper reviewed by the Board of Trustees in 2013, and Open Access principles have already been incorporated into the Museum's intellectual property policy. The policies in this document are intended to build upon that position.*

Open access principles are applicable to the vast majority of digital assets generated by the Natural History Museum's ongoing digitisation activities. There are exceptions to this rule (discussed in Section 5 and Appendix A) where an entirely open licensing model is not appropriate, and a consistent and quantifiable approach is needed to handle these. Natural History Museum data policies should be embedded within the systems and protocols for managing internal datasets and serving them up to the public domain.

This document defines a digital data access and citation policy which will apply to DCP activities, is promoted for wider use within Science, and contributes to informing strategy across the museum as a whole.

Key items (detailed later in the document(s))

1. *The default position is to publish collection and scientific research data sets under the Creative Commons Zero (CC0) waiver. Attribution and citation of those datasets are encouraged in line with community best practice.*
2. *The default position is to publish digital media assets under the Creative Commons BY Attribution license (CC BY 4.0).*
3. *Exceptions to the 'open by default' policy should be handled by a clear, efficient oversight and decision support structure, with particular regard to protecting the NHM's commercial interests and research competitiveness, and third party rights.*
4. *Data publication, licensing and citation policy are made easily and universally accessible on the Data Portal. Preferred citation formats are provided that incorporate stable URIs/DOIs.*
5. *Software developed as part of DCP activities is published under an appropriate open source license.*
6. *Citation and publication of datasets derived from Museum data assets are encouraged via a clear statement of expectations.*

The Data Protection Policy *sets out the Natural History Museum's commitment to data protection legislation (meaning any UK Data Protection Act in force from time to time, and the General Data Protection Regulation) and good practice in handling personal data. It is intended primarily as an internal document, which may be made available publicly, complementing the external privacy notice on the Museum website (www.nhm.ac.uk/privacy-notice). Data protection legislation provides a framework for organisations to ensure that personal data is handled properly and gives individuals important rights in relation to their personal information, including being able to find out what is held about them.*

The Staff and Volunteer Data Protection Policy determines how their personal data is handled at NHM.

The Open Data Exceptions Policy is a long and detailed document. We include selected excerpts here. The policy *translates the principle of openness in relation to all* *information – data, documents, images – created and held in the course of Museum activities, whether scientific, public engagement or administrative, into practice by*

- *connecting the policy to data protection, freedom of information, third party rights and official marking*
- *setting out the exceptions to release, and how to interrogate whether they are applicable*
- *signposting to key personnel who can assist with queries.*

This policy documentation extends the work in the DCP Digital Licensing and Citation Policy and the Science Data Embargo Policy to cover all the exceptions initially listed therein plus an additional exception covering legal/regulatory compliance.

The Freedom of Information Act (FOI) and the Environmental Information Regulations (EIR) take precedence over Museum exception decisions.

When a request for information is received under these pieces of legislation, any internal exceptions will be taken into account when considering the response. However, only specific FOI exemptions or EIR exceptions can be used to withhold information from release. Even if it is proposed to withhold requested information under an FOI exemption/EIR exception, the 'public interest test' also has to be applied before a final decision is made. There is rarely an absolute guarantee that information that the Museum might wish to withhold will not be released. Nevertheless, it should also be noted that some FOI exemptions can remain in place for a long period of time e.g. section 40 for personal data can apply for the lifetime of the individual concerned.

Exceptions to the 'open by default' principle

i. Institutional profile and brand value
ii. Commercial value
iii. Research competitiveness
iv. Third party rights
v. Sensitive collection information
vi. Donor or funder conditions
vii. Confidential information
viii. Legal or regulatory compliance

Consideration of all possible exceptions must be taken into account before information is enabled for public access.

Declarations of exceptions, accompanied by a clear justification, must be sent to the Information Manager for inclusion in the Open Data Exceptions register.

We include below the description of Sensitive collection information

To be used for information relating to sensitive or valuable collections which is not suitable for sharing in the public domain, or would compromise collection or personnel security

For example: information on rare or endangered species, sensitive human remains and valuation data

Before information is enabled for public access, the information owner must consider its sensitivity. A list of questions to facilitate decision-making can be found in the Appendix.

Generally, data of this nature should be actively excluded from release into the public domain (e.g., removed from datasets before publication). Sensitive datasets shared with collaborators should be accompanied by a confidentiality agreement that prohibits re-sharing with any parties not explicitly permitted as part of the agreement.

If the information owner is unsure, the Registry or an appropriate senior member of the department (e.g., Head of Collections) should be consulted.

Potential FOI exemptions: section 31 – law enforcement; section 38 – endangering health and safety; section 39 – environmental information; section 40 – data protection

What happens when an exception is declared?

If data is considered to fall within any of these exceptions, one or more of four basic actions will occur:

1. *No release (this would be the default for sensitive/confidential information)*
2. *Time-delayed release (i.e. embargo)*
3. *Release with conditions*
    a. *Release under licence from the Museum*
    b. *Release with third party conditions*
    c. *Release according to legal/regulatory conditions*
4. *Release with all rights reserved*

When an exception is declared, it will be recorded in the Open Data Exceptions register by the Information Manager, and elsewhere as appropriate, for example within a system.

Once an embargo period has expired, the exception should remain marked unless it is no longer applicable, in which case it should be cleared.

How to manage embargoes can be tricky. Experience shows that it can stop short any open access policy, at least in universities, where scientists manage their own data, with little interest from the side of the management. The NHM has a Managing Science Data Embargoes Policy, which tackles these issues. Interestingly, with 16 pages, this is the longest text in the comprehensive set of NHM data policies. It opens with the following:

Open access principles will be applicable to the vast majority of digital assets generated by the Museum's on-going digitisation activities. However, there are exceptions to this rule where an entirely open access and licensing model is not appropriate, and a consistent and quantifiable approach will be needed to handle these. One such exception is designed to protect the Museum's research competitiveness, primarily by the application of appropriate embargoes to delay public release of the data.

This document describes the Natural History Museum's policy and processes for applying of science data embargoes for the Museum. It should be regarded as an exception to the 'Open by Default' rule, as defined by the overarching DCP data policy framework.

Key items

1. *Review body: Science Strategy Group, supplemented by the Registrar and an IP specialist and with reference to specialist advisory groups as appropriate, to act as the primary authority for reviewing applications to invoke an embargo exception.*
2. *Duration: Initial applications for embargos to have the option of an initial 12 month or 36 month duration from the delivery of a research-ready dataset.*
3. *Application process: Initial embargo applications to be approved without challenge unless there are exceptional circumstances. If no application is received, it will be assumed that the data can be released under default open terms (subject to any additional exceptions).*
4. *Assessment: Applications to be based on an assessment of the 'value' that the embargo will bring to the Museum, balanced with our obligations to release data under EU, UK government and funder expectations.*
5. *Renewals: Applications to renew embargos to require increasingly compelling arguments for the 'value' that an extension would deliver.*
6. *Management: Full details of current and historic embargos to*

We will not cite nor further comment the details in this elaborated document but encourage all readers to study it carefully.

DMP:

Conclusion:

NHM has a comprehensive set of data policies that cover the most important issues.

The fact that the Embargoes Policy is the most detailed one is particularly intriguing. The length of this document indicates that this might be an important factor also elsewhere that hampers open access but has rarely been put on paper. This topic should be investigated more.

Naturalis Biodiversity Center

Partner abbreviation: Naturalis

Size of collections: About 40 million specimens

Records on own site: 5,300,000 (13% of specimens)

GBIF records: 8,352,000 (158% of records on own site)

Licences: Not mentioned

Data policy: 'Research Data Management Policy.pdf'
Policy name: Research data management
Authors: Hannco Bakker & Rutger Vos
Accepted by BO: 28 November 2017
Version: 1.0

At Naturalis we describe, understand, and explain biodiversity. We do this by providing an open environment to science and by sharing our knowledge as widely as possible. We believe that *openness in science is of utmost importance to facilitate the dissemination of knowledge for the advancement of science and society.*

Current developments in scientific research practices surrounding data require institutional support. The rise of a societal demand for scientific openness, the pursuit of scientific excellence, and the internal need to use our organization's resources (human power and infrastructure) sensibly and efficiently, raise the need for a well-considered internal policy on data management. In addition, funding agencies such as NWO and EU have introduced increasingly explicit requirements for data management and research data stewardship.

To address these developments, this document establishes our policy on research data management, with the aim to more clearly evaluate and control current and future data storage usage and data management practices. The policy reflects the requirements established by funding agencies (with specific attention to NWO and EU), adheres to the FAIR guiding principles for scientific data management and stewardship, dovetails with the European Open Science Cloud and its preparations for the GO FAIR initiative, and adopts KNAW's recommendations on responsible research data management and the prevention of scientific misconduct . The policy also aligns with our institutional mission and the National Plan for Open Science.

DMP:

Naturalis has a number of DMP templates for use by specific projects. Accepted projects require a DMP. Project proposals contain a data management section.

Process of publishing:

Research data is published after a research project is completed, either on institutional infrastructure or at an external data repository that complies with the regulations and requirements in this policy. When published on an external data repository a backup will be placed on the institutional infrastructure.

A "Data Steward" is responsible for adherence to FAIR data principles. This role is also responsible of other tasks, such as dissemination, training, support, and advice on data management practices and regulations; as well as *for providing methods and procedures for controlled access to research data.*

"Researchers" are responsible in cooperation with the Data Steward for storing relevant and reusable research data … and making it accessible during and/or after the project.

"Head of Laboratory" is responsible in cooperation with the Data Steward for managing lab data and making it accessible.

"Collection- and Library- Managers" are responsible for the management of research data and metadata coming from or linked to the collection in the CMS (collection management system).

Annotations, feedback: Not mentioned.

Exceptions to openness policy:

Exceptions will be made for sensitive data, e.g. locations of red list species, patient data, intellectual property, etc. In the case of data ownership within Naturalis (as per 2.10), the ICT department facilitates and supports data storage on institutional infrastructure indefinitely.

Research data is published post-project subsequent to any embargos as agreed with involved funding agencies and project partners. When no such parties are involved, research data is published after the embargo period specified in the applicable DMP.

Access to research data post-project is subsequent to agreements with involved funding agencies and project partners. When no such partners are involved, research data is accessible as per the terms described in the applicable DMP.

Terms of use: Not mentioned

Conclusions:

Naturalis has a well-formulated data policy. It makes clear separation of the data policy and DMPs. The data policy is clear on who is responsible in the process of data publishing. Interestingly, there seems to be more data available on GBIF than on institutional site.

Royal Botanic Garden Kew

Partner abbreviation: RBGK

Size of collections: About 8.5 million specimens.

Records on own site: 1,830,000 (20% of specimens)

GBIF records: 924,700 (50% of own records)

Licences: CC-BY

Data policy: Briefly in the framework document[18]

Embrace digital and data

We will migrate our customers to online channels, building common ICT platforms and sharing data to improve and simplify the customer experience. All digital services will be designed according to Government Digital Service best practice and the Digital Service Standard.

List of policies does not include explicit data policy[19].

During 2015/16 RBG Kew participated in the Defra Open Data Programme, develop an open data policy and started to plan for implementation[20].

[18] https://www.kew.org/sites/default/files/Kew_FD_June18_Final%20-%2026%20July%202018_0.pdf

[19] https://www.kew.org/about-our-organisation/our-policies

[20] https://www.kew.org/sites/default/files/annual-report-accounts-1516.pdf

DMP:

Process of publishing:

Annotation, feedback:

Exceptions to openness policy: Not very specific, ref [21].

Terms of use: Well documented [22].

Conclusion: Kew has many policies. Data policy seems to be under development and is not yet publicly available.

University of Helsinki, Finnish Museum of Natural History (LUOMUS)

Partner abbreviation:	UH
Size of collections:	About 13 million specimens.

There are about 80 different databases at LUOMUS, which are being consolidated to the collection management system KOTKA.

Records on own site:	1,617,000 (12% of specimens)
GBIF records:	609,000 (38% of own records)
Licences:	CC-BY 4.0

Data policy:

> *The data policy[23] of the Finnish Biodiversity Information Facility is guided by the same acts and obligations as the Finnish Museum of Natural History.*
>
> *In its data policy, FinBIF complies with policies, strategies and programmes related to the open access to digital data agreed upon both at national and international levels including the EU, specifically the policy to increase the openness of data by the Finnish government.*
>
> *The public nature of the museum's operations is governed by the Universities Act and the Act on the Openness of Government Activities, which has a direct impact on FinBIF's data policy as well. The data policies of the museum and, therefore, FinBIF are in line with the data policy of the University of Helsinki, the parent organisation of the museum.*

[21] See https://www.kew.org/about-our-organisation/our-policies/access-to-information

[22] https://www.kew.org/science/data-and-resources/science-terms-and-conditions

[23] https://laji.fi/en/about/960

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

Other regulations that have an effect on FinBIF's data policy are the following acts and regulations concerned with online services and government as well as related to the creation, use, distribution, utilisation and storage of digital data: the Nature Conservation Act, the Administrative Procedure Act, the Act on Information Management Governance in Public Administration, the Personal Data Act, the Act on Electronic Services and Communication in the Public Sector, the Act on Electronic Signatures, the Archives Act, the Act on Collecting and Preserving Cultural Material and copyright regulations.

The data policy of the Finnish Biodiversity Information Facility also takes into consideration the policies in Finnish legislation and strategy influenced by the EU's PSI (Public Sector Information) and the INSPIRE (Infrastructure for Spatial Information in Europe) directives. In accordance with the act (421/2009) and decree (725/2009) on spatial data infrastructure, the museum has duties as regards the shared use of location data related to species distribution and biogeographical regions. Datasets must be maintained and shared in accordance with stipulations set in the act and the implementing rules of the INSPIRE directive.

Data policy scope: The data policy covers all data managed by FinBIF. Datasets are received for management, storage and distribution on the basis of agreements.

DMP: Only briefly translated to English, see [24].

Process of publishing:

Not specified for LUOMUS in its DMP, but for FinBIF at large data publication is described in general terms [25].

Annotation, feedback: Records can be publicly commented by registered users of FinBIF.

Exceptions to openness policy:

Openness can be restricted in several ways, the most common of which is making location data less specific. Access to data dependent on time or related to a location other than the observation location, such as overwintering or nesting sites, can be restricted with good reason.

In addition, FinBIF may have datasets whose openness and availability may be restricted at the request of data owners or managers as agreed. The rights related to sharing this type of data are also defined in accordance with the user access licenses stated above on a case-by-case basis.

Data will be stored in the FinBIF data warehouse in the form they were received in from the data owners. The restrictions stated above will only come into effect when data is transferred from the data warehouse to the public service of the Laji.fi website and made available for use. In accordance with the goals promoting the open use of FinBIF data, authorities and parties classified as authorities have unlimited access to the data for official duties.

Data disclosed voluntarily through FinBIF for use by authorities is protected under section 24, subsection 1, item 16, and section 26, subsection 1, item 2 of the Act on the Openness of Government Activities[26]. In practice, data disclosed for official use by authorities cannot be further disclosed without the consent of the data owner. This is ensured by mutual agreements

[24] https://laji.fi/en/about/864
[25] https://laji.fi/en/about/845
[26] https://www.finlex.fi/en/laki/kaannokset/1999/en19990621

> *between FinBIF and the authorities when agreeing on how authorities will have access to all FinBIF data.*
>
> *Data openness must not violate data protection. Restrictions to the openness of data and their justification are described in the metadata. Open data is made available to the academic community, public administration, organisations, businesses and private citizens for free use through the Finnish Biodiversity Information Facility portal (Laji.fi), managed by the Finnish Museum of National History. Other public distribution channels for open data may be established.*
>
> *The open and public nature of FinBIF data is promoted by sharing data and metadata from the data warehouse for publication in the Global Biodiversity Information Facility portal and other similar services.*

Terms of data use[25]:

> *As a rule, digital data managed by the Finnish Biodiversity Information Facility is public and free to use, or open data. As a rule, data is shared in compliance with the principles set by the Creative Commons licenses. Selection and implementation of CC licenses is conducted in accordance with agreements made with data owners.*

Conclusion:

> LUOMUS has a well-defined data policy. Its implementation is progressing but does not yet cover all data at LUOMUS. It is not easy to find out of how many digital specimen records actually exist at LUOMUS, and digitization percentage of 12% probably is an underestimation. The DMP is not a separate document but embedded in the data policy. It is not stated how the data from the ongoing mass-digitisation efforts is published. Publishing towards GBIF is not done according to any schedule, and currently covers only 37% of FinBIF specimens.

Discussion

The difference of a data policy and a data management plan seems not to be clear. One or the other is typically missing, for each institution, or we still have not received them.

DMP is a tool that helps to implement the data policy. Bakker & Vos (2017) define a DMP as follows:

> "*A data management plan (DMP) outlines the questions and approaches surrounding data management and, when applicable, within the scope of a funded research project. A DMP could be either a project-based DMP or a personal DMP. Where a project-based DMP is typically drafted and submitted to the funding agency early on in the project and may be periodically evaluated and/or updated, a personal DMP is typically drafted for research data outside of projects. In a DMP, the main types of data, their formats, and their volumes are enumerated. A DMP describes all regulations and practices regarding sharing, archiving and reuse of data. A DMP will consider the best practices as established within the applicable research domains, meet the requirements of the funding agency, and/or comply with the Naturalis DMP templates.*"

This definition highlights the need for DMP templates. Logically, there is a need to define one or more DMP templates for DiSSCo institutions. The European Commission has a DMP template for H2020 projects[27], which has been followed for ICEDIG DMP[28] (D6.1). Could this template be recommended for all DiSSCo institutions? Templates for data policy might be useful, since such document does not exist for all institutions.

NHM, London, has one digital collections data policy and four specific policies for details on sensitive data, embargoes and personal data protection. This illustrates the complexity of data policies issues. It is a tall order to expect that all the 115 DiSSCo institutions go through the same exercise.

In many cases written policies do not exist in the collection institutions, but there are national regulations that are being followed. There may be advantages of flexibility in avoiding a written data policy. Nevertheless, DiSSCo cannot move forward with data integration without some clarity of the rules. But it could be argued that some flexibility be built into a DMP. The DMP should also include a mechanism to change the DMP. Most DiSSCo institutions probably fall in between no written policies and a whole suite of them. This will be explored in the remaining for of ICEDIG Task 6.1.

The embargoes policy in particular seems to be complicated. Although not much has been written about it, embargoing has been one of the most contentious issues with scientists. We have not fully resolved how it should be policed.

Currently, it is also complicated to deal with data that is sensitive for biological, cultural and political reasons. Despite of their best efforts, staff working in country A in continent B do not know what is 'sensitive' in country C in continent D. There have actually been cases of accidental release of sensitive data this way. Can this be taken care of through some kind of clearing-house, and how might that work? National portals for citizen observations already have lists of species for which data must be hidden or generalised. If these lists were published in a standardised machine-readable form and gathered in one place such as DiSSCo portal, filtering sensitive data internationally could perhaps be achieved.

Nevertheless, as a fundamental principle, coordinates of specimens (where known) should be published to DiSSCo. There are two ways to deal with coordinates of e.g., sensitive species. First is not to publish them to DiSSCo but that is restrictive for those that might have a legitimate reason for knowing them. The second way is to encrypt the coordinates and give keys to those authorised to know the coordinates. This has been nicely implemented at APM (see the Section of institutional policies above).

The Nagoya agreement is trying to sort this out for data use in the case of DNA sequences. Everything is forbidden by default. Then special rules are applied to release some data for use in some circumstances. This is how firewalls on Internet work. Does it work like this for biodiversity data which are needed for open science? Nevertheless, species names, collectors and locations of specimens are not genetic resources and should not be affected by the Nagoya Protocol, although some countries have different opinions about that.

[27] http://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf

[28] https://cetaf.org/sites/default/files/d6.1_icedig_deliverable_report_d6.1.pdf

Our findings have exposed many questions that the DiSSCo Research Infrastructure must solve. One important conclusion is emerging, though. Data policies and DMPs can be complicated. We need some coherence in data policies across members of DiSSCo and that these need to be agreed collectively. It would be futile for DiSSCo to expect that all its individual members independently come up with their own data policies, which would then be somehow put in coherent action through their individuals DMPs, for implementation of the emerging new DiSSCo Research Infrastructure.

We have seen this before. GBIF, initially, did go through this path, allowing free-form data sharing agreements. This helped to mobilise data fast, but after ten years of operation, the situation became unwieldy. In 2014 GBIF embraced standard, machine readable licensing of data. GBIF took a real leap-of-faith and DiSSCo should not underestimate the risk they took. Massive loss of data was predicted by some people when they did this, but by and large fears were not realised and some data providers have returned afterwards. The situation with GBIF was much simpler, though, than what DiSSCo will face. DiSSCo will not only build a portal of copies of data which has been voluntarily contributed by its participants. DiSSCo will manage and curate data on behalf its participants, which are subject to European and national regulations.

Public sector data cannot be restricted arbitrarily, but exceptions must be based on objective criteria in legislation and justified. However, although the data in principle is open, there is no clear requirement to provide on-line service, except for some spatial data covered by the INSPIRE Directive. It requires for habitats and species implementation of data and on-line service, but only on an aggregated basis, not raw data.

Conclusion

The future data flows and storage infrastructure for DiSSCo must fit into the policy framework of the supporting entities. This framework is defined at a supranational level by the European Union and by the Convention of Biological Diversity, especially the Nagoya Protocol to which the European Union is party as well. It implies, among other components, the following policy guidelines:

According to article 6 of the Nagoya Protocol, access to genetic resources for their utilisation shall be subject to the prior informed consent of the Party providing such resources that is the country of origin of such resources or a Party that has acquired the genetic resources in accordance with the Convention, unless otherwise determined by that Party. Access to genetic resources must therefore be legally authorised by the country of origin of such resources.

The access to and the reuse of public sector information is subject to Directive 2013/37/EU. As a general principle, this Directive states that Member States shall ensure that documents to which the Directive applies shall be re-usable for commercial or non-commercial purposes in accordance with the conditions set out in Chapters III and IV. For documents in which libraries, including university libraries, museums and archives hold intellectual property rights, Member States shall ensure that, where the re-use of such documents is allowed, these documents shall be re-usable for commercial or non-commercial purposes in accordance with the conditions set out in Chapters III and IV. Biodiversity data in publicly funded collection institutions are „documents“ in the sense of this

Directive. Member States must ensure that they shall be re-usable for commercial and non-commercial purposes.

The Guidelines to the Rules of Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 apply these general principles for the sector of publicly funded research. The EU „wants to improve access to scientific information and to boost the benefits of public investment in research funded under Horizon 2020“ (European Commission 2017).

This legal framework is binding for every publicly funded collection institution, whether it is explicitly transformed into an institutional policy or not. It makes clear that biodiversity data collected in these institutions must be findable, accessible, interoperable and reusable by default. From our analysis of data policies from ICEDIG institutions, we conclude the following:

- Digital Specimen Objects (DSO) must be findable and accessible at MIDS-0 level. Data should be deposited in a trusted public repository, such as EUDAT, Zenodo, national system, or DiSSCo’s own public repository.
- As far as possible, projects must enable third parties to access, mine, exploit, reproduce and disseminate this data by using a copyright waiver such as CC0 or an open access licence such as CC-BY.
- Exceptions to openness policy must be stated clearly and strictly limited to reasons of national security, legal or regulatory compliance, sensitivity of collection information, and third party rights.

References to literature

Bakker H, Vos R (2017) Research Data Management Policy. Version 1.0, 9 pages. Naturalis Biodiversity Center, Leiden, The Netherlands.

Dillen M, Groom Q, Hardisty A, et al. (2019) Interoperability of collection management systems. ICEDIG Deliverable Report D4.4. 57 p. https://icedig.eu/sites/default/files/deliverable_d4.4_icedig_interoperability_with_collection_management_systems.pdf

European Commission (2017) Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2, 11 pages. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

GBIF (2015) New GBIF task force to speed up access to natural history collection data. Group of experts will help collections set priorities for digitizing more than 2 billion specimens. https://www.gbif.org/news/82381/new-gbif-task-force-to-speed-up-access-to-natural-history-collection-data

Hardisty A, et al. (2019) Provisional Data Management Plan for DiSSCo infrastructure. ICEDIG Deliverable report D6.6 (Version 0.8, Final draft) 80 p.

Kahn RE, Wilensky R (2006) A framework for distributed digital object services. International Journal on Digital Libraries 6(2): 115–123. https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf

Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, ..., Desmet P (2014) The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PloS one, 9(8), e102623.

Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, et al. (2014) Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. PLoS ONE 9(3): e89606. https://doi.org/10.1371/journal.pone.0089606

References to data policies reviewed

APM

”Terms of use” https://www.plantentuinmeise.be/en/informatie/Disclaimer

17 Databases https://www.plantentuinmeise.be/en/wetenschap/Databases

MNHN

Data https://www.mnhn.fr/en/collections/scientific-databases

Scientific data https://www.mnhn.fr/en/collections/scientific-databases/museum-national-histoire-naturelle-databases

Scientific database https://science.mnhn.fr/all/search

NATURALIS

https://github.com/naturalis/naturalis_data_api/blob/V2_master/LICENSE

Research Data Management Policy
https://docs.google.com/document/d/1K4Ve9zqWzqNiI63B8L7vzUuHfLhksbPQJzkRyz1xksI

NHM

http://nhm.openrepository.com/nhm/pages/about.html

https://www.ukri.org/funding/information-for-award-holders/data-policy/

http://www.nhm.ac.uk/about-us/privacy-notice.html

RBGK

https://www.kew.org/cookies-policy-0

Terms and conditions https://www.kew.org/terms-and-conditions

Data and resources terms and conditions https://www.kew.org/science/data-and-resources/science-terms-and-conditions

Data access https://www.kew.org/science/data-and-resources

https://www.kew.org/sites/default/files/annual-report-accounts-1516.pdf

UH (LUOMUS) /FinBIF

Terms of data usage and publishing https://laji.fi/en/about/845

Data management https://laji.fi/en/about/864

Data policy https://laji.fi/en/about/845

Collections https://kotka.luomus.fi/view?uri=http://tun.fi/HR.128 (requires login)

Databases on FinBIF https://laji.fi/en/observation/stats (Luomus and other institutions)

80 Databases of LUOMUS only https://laji.fi/en/observation/stats?collectionId=HR.128

UTARTU

DMP https://plutof.ut.ee/

Privacy and terms https://plutof.ut.ee/#/privacy-policy

https://unite.ut.ee/workbench.php