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 D6.6 / R / PU
Work package 6 / Lead: Naturalis
Delivery date M22
DOI: 10.5281/zenodo.3532937

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Provisional Data Management Plan for DiSSCo infrastructure

DELIVERABLE D6.6

(Version 1.0, Final)

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

ICEDIG.EU

DOCUMENT INFORMATION

Date and version no.	Author	Comments/Changes
21 June 2018, v0.1	Alex Hardisty, CU	Structuring and outlining of expected content.
26 June - 4 July 2018, v0.2	Alex Hardisty, CU	Adding notes, text, etc.
4 July – 3 Aug 2018, v0.3	Alex Hardisty, CU	Ditto. Added figures 1 and 2, tables 1 and 2 and definitions of core structural concepts to help us set the scope of the DMP, the major data classes/types to be handled and the main protected characteristics.
4 Aug – 22 Oct, v.04	Alex Hardisty	Added authenticity as a protected characteristic.
23 Oct – 20 Nov, v0.5	Alex Hardisty	Continued drafting; especially in sections on FAIR, levels of digitisation and definition/use of terminology. Settlement on adoption of Digital Object Architecture as the basis for DiSSCo. Version for AHM2 discussions.
11 Dec, v0.6	Alex Hardisty	New version, considering outcomes from AHM2 discussions and further progress in developing data management principles for DiSSCo.
14 Jan – 17 Apr 2019, v0.7	Alex Hardisty	Further work, aligning also with content of draft v4 of MS35 milestone report; adding further details under lifecycle, FAIR and identification sections, etc. Introduction of numbered "Data Management Principles" to highlight important requirements of DiSSCo data management. Improvements throughout, especially in data types, roles, and identification sections, and addition of editor's notes to highlight outstanding points and works needed. Distributed to project participants at Easter 2019 for review and discussion at AHM3, June 2019.
18 Apr – 26 Sep, v0.8	Alex Hardisty	Further work following Easter 2019 technical review to resolve all comments received and to prepare a final draft for review by the consortium. Section on position of DiSSCo in the global landscape has been removed as no longer relevant to the DMP directly.
31 Oct 2019, v1.0	Alex Hardisty	Completion following review by consortium beneficiaries and incorporation of final remarks.

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

Table of Contents

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

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

List of Tables

List of Figures

List of Data Flow Diagrams

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

ICEDIG.EU

Blank page

Preface to the present version

Status
The present provisional Data Management Plan has been prepared as deliverable D6.6 of task 6.2 of the EU funded ICEDIG (777483) project, with the intention that it should be taken up by relevant tasks of the DiSSCo Prepare project. In this context, 'taken up' means both that it should act as guidance to the further work to be done by DiSSCo Prepare and that it should also be further developed and improved by that project. The expectation is that DiSSCo subtask 6.4.3 will bring this Data Management Plan (DMP) to maturity for final approval and acceptance by the DiSSCo General Assembly as the basis for DiSSCo data management.

For further study
Several topics have been left for further study. They are marked as such in the text. This is because either they are not essential to the content of this provisional Data Management Plan or because further actions and decisions are needed before principles and requirements can be established.

Hidden text
The Microsoft WORD version of the present document contains hidden text acting as a guide to the required structure and content of the DMP; also including "editor's notes" to guide future authors. Hidden text can be revealed by using the WORD function to show/hide paragraph marks and other hidden formatting symbols (Ctrl-Shift-*).

The hidden text is not a normative part of the DMP. Its aim is to serve as reminders and to aid further improvement.

Acknowledgements
The following individuals are gratefully acknowledged as having contributed to the present document through their technical review of it and their participation in technical discussions:

Wouter Addink, Donat Agosti, Mathias Dillen, Willi Egloff, Pierre-Yves Gagnier, Quentin Groom, Anton Güntsch, Donald Hobern, Sharif Islam, Paul Kersey, Dimitris Koureas, Abraham Nieva, Niels Raes, Tim Robertson, Hannu Saarenmaa, Holger Thüs, Claus Weiland, Noortje Wijkamp.

Blank page

Executive summary

DiSSCo, the "Distributed System of Scientific Collections, is a pan-European Research Infrastructure mobilising, unifying and delivering bio- and geo-diversity digital information to scientific communities and beyond as a single digital virtual collection. With approximately 1.5 billion objects to be digitised, bringing natural science collections to the information age is expected to result in 100 petabytes of new data over the next two decades, used on average by 5,000 – 15,000 unique users every day. The DiSSCo Data Management Plan (DMP) is a living document reflecting the active data management planning and stewardship philosophy of DiSSCo, with focus on achieving maximum accessibility and reusability of data according to core principles of 'findable, accessible, interoperable and reusable' (FAIR), longevity of data and data preservation, community curation, linking to third-party information and reproducible science.

The DiSSCo DMP offers unified data management principles for data providers, data managers and users, and guidance to engineers and programmers on technical standards and best practices. It applies to data management activities (production and acquisition, curation, publishing, processing and use) of the geographically distributed collection-holding organisations (the DiSSCo Facilities) and to all DiSSCo Hub activities.

DiSSCo adopts Digital Object Architecture (DOA) as its foundation because of its future-proof flexibility over long timescales in the face of technological change, and because DOA has been shown to offer adherence to the FAIR principles as an integral characteristic, providing mechanisms inherently that directly address the specific principles to be promoted. In DOA the core concept is the 'digital object'.

Digitisation is the process of making data about physical objects digitally available, and the output of that process is Digital Specimens and Digital Collections. Digital Specimens and Digital Collections are specific types of 'digital objects', which are the fundamental entities to be the subject of data management in DiSSCo. Each instance of a digital object collects and organizes all the core information about the physical things it represents. These identified objects are amenable to processing and to transport from one system to another, making DOA a powerful yet simple extension of the existing Internet. A link must be maintained by the Digital Specimen to the physical specimen it represents. This link is the identifier of the physical specimen. These Digital Specimen objects are the principal data that DiSSCo manages.

Each Digital Specimen or other digital object instance handled by the DiSSCo infrastructure must be unambiguously, universally and persistently identified by an identifier (Natural Science Identifier, NSId) which shall be assigned when the object is first created. Each DiSSCo Facility shall be responsible for creating (minting) and managing their own NSIds in accordance with the DiSSCo policy for NSIds, and for registering their own Digital Specimens with the DiSSCo Hub infrastructure. Resolution of an NSId shall always return the current version of an object's content, as well as any interpretations and annotations associated with it.

The principle object types in DiSSCo (Digital Specimens, Digital Collections) are treated as mutable objects with access control and object history (provenance), meaning that they can be updated as new knowledge becomes available. Provenance data must be generated and preserved by all operations acting upon DiSSCo data objects. Timestamped records of change (provenance data) allow reconstruction of a specific 'version' of a digital object at a date and time in the past.

Information about Digital Specimens and Digital Collections must be published and managed as part of the European Collection Objects Index. DiSSCo Facilities are encouraged to publish the fullest available digital data about their individual specimens and collections at the earliest opportunity, aiming as best practice to achieve at least MIDS level 2 for Digital Specimens and MICS level 2 for Digital Collections information.

Several characteristics, such as centrality, accuracy and authenticity of the Digital Specimen, protection of data, preservation of readability, traceability/provenance, and annotation history are essential for developing long-term community trust in DiSSCo. They are the protected characteristics of DiSSCo that must be protected throughout the DiSSCo lifetime. Thus, all design decisions (technical, procedural, organisational, etc.) must be assessed for their effect on the protected characteristics. Such decisions and changes must not destroy or lessen the protected characteristics.

Summary of DiSSCo data management principles

The DiSSCo data management principles can be summarised as the following:

DMpr 1: All design decisions (technical, procedural, organisational, etc.) must be assessed for their effect on the protected characteristics. Such decisions and changes must not destroy or lessen the protected characteristics. (page 12)

DMpr 2: Digitisation is the process of making physical objects digitally available, and the output of that process is Digital Specimens and Digital Collections. (page 14)

DMpr 3: DiSSCo treats data as digital objects, each having a persistent identifier (pid) and a type. (page 15)

DMpr 4: A link must be maintained by the Digital Specimen to the physical specimen it represents. This link is the identifier of the physical specimen. (page 15)

DMpr 5: DiSSCo Facilities are encouraged to publish the fullest available digital data about their individual specimens and collections at the earliest opportunity, aiming as best practice to achieve at least MIDS level 2 for Digital Specimens and MICS level 2 for Digital Collections information. (page 17)

DMpr 6: The principal digital object types to be managed by DiSSCo are: Collection and DigitalSpecimen. Other object types include: StorageContainer, SpecimenCategory, Presentation, Gathering, Annotation, Interpretation, and Provenance. (page 19)

DMpr 7: Management of the DiSSCo digital object types shall be based on the general principles of the DiSSCo Data Management Plan, supplemented where necessary with additional management requirements for specific object types. (page 20)

DMpr 8: Provenance data must be generated and preserved by all operations acting upon DiSSCo data objects. (page 21)

DMpr 9: DiSSCo digital objects must be serialized as JSON [ECMA-404], as specified in section 4 and appendix A of the Digital Object Interface Protocol specification (DOIP) [DOIP 2.0 2018]. (page 22)

DMpr 10: Information about Digital Specimens and Digital Collections must be published and managed as part of the European Collection Objects Index. (page 26)

DMpr 11: Each Digital Specimen or other digital object instance handled by the DiSSCo infrastructure must be unambiguously, universally and persistently identified by a Natural Science Identifier (NSId), which shall be assigned when the object is first created. (page 35)

DMpr 12: Each DiSSCo Facility shall be responsible for creating (minting) and managing their own NSIds in accordance with the DiSSCo policy for NSIds, and for registering their own Digital Specimens with the DiSSCo Hub infrastructure. (page 36)

DMpr 13: Resolution of an NSId shall always return the current version of an object’s content, as well as any interpretations and annotations associated with it. (page 36)

DMpr 14: The principle object types in DiSSCo (Digital Specimens, Digital Collections) are treated as mutable objects with access control and object history (provenance). (page 37)

DMpr 15: Timestamped records of change (provenance data) must be kept, allowing reconstruction of a specific ‘version’ of a digital object at a date and time in the past. (page 38)

1 Introduction

1.1 Background

> *"Data is most reusable where data types are simple and easy to describe, and when the community is organized and collaborative."*
>
> Quote from "FAIR in practice"[1]

> *"FAIR is not limited to its four constituent elements: it must also comprise openness, the accessibility of data, long-term stewardship, and other relevant features."*
>
> Quote from "Turning FAIR data into reality"[2]

DiSSCo, the "Distributed System of Scientific Collections, is a pan-European Research Infrastructure mobilising, unifying and delivering bio- and geo-diversity digital information to scientific communities and beyond. With more than 115 institutions across 21 countries, DiSSCo combines uniform access to carefully curated hard evidence in natural science collections (i.e., data about physical specimens) in a community curation space that is open to contributions from multiple sources of expertise. Contributions to data will no longer be limited by the capacities of collection-holding institutions and will widen the base for developing new scientific information. Indexing the data from diverse natural science collections makes it available as one virtual collection. This combination of uniform access to hard evidence as one virtual collection and community curation and improvement of data differentiates DiSSCo substantially from other natural science data infrastructures. DiSSCo appeals at a national/local level as well as on a European level not only providing an index of all European collection objects but also by providing direct links to the local data holders (the collection-holding institutions), common mechanisms for arranging loans and visits, an annotations system, access to local sources of taxonomic and other scientific expertise, and wider public access to hidden collection holdings.

In addition to serving as the future framework for interpreting, validating and improving specimen data, DiSSCo connects the historical and contemporary collection data with the c. 500 million pages of literature in which species are described. It also connects to data emerging from new techniques. These new data include DNA barcodes, whole genome sequences, proteomics and metabolomics data, chemical data, trait data, and imaging data (Computer-assisted Tomography (CT), Synchrotron, etc.). DiSSCo will deliver the diagnostic information required for novel approaches and technologies for accelerated field identification of species, contributing to the development of datasets at adequate scale to support regular environmental monitoring, trend analysis and future prediction. The human discoverability and accessibility of the DiSSCo knowledge base will enable researchers across disciplines to tap into a previously inaccessible pool of quality assured data, while the machine readability will enable users to automatically digest these datasets into analytical workflows and tools. This focus on the quality (i.e., fitness-for-purpose) of data, along with the reproducibility and validation capabilities of science is enabled by persistent and robust linkages of digital information back to the physical objects that data represents.

With approximately 1.5 billion objects to be digitised, bringing natural science collections to the information age is expected to result in 100 petabytes of new data over the next two decades, used on average by 5,000 – 15,000 unique users every day. This requires new skills, clear policies and robust procedures to create, work with and manage large digital datasets over their entire research data lifecycle, including their long-term storage and preservation and open access. Such processes and procedures must match and be derived from the latest thinking in open science and data management, as epitomised by the two quotations at the

---

[1] FAIR in practice – Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles, https://doi.org/10.5281/zenodo.1245568.

[2] Turning FAIR data into reality – Interim report from the European Commission Expert Group on FAIR data, https://doi.org/10.5281/zenodo.1285272.

beginning of this section concerned with realising the core principles of 'findable, accessible, interoperable and reusable' (FAIR).

The present document, the DiSSCo Data Management Plan (DiSSCo DMP, or DMP for short) is a living document reflecting the active data management planning and stewardship philosophy adopted by DiSSCo, with focus on achieving maximum openness and reusability of data, longevity of data and data preservation, and reproducible science. Of course, no scientific community is static, with needs, capabilities and capacities evolving. As such, this DMP will be revised as necessary to reflect current DiSSCo data management policy, associated decisions and procedural changes.

1.2 Scope

The DiSSCo DMP offers unified data management principles for data providers, data managers and users, and guidance to engineers and programmers on technical standards and best practices to be applied. The scope of the DMP is shown by the red-bounded area over the key building blocks of DiSSCo (Figure 1).

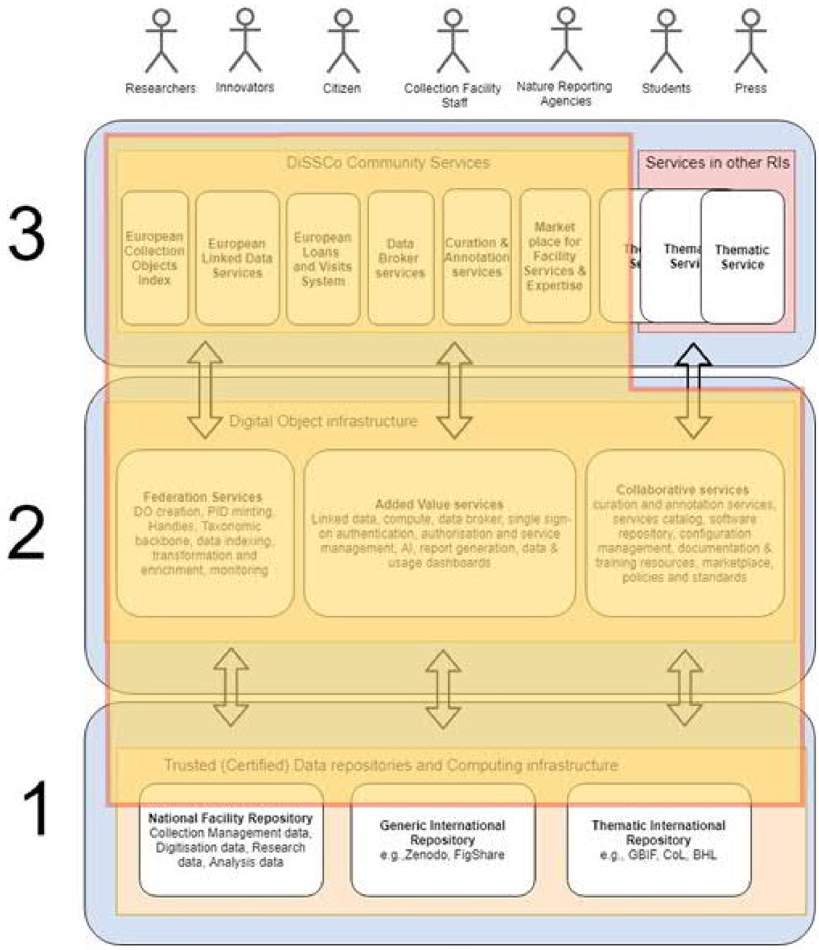


Figure 1: Three building blocks of the DiSSCo architecture

The three building blocks of DiSSCo are:

1. Trusted data repositories and computing infrastructure: The geographically distributed collection-holding organisations (i.e., natural science/history collections) that deliver data and expertise into the DiSSCo infrastructure, and which can be accessed by users via the DiSSCo infrastructure are referred to as DiSSCo Facilities. DiSSCo will link data that is provided by the DiSSCo Facilities in trusted repositories. These can include local institutional repositories and national repositories, as well as global thematic repositories such as the Global Biodiversity Information Facility (GBIF). All data that can be linked to collection objects (specimens) are in scope. DiSSCo provides access to computing infrastructure to exploit these data.

2. Digital object infrastructure: The infrastructure of integrating services, information technology components (hardware and software), human resources, organisational activities, governance, financial and legal arrangements that collectively have the effect of unifying natural science collections through a holistic approach towards digitisation of and access to the data bound up in those collections is referred to as the 'DiSSCo Hub'. DiSSCo Hub implements the digital object infrastructure, which includes tools for federation and linkage as well as services to support annotation and enrichment of the data by the scientific community. It draws upon common services provided by third-parties e.g., global and European Open Science Cloud services for authentication and authorization.
3. DiSSCo Community Services: The infrastructure provides community services to discover, consume and interact with the federated Digital Specimen data. Part of these services are provided in collaboration with other research infrastructures to enable innovative services for multi-disciplinary science.

The DiSSCo DMP applies to all DiSSCo Hub activities and to those activities of DiSSCo Facilities where they interface and interact with DiSSCo Hub. The DMP applies to all projects in the DiSSCo programme.

1.3 Intended audience and guidance on reading

The present document is relevant to three different audiences according to their respective roles within the DiSSCo programme. These audiences, which can be found in DiSSCo Facilities and DiSSCo Hub are:

1. Engineers, technologists, programmers, operator/administrators concerned with developing and operating the DiSSCo data infrastructure (it's IT systems, software, services, processes, procedures, etc.);
2. Digital collection managers, digitisation project managers, digitisation leads, digital curators concerned with digitisation, related curation and publishing activities and management of digitised data; and,
3. General collection managers and heads of curation responsible for coordination and management of collection activities, those with legal/regulatory compliance responsibilities and scientists/researchers working with data.

It is beyond the scope of the present document to provide full technical background to each data management principle. Enough technical background is given to provide context and it is assumed that readers, each according to their role will have sufficient relevant technical knowledge.

The executive summary (page ix) and the summary of DiSSCo data management principles (page x) are relevant for all readers.

Sections 1 – 3 are explanation and context for the DMP. These sections give the background, scope, digital object basis of data management, and the lifecycle of DiSSCo data. Together with sections 14 (glossary) and 15 (references) they are relevant for all readers. Section 3 specifically will be of interest to general collection managers and heads of curation to gain an overview of the main activities of data management.

Sections 4 – 13 specify the main data management principles and technical requirements of data management. They are relevant to both IT infrastructure personnel and digital collections/programmes technical personnel.

Section 4 (page 12) establishes the essential characteristics of DiSSCo infrastructure to be protected throughout its entire lifetime, and from which many of the subsequent data management principles derive. This section is relevant for systems and other engineers with responsibility for design and implementation of the overall infrastructure. It is relevant also to digital collection managers, etc. (category 2 above) responsible for DiSSCo policy implementation where implementation of those policies is through data management principles and the IT infrastructure of DiSSCo Facilities.

Section 5, DiSSCo data summary deals with the purpose of data collection and management, and the principal categories of data to be handled by DiSSCo and their formats. It introduces the notion of minimum levels of information about specimens and collections.

Section 6 deals with the main data management principles and requirements necessary to make DiSSCo data 'FAIR' (Findable, Accessible, Interoperable, Reusable). It is the prelude for sections 7 – 13 dealing in detail with specific technical elements necessary to make data management work.

1.4 Conventions used in the present document

1.4.1 Numbered statements of primary data management principles

This DMP covers complex topics of data management and stewardship with multiple interrelated requirements that must be implemented and complied with. To give context to the main principles of the DiSSCo Data Management Plan to be applied, explanatory paragraphs form a substantial part of the plan. Where these explanations lead to a principle or policy of data management to be applied throughout DiSSCo, this is summarised as a numbered and bolded statement of a primary data management principle, "DMpr", as in the following example:

DMpr n: A statement of the principle.

These DMpr can be found throughout the document and are collected together in the summary of DiSSCo data management principles (page x). They are not a substitute for reading the full text, where many details can be found.

1.4.2 Use of verbs

Four kinds of verb are used consistently throughout the present document: 'must/shall', 'should', 'may' and 'can', with meanings as follows:

'Must' (or occasionally, 'shall') indicates a mandatory requirement of this DMP, as in: "Each Digital Specimen or other digital object class handled by the DiSSCo infrastructure must be unambiguously, universally and persistently identified …". See section 7.1 for this use in context.

'Should' indicates a recommended practice, as in: "The policy should be machine-readable …". See section 3.6 for this use in context.

'May' indicates optionality, a choice of whether or not to do something, as in: "… data may be identified temporarily …". See section 7.4.1 for this use in context.

'Can' indicates the presence of an ability or that something is possible, as in: "Referenced third-party data can include biodiversity literature" (but doesn't have to). See section 5.2.3 for this use in context.

1.4.3 Use of capitalized terms and abbreviations

Capitalized terms (e.g., Digital Specimen) and abbreviations are used with the meanings defined for them in the glossary of terms and abbreviations. (section 14, page 44).

1.4.4 For further study, to be defined

The present document is provisional. Where the data management requirements are undefined, unclear, or yet to be determined, the phrases 'for further study' and 'to be defined' are used. This indicates that a later version of the present document may address the matter. See section 7.5.3, page 38 for this use in context.

2 Digital Object Architecture as the basis

Responsible data stewardship is essential to the modern scientific process. Journal publishers mandate the provision of supplementary materials, including data to published articles and the deposition of datasets, workflows and software in trusted repositories. Citation of samples used, including natural science specimens is on the increase. Research funding agencies require projects to maintain Data Management Plans. Universities, research organisations and other institutions organise their research data internally. Re-use of scholarly outputs, especially open access to research data is highly desirable. Against this background, the FAIR guiding principles (Findable, Accessible, Interoperable, Reusable) for managing scientific data [Wilkinson

2016, Mons 2017] aim to enhance the re-usability of research data. Emphasising machine actionability and infrastructure support – because scientists increasingly rely upon computational and data infrastructure capabilities and capacities to assist them with modern-day science – the principles are increasingly widely adopted across research communities. The FAIR principles are a key element contributing towards responsible data stewardship and thus they are an essential consideration for DiSSCo data management.

DiSSCo data management principles expressed in the present DMP aim to be technology agnostic to the greatest extent possible, expecting that over the DiSSCo lifetime specific data management and processing technologies can evolve and will be replaced. A framework for data management must accommodate this and one such framework is Digital Object Architecture (DOA) [Kahn 2006, Wittenburg 2019a]. DiSSCo adopts DOA as its foundation because of this future-proof flexibility and because DOA has been shown to offer adherence to the FAIR principles as an integral characteristic, providing mechanisms inherently that directly address the specific principles to be followed [Lannom 2020, Wittenburg 2019b].

Digital Object Architecture (DOA) is technology neutral, meaning there is considerable flexibility to decide how to implement data management and to change that over time. The core concept in DOA is ‘digital objects’ as the fundamental entities to be identified and manipulated by systems. Digital objects are open, editable, interactive items collecting all the core information about the thing they represent in one place [Kallinikos 2010]. In DiSSCo, these digital objects are principally Digital Specimens and Digital Collections. These identified objects are amenable to processing and to transport from one system to another, making DOA a very powerful yet simple extension of the existing Internet Architecture.

Persistent identifiers (PID) are the mechanism for identifying digital entities (including digital objects, datasets, workflows, software programs, journal articles, and more) involved in and produced by modern-day research. In the world of open data and open science this ability to uniquely and unambiguously identify such entities is essential to citation in scholarly outputs to support claims made and to aid reproducibility. The abilities to create meaningful links between entities based on PIDs and to record provenance back to the data producers increases the value of those entities to research and gives credit to those producing them. PIDs are an integral element of DOA that contribute towards the DiSSCo characteristic of ‘FAIRness’ (4.3 below), an essential characteristic of the infrastructure that is protected throughout the DiSSCo lifetime. PIDs (section 7) play a prominent role in the DiSSCo infrastructure, being used to identify everything from Digital Specimens and Digital Collections, through the transactions (such as loans and visits, annotations and interpretations) associated with those specimens and collections, to the people and organisations involved. As far as possible, DiSSCo follows best practices in relation to identification and citation using PIDs as set out, for example by the environmental sciences research community in Europe [ENVRI 2017].

3 Lifecycle of DiSSCo data

3.1 Origin of the data

The origin of DiSSCo data is digitisation lines/factories, as illustrated in Figure 2.

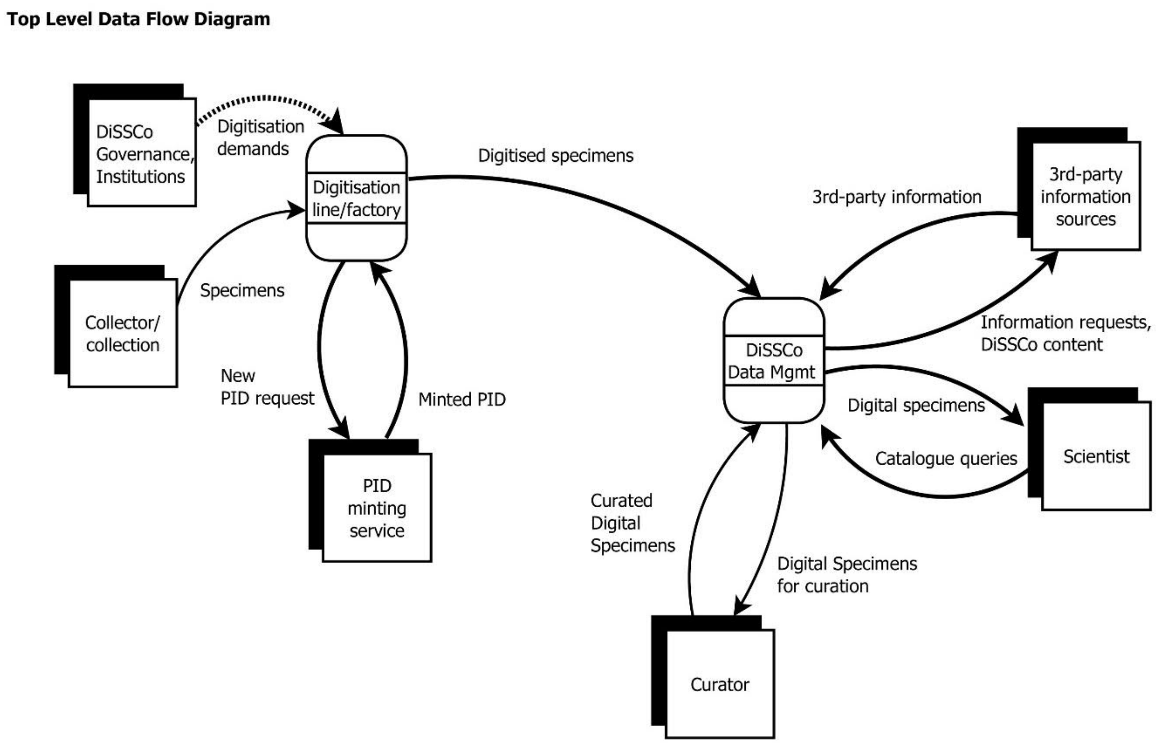


Figure 2: Top-level data flow diagram for DiSSCo
See Appendix C (page 53) for explanation of symbols (section C.1.2) and further levels of detail.
Key: PID = Persistent Identifier; Mgmt = Management.

Digitisation lines/factories are part of the digitisation programmes of DiSSCo Facilities and can be operated by DiSSCo Facilities themselves (when digitisation is carried out 'in-house') or can be external (i.e., a sub-contract or funded consortium arrangement). Data produced by digitisation has a lifecycle.

3.2 Digitisation activities and major lifecycle phases

Data flow diagrams in Appendix C (Figures C.2 – C.7) illustrate typical component activities around digitisation of specimens/collections and subsequent use of the digital data, including:

a. Pre-digitisation curation[3];
b. Imaging station(s) setup;
c. Imaging;
d. [Specimen] conservation;
e. Image processing;
f. Image archiving;
g. Optical character recognition;
h. Manual data entry and correction;
i. Data transcription;
j. Data quality control and,
k. Data publishing (local, to DiSSCo users, externally)[4]
l. Use, including experiments and analyses
m. Annotation.

Each component activity belongs to one of five major activities (Figure 3) corresponding to lifecycle phases of DiSSCo data (Figure 4):

<u>Major activity</u>
Acquiring data
Storing and preserving data
Making the data publicly available
Providing services for further data processing
Using the data to derive new data products and analyses

<u>DiSSCo data lifecycle phase</u>
Data Acquisition
Data Curation
Data Publishing

[3] Transport from collection to local imaging or digitization factory can also occur as part of this activity. Assigning an institutional identifier also takes place during this activity, if needed.
[4] During data publishing a persistent identifier can be created (minted) to uniquely and unambiguously identify the data.

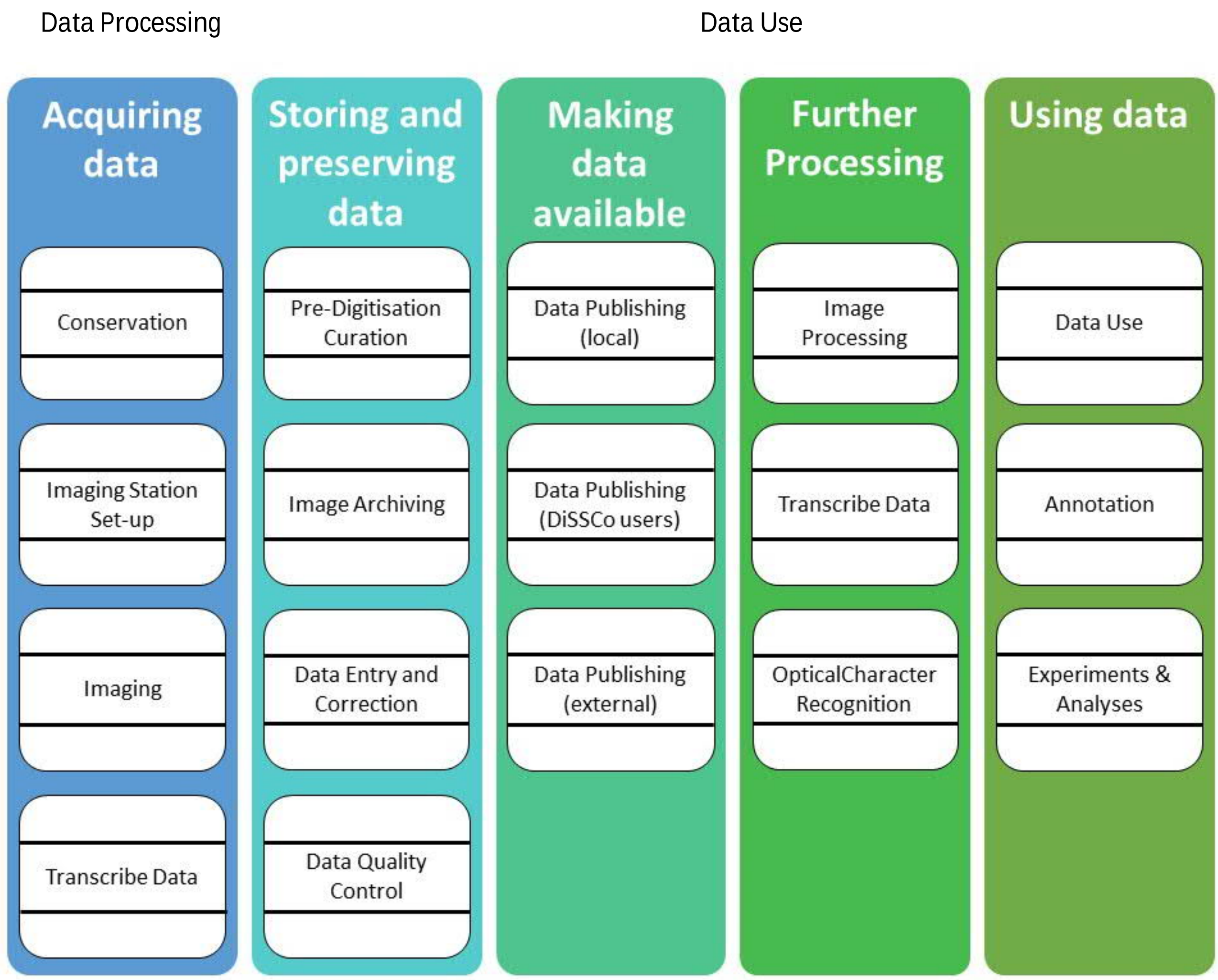


Figure 3: Relating data lifecycle phases to DiSSCo digitisation and data management

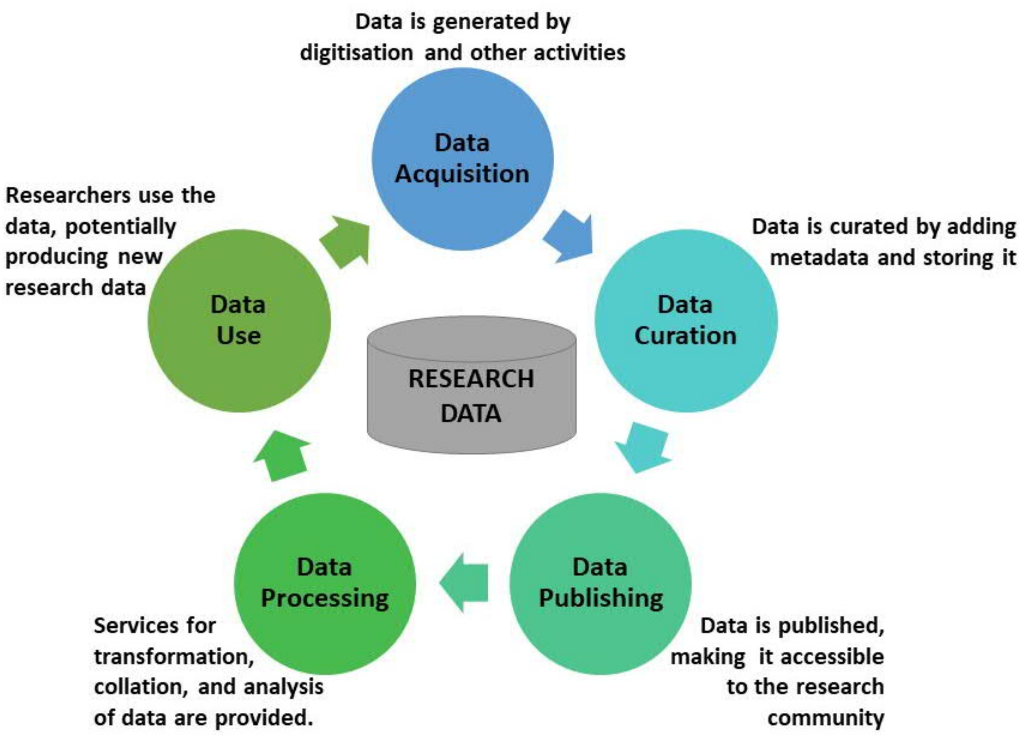


Figure 4: Simplified lifecycle of DiSSCo data

All activities, data management principles, applications, services and software tools of the DiSSCo infrastructure are designed and implemented to support this DiSSCo data lifecycle.

The data lifecycle begins with the acquisition of data through the principal activity of digitisation of physical specimens by a network of dedicated digitisation lines/factories. Digital specimen data is registered and curated within a data store(s) belonging to the DiSSCo infrastructure, either located in one of the DiSSCo Facilities or located as part of the DiSSCo Hub. Curated data is published to DiSSCo users and parties external to the infrastructure, as well as directly to other services within DiSSCo. This sequence results in a natural partitioning into data acquisition, data curation and data publishing. In addition, DiSSCo provides services for further processing of data that can produce new data to be stored within the infrastructure. Finally, the broader research community uses DiSSCo data and can design experiments and analyses acting on the published data that produce results (new data), which in turn can be passed back into DiSSCo for curation, publishing and processing; thus, restarting the lifecycle. Each phase is explained further below (3.4 – 3.8).

Throughout the data lifecycle, information technology (IT) support, including development and operations and often involving knowledge of biodiversity informatics and/or natural history collections is a core contributor for successful digitisation and data management. In addition to traditional management of physical collections, positive management of digitisation and data is essential to ensure timely, efficient and cost-effective throughput and stewardship in compliance with policies, procedures and quality standards. This demands clear allocation of roles to resources involved in digitisation and data management.

3.3 Roles within lifecycle phases

Within each lifecycle phase, data management responsibilities are assigned to different roles, summarised in Table 1 and described in sections 3.4 – 3.8 below. These active roles are performed by the various actors in the DiSSCo community and can be allocated in variable combinations, as reflected by individual job descriptions in different collection-holding and other stakeholder organisations.

Table 1: Responsibilities/accountabilities of major roles in DiSSCo

Lifecycle phase	Community Role	Sub-roles (Note 1)	Data management responsibility
Acquisition	Digitisation (mass): Working with the physical processes of digitising specimens (capturing information, transcription, imaging) as a project-based activity for mass digitisation of collections. On-demand digitisation is a subset of this role.	Collection curator	Together with coordination role, agrees the specimens and collections to be digitised.
		Technician, Collection technician	Retrieves, transports and prepares specimens or containers from collection storage to be processed for digitisation, cleaning specimens to be digitised, minting and attaching identifiers, returning to storage, etc.
		Digitisation operator, Digitiser (Note 2)	Operates digitisation equipment to digitise specimens (imaging, capturing information, transcription)
	Digitisation (regular study): Working with collections (visits, loans) for scientific purposes where studying specimens extracts and records data.	Scientist Researcher	Capturing information, transcription, data entry, log book as part of a scientific process that includes digitisation.
Curation	Curator: Those responsible for curating collections and information, either entirely physical collections, digital collections or hybrid physical/digital collections. (Note 3)	Data curator, Data manager, Data registrar, Data steward, Data guardian, Digital curator	Takes care of data resulting from digitisation processes e.g., checking, cleaning and improving data i.e., quality control, assigning identifiers, depositing in databases/image stores, long-term archival.
Publishing	Data specialist: Working specifically with the data resulting from digitisation processes e.g., checking, cleaning and improving data i.e., quality control; processing data; making data available. (Note 3)	Data publisher (Note 4)	Makes data publicly available by following a data publishing procedure that leads to data becoming publicly available in a database and/or a scholarly data paper.
Processing		Data specialist	Further processing and checking of data and/or images, improving data, producing new data such as annotations or derivative images at lower resolution.
Use	Scientist/Researcher: Those exploiting collections and their physical and digital content.	Scientist, Researcher, Other users i.e., educator, trainer, policymaker, citizen, citizen scientist, commercial, etc.	Exploits collections and their physical and digital content. Uses data for specific purposes, perhaps generating new data in the process. Ensures compliance with Nagoya protocol.
All phases	IT development and operations: Concerned with developing and operating IT systems, software, etc. often involving a knowledge of biodiversity informatics and/or natural history collections.	Engineer, Technologist, Programmer, Administrator	Develops/operates IT systems, software, etc. often involving a knowledge of biodiversity informatics and/or natural history collections.
	Coordination/management: Coordination of various activities or things; more generally, management.	(General) Collection manager, Head of curation	Coordination. Manages/curates collection(s), with overall responsibility.
		Digital collection(s) manager, Digitisation project manager, Digitisation lead	Leads digitisation and related activities, with overall responsibility to ensure this is carried out in a timely, efficient and cost-effective manner in compliance with policies/procedures.

Note 1: Sub-role names are indicative alternatives for the same or similar roles. More than one (sub-)role can be covered by any single job description.

Note 2: The Digitisation operator/digitiser sub-role can be partially or completely performed by volunteers, as well as by professional employees of a collection-holding organisation, or by outsourcing to a suitably experienced commercial organisation. Transcription of label information and operation of herbarium digitisation lines are common examples.

Note 3: The Curator and Data Specialist roles have different behaviours and span several lifecycle phases. The Curator role looks after data whereas the Data Specialist role works with and processes data. The roles can overlap and one or both can include the quality control component.

Note 4: Sub-role of the Data curator/manager role.

3.4 Data acquisition

Data acquisition is concerned with the principal activity of digitising physical specimens by dedicated digitisation lines/factories. This involves prioritising the specimens and collections to be digitised; retrieving, transporting and preparing specimens or containers from collection storage to be processed for digitisation, cleaning specimens to be digitised, and returning to storage afterwards; and operating digitisation equipment to digitise specimens (capture information, transcription, imaging). Planning of data acquisition means committing and spending formidable resources, which requires full attention of senior management and the head of curation (i.e., coordination and management role). The work in data acquisition is carried out by the Digitisation role; typically, by technicians and digitisation operators. In some cases, these can be trained volunteers rather than professional employees. Transcription of label information and operation of herbarium digitisation lines are common examples where volunteer effort can be used.

For data acquisition, key strategies for effective data production and management include:

- Developing an optimal de/centralisation strategy whether to digitise in-house or out-source the functions;
- Developing comprehensive procurement requirements, procedures, and quality assurance of the deliveries;
- Minting identifiers and attaching tags to specimens in an efficient procedure. At the same time, entering bulk data from the collection cabinets, boxes, folders, drawers, unit trays, and jars, for multiple specimens;
- Automation and operation of imaging pipelines;
- Transcription and quality control of data, and ingesting it to relevant repositories;
- Providing stronger support (e.g., by means of computer tools) for earlier creation of rich data i.e., a strategy for more easily reaching higher levels of digitisation; and,
- Not only dealing with newly digitised specimens but also dealing with specimens and collections that have already been digitised at the time DiSSCo becomes operational. How to bring these specimens into the DiSSCo environment?

3.5 Data curation

Data curation is concerned with curating digital specimen and collection data. This involves caring for and improving the data resulting from digitisation processes in the data acquisition phase e.g., checking, cleaning and improving data i.e., quality control, assigning (persistent) identifiers, depositing in databases/image stores, etc. The work is carried out by the Curator role; typically, by data curators, data managers or similar with support as necessary from data specialists where specific processing and manipulating is required prior to or as part of the overall curation process.

Beyond acquisition, curation is the central part of the DiSSCo data lifecycle. Curation is the process to keep data FAIR for the long-term, involving a combination of IT infrastructure and services, processes and procedures, and organisational stability and governance. Insofar as OAIS (Open Archival Information System) principles are relevant to DiSSCo, the key requirements are enshrined in the present DMP. DiSSCo infrastructure must act as a certified, trusted infrastructure to keep data authentic and available over time, maintaining the history of the provenance of data and transactions on data.

Annotations (assertions made on or about published data) must be fed back to the curation process to close the loop for improving data quality and for enhancing data. Eventually, the flow of data between CMSs and ECOI is expected to become bidirectional with mechanisms to maintain synchronisation between the systems. However, the timescale is not presently indicated for this.

For data curation, key strategies for curation and stewardship of data include:

- Taking care that changes to the physical specimen or collection after digitisation, such as a new physical annotation to labels are also applied to the digital specimen or collection data as part of an update procedure; and,
- Long-term archiving of lossless 'master' image files (the Authentic Image of Record) and preservation of the readability and retrievability of such files (protected characteristic C5, see 4.5), including determining whether to store data locally, externally in the cloud, or in specialised repositories.

Data produced by digitisation is normally curated in the Collection Management Systems (CMS) of DiSSCo Facilities or may be curated by the DiSSCo "CMS-as-a-Service" (CMSaaS) service.

3.6 Data publishing

Data publishing is concerned with making curated data accessible to DiSSCo users and publicly to parties external to the DiSSCo infrastructure, as well as directly to other services within DiSSCo. The work is carried out by Data Publishers (a sub-role of the Curator role) following a data publishing procedure that leads to data becoming publicly available in a database and/or a scholarl data paper.

Data publishing is based on the DiSSCo open access policy (see 6.3.1) and must adhere to relevant quality control criteria, including relevant minimum information standards (e.g., MIDS) for the type of data being published (section 9). The data lands in one or more DiSSCo open-access repositories (see 6.3.4).

Key strategies for effective and timely publication include:

- Following a data publishing procedure that implements the DiSSCo open access policy automatically; ideally, an automated procedure (daily, weekly, monthly, etc.) so that data does not remain inaccessible unnecessarily;
- Maintaining a publishing log (provenance, section 11); and
- Digitally signing published data to guarantee its integrity (section 7.10.1).

CMS data must be published as Digital Specimen objects and Digital Collection objects by DiSSCo Facilities and must be managed as part of the ECOI by DiSSCo Hub.

3.7 Data processing

Data can be further processed after it has been curated and published. This is the concern of the data processing phase. DiSSCo provides a range of services (expected to grow over time) for transforming data from one form to another, for collating and aggregating data e.g., to produce data summaries, and for analysing data in various ways. The results of data processing can be new data (annotations by external users, for example) that itself must be curated within the infrastructure and subsequently published.

The work of data processing is carried out by the Data Specialist role, which can be combined with either or both the Curator and Scientist/Researcher roles.

Key strategies for adding value during the data processing phase can include:

- Determining whether to process data locally, in the cloud, or elsewhere;
- Serving data in a variety of useful representation formats e.g., JSON, RDF, XML, CSV, etc.;
- Publishing details of simple to use Application Programming Interfaces (API) that make it possible to develop new and innovative services for processing DiSSCo data;
- Link identification/building between objects in the DiSSCo Unified Knowledge Graph (see 6.4.4); and,
- Synchronisation between DiSSCo Hub and the Collection Management Systems (CMS) of DiSSCo Facilities.

3.8 Data use

In the final lifecycle phase, data use, the broader research community (represented by the Scientist/researcher role) can exploit the digital and physical collections for science and can design digital experiments and analyses acting on the published and processed data. These experiments produce results (new data) that in turn can be acquired by DiSSCo for further curation, publishing and processing, thus restarting the lifecycle.

Much of the data managed by DiSSCo requires interpretation and validation. Information can change as new knowledge becomes available. This presents challenges for research reproducibility from the perspective of tracking workflow and data integrity.

Key strategies for adding value during the data use phase can include:

- Serving data in a variety of useful representation formats e.g., JSON, RDF, XML, CSV, etc.;
- Publishing details of simple to use Application Programming Interfaces (API) that make it possible to develop new and innovative services that use DiSSCo data;
- Providing useful tools and services that help users to build and execute digital experiments based on DiSSCo data; and,
- Determining how to meet the requirements of the Nagoya protocol.

4 Protected characteristics

There are nine characteristics (Cn) of DiSSCo data management that are essential to protect throughout and ultimately beyond the lifetime of the DiSSCo data infrastructure. This lifetime is expected to be 25 – 30 years. These characteristics are essential for engendering community trust in the value, veracity and reliability of the data to be managed.

This means that proposals for design decisions and changes, technical, procedural and organisational) must be assessed for their effect on the protected characteristics. Ideally, all design decisions and changes must not destroy or lessen any of the protected characteristics and should aim to enhance one or more of the characteristics.

DMpr 1: All design decisions (technical, procedural, organisational, etc.) must be assessed for their effect on the protected characteristics. Such decisions and changes must not destroy or lessen the protected characteristics.

4.1 Centrality of the digital specimen (C1)

The digital surrogate of the physical specimen, as represented by a Digital Specimen (DS) object type is the central asset of interest in the DiSSCo infrastructure and is the design unit/concept that all other design decisions must respect.

4.2 Accuracy and authenticity of the digital specimen (C2)

The digital specimen, as represented by a Digital Specimen object type is the best available digital representation (surrogate) for a physical specimen in a natural science collection. The possibility of data re-use implies that digital specimen data can be adapted, remixed, transformed, and built upon for different purposes. It might not be obvious when this happens, so it is important that the accuracy, authenticity and meaning of the digital specimen data is clear and preserved throughout its lifetime. Sections 7.5.1 and 7.10 on authenticity and mutability of digital objects deal with the data management principles supporting this protected characteristic.

4.3 FAIRness (C3)

FAIRness is a characteristic exhibited by an infrastructure (component) and the data it manages when that infrastructure maintains compliance with the principles of FAIR (Findable, Accessible, Interoperable, and

Reusable. This characteristic must be protected throughout the DiSSCo lifetime. The principles of the present DMP are oriented around maintaining FAIRness and, as explained in section 2, this is substantially aided by adoption of Digital Object Architecture and treatment of DiSSCo data as digital objects. Many sections of the present document deals with elements of the data management principles supporting this protected characteristic but see section 6 specifically.

Achievement of FAIRness is demonstrated, for example by achieving a score (passing a threshold) in an assessment against an agreed set of maturity indicators.

4.4 Protection of data (C4)

Data held and managed by DiSSCo must be protected in accordance with legal regulations and community norms for the kind of data concerned. For example, data considered to be sensitive by the community, such as that revealing the geographical location of endangered species must be protected from unauthorised access. Data revealing personal details of individuals such as researchers and collectors must be protected in accordance with the EU's General Data Protection Regulation 2016/679. Section 6.3.1 on the DiSSCo policy on accessibility deals with the data management principles supporting this protected characteristic.

Note 1: Protection of data, as used here means protection in the legal sense of the term. Protection in the sense of protection against loss and calamities is covered in section 10.1, page 41.

4.5 Preserving readability and retrievability (C5)

Maintaining the ability to retrieve categorical reference images from 'deep archive' and preserving the readability of the image and other file formats over long periods of time is essential. This means maintaining the ability to read and use:

a) Lossless, archived TIFF files and the related JPEG files used for data sharing and data processing;
b) Files in other specific formats, even after those formats become obsolete.

It means (potentially) systematically replacing image files in image archives/repositories with newer files using up-to-date file formats.

Note 2: Whether to maintain an ability to read and process the RAW formats of cameras used during digitisation processes is a local decision.

4.6 Traceability (provenance) of specimens (C6)

Traceability (provenance) of specimens, their digitisation and change history, annotation and usage must be maintained consistently through the entire lifetime of the DiSSCo infrastructure. Storing and managing provenance information is a shared responsibility of the DiSSCo Facilities and DiSSCo Hub. Provenance shall be based on the W3C PROV framework [W3C PROV 2013], with provenance stored as part of the digital specimen itself. The DiSSCo Trace subsystem is how this is achieved, requiring the adoption of a standard provenance framework (W3C PROV) and recording library wherever provenance must be recorded[5].

4.7 Annotation history (C7)

and interpretations attached to specimens and collections are an important part of the scientific and historical record (provability) and must not be lost or altered.

[5] In the same way that a standard logging framework is made available for to all Java™ software applications. A DiSSCo provenance logging framework must be developed, based for example on ProvToolbox (http://lucmoreau.github.io/ProvToolbox/).

4.8 Determinability (status and trends) of digitisation (C8)

Information (statistics) about the volume and scope (description) of natural history collections and their state of digitisation is needed, both to show progress and to assist with prioritisation of digitisation activities. This information must be maintained consistently to ensure a common basis for comparison is maintained over time. The Collections Digitisation Dashboard (CDD) is the subsystem by which this is achieved.

4.9 Securability (C9)

Securability (authentication, authorization, accounting, auditing) of multiple levels of access for users according to their authority and permissions, including access to retrieve and/or modify sensitive information must be maintained over the DiSSCo lifetime; considering that potential users from come from many communities, not only the European research and education community. Section 10 deals with the data management principles supporting this protected characteristic.

5 DiSSCo data summary

5.1 Digitisation and its relation to DiSSCo objectives

The scientific vision and mission of DiSSCo includes mobilising and harmonising natural science collection data as a single European digital virtual collection available in human and machine-readable forms via the Internet. At the same time, it includes connecting that historical collection data with data emerging from new techniques including imaging, tissue banking, DNA barcoding, whole genome sequencing, and legacy literature digitisation.

At the heart of DiSSCo are 'Digital Collections' and 'Digital Specimens', acting as digital representations in computer systems for collections and specimens in the real world. These digital representations are specialisations by DiSSCo of the more general-purpose notion of 'digital objects'.

As surrogates for the physical specimens in collections, the digital specimen concept organises information about what the specimen is, where and when it was collected and by whom, and (through the inclusion of images) what it looks like. A digital specimen includes a pointer to where the physical specimen can be found, as well as containing the history of the specimen as it was collected, identified, assigned to a specific taxon and named. Digital specimens record changes to that understanding over time as annotations that can be recalled for inspection. This rich information can be extended to include references to relevant scientific literature, tissue samples, genetic sequences and trait data, agricultural, toxicology and ecosystem data and more. Relations such as ‘isDuplicateOf’, ‘isHolotypeOf’, ‘hasSequence’, etc. can be established between specimens. Through linkages at the classification (taxon) level, further information applying to the specimen concept (i.e., at the species level for biological specimens) can be accessed.

Digital Specimens can be thought of as a dynamic storage folder where traceable, directed links to the core information about the specimen can be gathered and organised in a single place. Information on the outside of the folder describes its contents. The process of creating Digital Specimens is known as digitisation and can involve multiple stages of work. Similarly, Digital Collections collate information about a collection of specimens.

DMpr 2: Digitisation is the process of making physical objects digitally available, and the output of that process is Digital Specimens and Digital Collections.

The scope of DiSSCo includes all kinds of natural science collections, including fossils, rocks and minerals, anthropological artefacts, preserved biological specimens (plants, seeds, animals, insects, etc.) and living biological collections. Digital Specimens are at the heart of an interconnected graph of diverse and dispersed data classes, equipping them for many research and teaching purposes that might not otherwise be possible.

DiSSCo acts as the core for delivering rich information about specimens in collections, including links to their identification and taxonomy, known and expected distribution, and their value in fields as diverse as medicine, toxicology, food security, environmental protection, and land-use planning. Making information about the objects in collections available in digital form (i.e., digitisation) changes the way scientists interact with collections, including changing the way they physically visit collections or borrow specimens. Not only does digitisation open the collections/specimens to wider groups of professional and amateur scientists, students, the general public, etc. but it also opens the way to new innovative uses of collections that hitherto might not have been possible prior to the era of 'big data'; for example, by allowing the application of artificial intelligence (machine learning) techniques to vast quantities of digital specimen images, accumulated genetic sequences or morphological information.

5.2 Types of data generated/acquired

DiSSCo accrues and manages data constituting digital collections and digital specimens from successively more comprehensive levels of digitisation (5.2.1), as well as other types of data relating to the DiSSCo sub-systems for managing and administering scientific use of collections and specimens (5.2.5). The principal data types (5.2.3) i.e., collection and specimen related data itself, the format standards by which such data is represented, and the persistent identifiers that uniquely identify units and subunits of collection and specimen data together provide the target for establishing DiSSCo data as 'findable, accessible, interoperable, and reusable' i.e., FAIR (section 6 below)[6]. The present DMP is principally concerned with collection and specimen data but the principles it establishes are relevant for all types of data handled by DiSSCo[7].

5.2.1 Digital objects

DiSSCo adopts 'Digital Object Architecture' (section 2 above) wherein digital objects can be thought of as the meaningful entities of an application domain that are exchanged between and processed in different information systems. Digital objects have content, such as data about specimens or collections and are defined by a type, such as Digital Specimen type or Digital Collection type. Each object is uniquely and ambiguously identified by a persistent identifier (pid). “Meaningful” in the above context implies that the content data of, for example a Digital Specimen object has value for and can be interpreted by a human and/or a machine.

DMpr 3: DiSSCo treats data as digital objects, each having a persistent identifier (pid) and a type.

A link must be maintained by the Digital Specimen (with its accompanying persistent identifier) to the physical specimen it represents. This link is via the identifier of the physical specimen, including any additional information (e.g., institution code, collection code) necessary to identify the physical specimen. This link must be maintained for the lifetime of the digital object. Note, Digital Specimens do not include asset management information such as specific storage location.

DMpr 4: A link must be maintained by the Digital Specimen to the physical specimen it represents. This link is the identifier of the physical specimen.

DiSSCo digital object types are defined by object type definitions in the openDS standard [openDS <date>] and comprise one or more (in combination) of the data categories outlined in sections 5.2.3 and 5.2.5 above. The formats of the five principle data categories, and thus for how the object types will be defined by openDS are given in the following subsections 5.3.2 – 5.3.7. The formats of the secondary data categories are for further study. Serialization and packaging of digital objects is discussed in section 5.3.8.

The basic FAIR principles of accessible, interoperable and reusable (see FAIRness protected characteristic C3, section 4.3 and section 6.4.1 must be respected by choice and use of standard data formats.

---

[6] For background see also Figure 6 in Turning FAIR data into reality – Interim report from the European Commission Expert Group on FAIR data, https://doi.org/10.5281/zenodo.1285272.

[7] Future editions of the DMP will deal more specifically with the other types of data.

5.2.2 Levels of digitisation and minimum information requirements

DiSSCo Facilities (i.e., collection-holding partners) generate and manage data resulting from the process of digitising collections of natural science specimens. Digital information can be created for collections as a whole (overviews), for sub-parts of collections (inventories of trays of insects or boxes of herbarium sheets, for example) and for individual specimens. The first two categories contribute towards providing coverage information about the holdings of an institution, in terms of scope and extent; whilst digitisation of individual specimens provides specific and precise details about each object curated in a collection.

At the level of individual specimens, digitisation can be characterised generally as three distinct stages with data being created at each stage:

1. Attaching an identifier to a physical specimen and creating a digital catalogue record;
2. Digitising information about the collection or specimen, including:
    a) Imaging the specimen and/or its labels;
    b) Extracting, processing and encoding information (from labels, images, etc.); and,
3. Enriching with supplementary information from third-party sources (links to literature, genetic sequences, etc.).

Depending on the features of a specific digitisation process and how that's organised, the amount of detail created at each discrete step can vary. How much detail is published openly, and when also varies and has historically been determined by policies of individual collection-holding organisations.

DiSSCo aims to mobilise, unify and deliver natural science (bio- and geo-diversity) information at the scale, form and precision required by scientific communities. Achieving the elements of this mission (highlighted as italics in the preceding sentence) is accomplished in part by harmonising policy into guidelines for DiSSCo Facilities about practical levels of digitisation to apply (Table 2 and Table 3 below), and by harmonising the information to be expected from each level of digitisation. This is specified by the Minimum Information standard for Digital Specimens (MIDS) and Minimum Information standard for Digital Collections (MICS) standards <add references to both, when available>[8].

Table 2: Levels of digitisation of specimens, according to Minimum Information about a Digital Specimen (MIDS) standard

MIDS level	Record extent	Purpose
0	Catalogue	A skeletal record making the association between an identifier of a physical specimen and its digital representation, allowing for unambiguous attachment of all other information.
1	Basic	A minimal record of specimen information, mainly deriving from labels in collection boxes, folders, drawers, jars etc., enabling similar discovery capabilities on-line as researchers and curators would have by more traditional means.
2	Regular	Key information fields from specimen labels that have been agreed over time as essential for most scientific purposes.
3	Extended	Other data present or known about the specimen, including links to third-party sources.

Note: No notion of completeness or of a full or complete record exists because new information is always valuable and can be added to an existing record at any time.

[8] At the time of preparing this provisional DMP, these standards are being prepared and written.

Table 3: Levels of digitisation of collections, according to Minimum Information about a Collection (MICS) standard

MICS level	Record extent	Purpose
1	Overview	Overview information about a collection and the organization holding it – the who and where of a collection and, broadly, its content and history.
2	Inventory	Key information fields describing the collection in its entirety – species-level (or tray/drawer/cabinet-level) information about the collection.

Harmonisation provides clarity about the minimum quantity and quality of information DiSSCo Facilities should be publishing to make collections and Digital Specimens useful for multiple purposes of teaching and learning, research, etc. Similarly, by harmonising a framework that clarifies what is meant by different levels of digitisation it becomes easier to consistently measure the extent of digitisation achieved (e.g., via a Collection Digitisation Dashboard[9]) and to set priorities for the remaining work.

As a general principle, DiSSCo Facilities are encouraged to publish the fullest available digital data about their collections and individual specimens at the earliest opportunity, expecting that such data is likely to become enriched and annotated over time. For collections, MICS Level 2 Inventory is the ideal standard to aim for; while for specimens, "what, where, when" i.e., MIDS Level 2 Regular is the minimum standard to aim for, with enrichment towards MIDS Level 3 Extended being the 'gold standard' to achieve. However, publishing information in accordance with lower MICS and MIDS levels is also acceptable as a precursor towards future continuous improvement of digitisation efforts. The information elements expected defined by the MIDS and MICS standards at any level of digitisation (MIDS or MICS level) should be the minimum amount of information to be published.

DMpr 5: DiSSCo Facilities are encouraged to publish the fullest available digital data about their individual specimens and collections at the earliest opportunity, aiming as best practice to achieve at least MIDS level 2 for Digital Specimens and MICS level 2 for Digital Collections information.

5.2.3 Principal categories of data associated with digital specimens and collections

In connection with digital specimens and collections, DiSSCo data management distinguishes several categories of data to be managed:

1. Specimen and collection data;
2. Annotations;
3. Interpretations;
4. Supplementary data (including third-party data); and,
5. Provenance data.

Specimen and collection data (also Content Data, Authoritative Data) is data about physical specimens and collections, such as images of those specimens/collections, information from specimen/collection labels (such as scientific name, location where collected, date collected, collector name, etc.), or measurements and other analyses of specimens. Specimen identifiers, including all earlier, historical or superseded identifiers are listed, as well as information derived indirectly by interpretation from other specimens or literature.

An essential characteristic of this data is that it is authoritative about a specimen or collection. That is, the information that this data represents has been determined by scientists and curators and it is they, as approved and authorised experts that hold authority to make changes to this data as knowledge and

[9] See van Egmond et al., (2019, March 31). Design of a Collection Digitisation Dashboard. ICEDIG Deliverable D2.3. doi: 10.5281/zenodo.2621055.

understanding about specimens and collections evolves. When clarity is needed, especially in the context of who can modify such data the term 'Authoritative Data' is used.

Annotations are assertions replicating the traditional written annotation of physical objects such as determination of the species or comments relating to label information. Based on external knowledge (which should be documented) they can be made on or about both specimens and collections. Automated text recognition results from automatically scanning and processing specimen labels can be considered as machine-made annotations until verified and structured. Annotations become Interpretations when processed and accepted by an authorised curator or other approved expert.

Interpretations are the application of expertise, based on facts at hand to define more precisely the exact meaning of ill-defined text describing specimens. Several kinds of interpretation are possible: such as verbatim transcription; translation, transformation e.g., of date time formats, mapping to controlled vocabularies and linking to authoritative sources. A specimen can have multiple interpretations associated with it concurrently e.g., when experts disagree.

Supplementary Data is additional data about a specimen, beyond the categories mentioned above that contributes to better understanding of and increased knowledge about the specimen. Supplementary data can be generated by specimen owners and/or by third-parties, and can include biodiversity literature references, DNA sequence data, trait data, acoustic recordings, or other information relating to specific specimens and collections. Supplementary data can be held outside of the DiSSCo infrastructure and referenced from DiSSCo.

Provenance Data provides a traceable record about data, its origins and the processing actions applied to it. DiSSCo provenance data records the history of Content Data, Supplementary Data, Annotations and Interpretations. Section 11 details with provenance in detail.

Putting together these high-level component data categories with levels of digitisation in common practice (Table 2 and Table 3 above) the principal object types to be managed by DiSSCo are shown in Table 4.

Table 4: Principal data categories and digital object types

	Collection overview (MICS level 1)	Collection inventory (MICS level 2)	Digital specimen (MIDS levels 0-3)
Content data	Information about the collection and its owner, and where it can be found.	Inventory describes collection in its entirety at storage unit level, including general classification of contents i.e., plant, animal, fossil, rock, etc. with presentation and preservation characteristics e.g., herbarium sheets, pinned insects, specimens in glass jars, etc. Collecting and sampling event data. <Align to TDWG CD, when available>	Data* describing attributes of an individual specimen, essential for most scientific purposes, including other identifiers, license, permission flags, quality control assertions.
Object type:	Collection	StorageContainer, SpecimenCategory, Presentation, Gathering	DigitalSpecimen
Annotations	Assertions made on or about the collection.	Assertions made on or about the collection.	Assertions made on or about the specimen.
Object type:	Annotation	Annotation	Annotation
Interpretations	Interpretations about what the collection represents.	Interpretations about what the collection represents.	Interpretations about what the specimen represents.
Object type:	Interpretation	Interpretation	Interpretation
Supplementary data	Links to external sources of related data.	Links to external sources of related data.	Links to external sources of related data.
Object type:	For further study. See 5.3.5.	For further study. See 5.3.5.	For further study. See 5.3.5.
Provenance	History of actions on the object i.e., provenance records, digitisation logs, usage logs, etc.		
Object type:	Provenance		

*Including images: 2D and 3D images, 3D models, stacked images, lossless original images, derived images

DMpr 6: The principal digital object types to be managed by DiSSCo are: Collection and DigitalSpecimen. Other object types include: StorageContainer, SpecimenCategory, Presentation, Gathering, Annotation, Interpretation, and Provenance.

5.2.4 Metadata in DiSSco and the use of standard vocabularies

Metadata is widely and generally understood to be additional data providing context around the principal data of value to which it applies. Metadata's function is to aid understanding of that data; such as what the data is, what it applies to, when and where that was created and by whom, the conditions under which it can be used, whether it has been modified, etc. In many cases, without metadata, the principal data may have no value at all. One typical example of where metadata is helpful is for image files. The images themselves are the data, whereas some specific characteristics about those files needs to be described, such as format (tiff, png, jpg, etc.), size and resolution, creation date and creator, etc. This body of information is often referred to as the metadata of the image/file. In addition, of course it is essential to know what the image is of and what it relates to. Such information is often obtained by linking the image in some way to the (meta)data describing it. However, distinctions between what is data and what is metadata can easily cause confusion because it depends upon how the data will be used at any given moment. For this reason, DiSSCo data management tries to reduce and remove distinctions between data and metadata as much as possible.

In DiSSCo, almost all digitised information about specimens and collections, whether directly about a specimen or a collection or about the context in which that exists are represented as attributes with values; for example, collector:personname, collectiondate:datetimestring, Country:countryname, institution:organisationname, etc.). DiSSCo does not differentiate such attribute:value pairs as being either data or metadata, thus allowing them to be treated as both, according to circumstance.

Nevertheless, attribute:value pairs and specifically the range of values attributes can take are not well harmonised. Usage practice varies widely and is inconsistent, even within single institutions and across communities. Sometimes, attributes are left with blank values, making it impossible to know what was intended. Is the information just missing or is it unknown? Did the data creator not know which value to use? Was it not part of the digitisation process to assign a value to the attribute at that time? Such open-endedness and the uncontrolled use of values lead to ambiguity in analysis and interpretation. In many cases it can render data unfit for use. It is thus helpful to standardise attribute definitions and specifically the potential range of values such attributes can take as a means of supporting and working towards more harmonised understanding and processing of such data in the future. This is an important added-value element of DiSSCo data management. The topic is covered in more detail in sections 6.2.5 and 6.4.3 below.

Of course, much harmonisation has already achieved with term definitions by DwC and ABCD, for example, and implementations of those in various data tools and portals. However, more consistent application of existing standards

5.2.5 Secondary categories of DiSSCo data

In addition to the principal categories of data above, multiple secondary categories of data are expected to be managed by DiSSCo or through interfaces with existing systems (e.g., for persons). These include:

- Central Index (Marketplace) of Expertise and Facilities;
- Collections Digitisation Dashboard (CDD) data;
- Details of researchers and collectors, including deceased ones;
- DiSSCo agreed vocabularies and ontologies;
- DiSSCo software, workflows, and associated documentation;
- Distinct defined operations on different object types;
- Geographic information related to collection sites;
- Images (of various kinds) and 3D models;
- Loans, visits, access requests, queries (i.e., European Loans and Visits (ELViS) transactions);
- Unified Collection Annotation System (UCAS) annotations;
- Usages of specimens, including On-demand Digitisation (ODD) requests and prioritisations;
- User authentication and authorization records.

Each of these data categories and subsystems to which they relate can (if necessary) have its own data management plan, which shall inherit general principles from the present document and be further detailed at the necessary level for managing that category of data.

DMpr 7: Management of the DiSSCo digital object types shall be based on the general principles of the DiSSCo Data Management Plan, supplemented where necessary with additional management requirements for specific object types.

5.3 Data category formats

5.3.1 Object type hierarchies

For indexing and relational purposes, the principle object types: Collection, StorageContainer, DigitalSpecimen, Annotation, Interpretation, Linkpacket and Provenance; and the secondary types (Presentation, SpecimenCategory, Organisation, Gathering) are organised into three hierarchy models, each of which has DigitalSpecimen at its root. These hierarchies are:

1. <u>Resource type hierarchy</u>: Specifies the different types of resource/information that can be part of or linked to a digital specimen, such as images and pointers to third-party data;
2. <u>Storage hierarchy</u>: Places the digital specimen in the context of the collection(s) to which it belongs, and the container(s) in which the physical specimen is stored (e.g., insect tray); and,
3. <u>Annotation and provenance hierarchy</u>: The hook on which to hang interpretations and annotations of the specimen and the provenance record/history of the specimen and actions applied to it.

5.3.2 Content Data format

There are a wide range of data and file formats for Content Data, of which JSON[10], DwC-A[11], ABCD[12], TIFF[13] and JPEG[14] are expected to be widely used by DiSSCo. More specialised Content Data, such as for 3-dimensional model representations and genetic data can make use of their own specialised data formats.

5.3.3 Annotations format

For further study. Annotation object type. The Web annotation data model (https://www.w3.org/TR/annotation-model/) with its specific JSON format for ease of creation and consumption of annotations should be considered. Also, the the applicability of the Web annotation vocabulary (https://www.w3.org/TR/annotation-vocab/) and the Web annotation protocol (https://www.w3.org/TR/annotation-protocol/). Also, the proposal by ICEDIG Task 5.2 for a data exchange standard to harmonize data generated by transcription or annotation platforms (MS28, Le Bras, Chagnoux and Dillen 2019, doi: 10.5281/zenodo. 2598413). A common approach towards the overall structure of annotation, interpretation and linkpacket objects is likely.

5.3.4 Interpretations data format

For further study. Interpretation object type. No standards presently exist for this. Something like the annotations format above is likely. A common approach towards the overall structure of annotation, interpretation and linkpacket objects is likely.

5.3.5 Link data format

For further study. LinkPacket or LinkEvent object type. See https://bit.ly/disscolinkbuilder. A common approach towards the overall structure of annotation, interpretation and linkpacket objects is likely.

5.3.6 Supplementary data format

Supplementary data formats/structures vary widely, according to the nature of the resource. The openDS specification [openDS <date>] supports a wide variety of data/object types, with provision for both known and unknown types.

5.3.7 Provenance format

Provenance data must comply with the W3C PROV Data Model (PROV-DM) [W3C PROV 2013] and must be generated and preserved by all operations acting upon DiSSCo data objects. Specifically, earlier, historical and superseded identifiers of physical specimens must be prominently retained.

DMpr 8: Provenance data must be generated and preserved by all operations acting upon DiSSCo data objects.

5.3.8 Serialization and data packaging

Serialization is the process of encoding data structures in a form that can be transmitted from one computer system to another such that decoding the received form allows the data structure to be faithfully reconstructed for further processing. CSV files, JSON and XML representations are examples of serialization approaches in common usage.

Data packaging facilitates the movement of data between systems (portability), including making data available to local systems without the need to maintain network connections to a central server/service. Data packages aggregate serialized data structures in a formal manner (as files in a folder hierarchy, for example).

---

10 https://en.wikipedia.org/wiki/JSON.
11 https://en.wikipedia.org/wiki/Darwin_Core_Archive.
12 https://en.wikipedia.org/wiki/ABCD_Schema.
13 https://en.wikipedia.org/wiki/TIFF.
14 https://en.wikipedia.org/wiki/JPEG.

Packages can be transferred between systems using well-known communication protocols. Darwin Core Archive is one example of a data packaging format widely used by the biodiversity science community.

The kinds of serialization and data packaging used depend on data type, purpose for which serialization/packaging is needed and the producing/consuming applications and services. The present data management plan does not cover all cases but establishes principles and requirements for the most important cases expected in DiSSCo. These are:

- Transfer of digital objects from one computer/software system to another;
- Bidirectional synchronisation of digital collection and/or digital specimen data between the CMS of a DiSSCo Facility and the European Collection Objects Index (ECOI) ;
- Retrieval of digital collection and/or digital specimen data by a scientist/researcher via the European Collection Objects Index (ECOI) to a local system for processing and use; and,
- Others, to be defined as necessary.

5.3.8.1 Transfer of digital objects from one system to another

Most DiSSCo services and applications will interact natively with one another using the Digital Object Interface Protocol (DOIP) [DOIP 2.0 2018] to transfer digital collection, digital specimen and other categories of data from one computer or software system to another. This is the most general case where serialization and packaging of DiSSCo digital objects is needed and the mandated serialization format is JSON as specified below. Packaging of multiple objects is dealt with in section 5.3.8.4.

Note 3: JSON (JavaScript Object Notation) [ECMA-404] is specified because it is the basis of typing and serialization in Digital Object Interface Protocol (DOIP), which is the key protocol for object exchange between systems in Digital Object Architecture (DOA), upon which DiSSCo is based. JSON is a lightweight format that is easy for humans to read and write, as well as being easy for machines to generate and process. In DOIP, a JSON segment acts as the description and wrapper of a digital object to be communicated, where such digital object consists of an identifier, a type, optional attribute:value pairs and optional data elements. The optional data elements can be encoded using any appropriate media type (https://en.wikipedia.org/wiki/Media_type), https://www.iana.org/assignments/media-types/media-types.xhtml).

DiSSCo digital objects must be serialized as JSON [ECMA-404], as specified in section 4 and appendix A of the Digital Object Interface Protocol specification (DOIP) [DOIP 2.0 2018]. The structure of the attributes and elements of the serialization segment representing a digital object must be as specified in appropriate standards or as in the respective sub-sections below (which may themselves refer to appropriate standards).

DMpr 9: DiSSCo digital objects must be serialized as JSON [ECMA-404], as specified in section 4 and appendix A of the Digital Object Interface Protocol specification (DOIP) [DOIP 2.0 2018].

For legacy and non-native applications, DiSSCo should support these and their data formats using appropriate format adapters, exporters and importers, converting wherever possible to the DiSSCo preferred formats.

5.3.8.2 Bidirectional synchronisation of collection and specimen data

Collection and specimen data must be serialised for synchronisation between the Collection Management Systems (CMS) of DiSSCo Facilities and the European Collection Objects Index (ECOI) in accordance with the requirements of openDS [openDS <date>].

For an interim period, CSV files packaged as a Darwin Core Archive (DwC-A) (e.g., using the Integrated Publishing Toolkit (IPT)) may be used for publishing towards the ECOI.

For an interim period, XML documents compliant with the ABCD or ABCD-EFG Schemas (e.g., as produced by BioCase Provider Software (BPS) may be used for publishing towards the ECOI.

Note 4: It is not expected that serialisations other than that specified by openDS will be used for updating CMSs to keep them synchronised to the ECOI.

5.3.8.3 Retrieval of collection and/or specimen data from ECOI

Collection and/or specimen data retrieved from the European Collection Objects Index (ECOI) for local or other processing must be serialised and packaged in accordance with the requirements of openDS [openDS <date>].

Adapter services may offer alternative serialisations such as: CSV files packaged as a Darwin Core Archive (DwC-A), RDF documents, XML documents compliant with the ABCD or ABCD-EFG Schemas.

5.3.8.4 Other data

Specific services (to be defined) can return subsets of available data customised to the requests of the user; for example, lists of collectors or plant names that can be used as look-up tables during large-scale transcription projects.

The format of such data must be defined on a service specific basis but in general, JSON should be used where possible as it is easy to generate and to read/process, both by humans and machines.

5.3.8.5 Packaging of multiple digital objects

For further study. Packaging of multiple objects serialized in accordance with the above sections needs further study.

5.3.9 Image file formats

Over the long-term DiSSCo aims to apply the recommendations of the International Image Interoperability Framework (IIIF, pronounced 'triple eye eff')[15] for access to digital images, particularly images from cultural heritage institutions, museums, libraries and archives, which includes natural science collections.

Tagged Image File Format (TIFF) (media type = image/tiff) should be used as the preferred format for high-resolution original images (Authentic Image of Record, Lossless Image) produced by digitisation processes.

JPEG File Interchange Format (JFIF) format, or simply, 'JPEG' (media type = image/jpeg) should be used for lower resolution or compressed images derived from original images.

5.4 Re-use of existing data

There is a wide range of pre-existing external data, such as biodiversity literature, genetic sequence and other molecular information, chemical composition and structure data, traits data, habitats data, alien and invasive species data, data about species of conservation concern, etc. that can be linked with specimens in collections. DiSSCo makes it possible to build links between such data and Digital Specimens to create a Unified Knowledge Graph (UKG) as explained in section 6.4.4.

Data from multiple external sources will be made available to DiSSCo users through the DiSSCo UKG, with the expectation that this will lead to wide re-use of existing data for multiple scientific purposes, both those known today and future purposes yet to be imagined.

5.5 Expected size of the data

With approximately 1.5 billion physical specimens in Europe to be digitised, bringing natural science collections to the information age is expected to result in petabytes of new data over the next decades. However, the numbers of links between digital specimens and other information can greatly exceed the number of objects themselves, perhaps by 3 – 5 times. This will continue to grow as digital specimens become

---

[15] https://iiif.io/.

more and more connected in the UKG to other pieces of information relating to the specimen. Appendix D contains estimates for the different data categories.

5.6 Data utility

There is a wide range of traditional and new user groups for DiSSCo data. Conventionally, this community includes all researchers engaged in discovering, describing and interpreting life on Earth, both past and present, as well as researchers studying the geological history of the planet.

European collections hold circa 80% of the 2 million species presently described and are at the forefront of efforts to describe what is estimated to be approximately 6 million new species that await discovery [Mora 2011]. DiSSCo collections also include extensive palaeontological and mineralogical collections, including concentrations of rock and ore samples, making them a valuable resource for the field of economic geology as well as for research focused on climate change and the origin of our solar system. Exceptional meteorite collections can be used to study the origins of our solar system and how meteorite impacts could affect our future. Ocean bottom samples are critical for studies of the ocean and ocean floor, including research looking at global change, climatic warming and marine pollution. Extensive gem collections are held.

Note 5: These extraordinary collections are physically distributed across institutions, but when taken together and represented as digital specimens and digital collections they present an unparalleled source of scientific evidence about the natural environment. They can be conceptually viewed along two fundamental time scales:

1. Deep Time: The dynamic history of change in the geology and life of our planet, spanning 4.56 billion years, including tectonic shifts in the continents, the rise and fall of major lineages of the tree of life, major shifts in species' geographic distributions, biomes, ecosystems and environmental signatures and trends.
2. The Anthropocene: The recent and accelerating changes in biodiversity and ecosystem functioning because of the impact of modern humans on land use and resource exploitation, including species extinction, shifts in the distribution and abundance of species, climate change, agricultural effects, and the emergence of new pests and diseases.

The unprecedented taxonomic, geographic, stratigraphic and historical coverage gathered within these collections, coupled with their increasing digital accessibility, is opening them up to entirely new user communities, and increasingly to the private sector/industry. These users are increasingly drawn to the time series and patterns represented within these collections, to make predictions about the sustainable exploitation of bio- and geo-diversity that inform practice and policy decisions. Table 5 provides some examples of typical DiSSCo data usages beyond academic/scholarship uses.

Table 5: Typical purposes for DiSSCo data usage

Environment	Agriculture	Health	Border control	Biobanking
• Urban planning • Environmental impact assessment • Deep-sea mining • Conservation planning & monitoring • Prospecting • Shifts in species geographic distributions and abundances • Biomes, ecosystems and environmental signatures and trends • Tectonics	• Species identification • Future domestication • Land use change • Industrial (insect) farming • Forestry • Agri-chemicals • emergence of new pests and diseases • Climate change, agricultural effects	• Pathogen identification • Medicine and food supplement verification • Pharmaceutical industry • Biotechnology	• Invasive species and pests • CITES protected species enforcement • Countering illegal wildlife trade identification • Shifts in species geographic distributions	Preserve genetic material (tissues & seeds) for: • Research • Government • Industry (medicine, biotech. & agriculture)
Education, virtual exhibitions, documentaries, citizen science, historians & artists.				

DiSSCo data is expected to be used on average by 5,000 – 15,000 unique users each day.

Since 2014, utilization of data about genetic resources has been controlled by the Nagoya protocol (see Note 10 in section 6.3.1 on policy for accessibility, and section 6.4.5 on legal access to data), which must be taken into account when designing access mechanisms.

6 FAIR

6.1 Setting the context

'FAIRness', meaning the attribute of data to be easily findable, accessible, interoperable and re-usable is a protected characteristic of the DiSSCo infrastructure (4.3 above). All sections of the present DMP contribute towards achieving and sustaining FAIRness, meaning that compliance with the procedures and rules set out herein must be maintained by DiSSCo and its contributing members throughout the lifetime of the infrastructure and throughout the lifecycle of identifiable data assets managed by the infrastructure.

'FAIRness' is also the central guideline of the legal framework that rules the DiSSCo infrastructure. Digital specimen data as well as digital collection data represent public sector information in the meaning of [Directive 2013/37/EU]. Furthermore, the Guidelines to the Rules on Open Access to Scientific Publications and Open Access in Research Data in Horizon 2020 [OA 2017] aim "to improve access to scientific information and to boost the benefits of public investment in research funded under Horizon 2020". This legal framework implies that natural science data collected in publicly funded institutions must be findable, accessible, interoperable and reusable by default. Exceptions must be justified by public interest concerns, such as risks of safety, conservation or security.

The following sub-sections 6.2 – 6.5 describe specific policy-driven requirements that must be implemented by technical means as data management procedures and rules. Statements are made for how DiSSCo implements each of the FAIR Guiding Principles [Wilkinson 2016, Mons 2017]. These statements are summarised in Appendix E as a statement of DiSSCo's implementation of the FAIR principles. Standards referred to here and throughout the remainder of the document are summarised in Appendix F. A DMP compliance checklist can be found in Appendix G.

6.2 Making data findable

6.2.1 Policy for findability

FAIR principle F4: (meta)data are registered or indexed in a searchable resource.

Information about Digital Specimens and Digital Collections, and their associated interpretations and annotations must be findable in publicly accessible and searchable index(es). Data contained in said index(es) must exclusively describe facts (identifiers, scientific name, status, etc.) and must not reveal any personal or otherwise protected data. There are legal and regulatory circumstances dictating that not all data is findable, accessible and reusable for all categories of users. See 6.3.1, Policy for accessibility.

DMpr 10: Information about Digital Specimens and Digital Collections must be published and managed as part of the European Collection Objects Index.

Data of each object type is indexed and must be searchable directly in the relevant DiSSCo index. These include: the European Collection Objects Index (ECOI) , the European Loans and Visits System (ELViS) , and the Unified Curation and Annotation System (UCAS) (and their relevant services) thus allowing relevant identifiers (handles) to be discovered when these are unknown.

Kernel information of each object must be sent to the relevant Local Handle Service servers during handle registration (6.2.6 below).

6.2.2 Data naming conventions

FAIR principle F1: (Meta)data are assigned a globally unique and persistent identifier.

A Handle (NSId, section 7) is issued to each object published in or by DiSSCo, allowing the object data to be found regardless of its location.

The principle data categories and digital object types are named as set out in Table 4 (page 19).

6.2.3 Metadata policy

FAIR principle F2: Data are described with rich metadata (defined by principle R1 below).

As explained in section 5.2.4, DiSSCo does not specifically distinguish between data and metadata, allowing all data fields to be used as metadata when this is needed for a specific purpose. As such, there are no policies relating to metadata only.

Digital specimen and digital collection data mainly take the form of 'named_attribute : value' pairs, with attribute names and definitions being sourced from appropriate standard vocabularies and schema where possible.

FAIR principle F3: Metadata clearly and explicitly include the identifier of the data it describes.

The NSId (Handle) is a top-level and mandatory field in the data of each DiSSCo object type, and is used wherever an object or reference to an object appears.

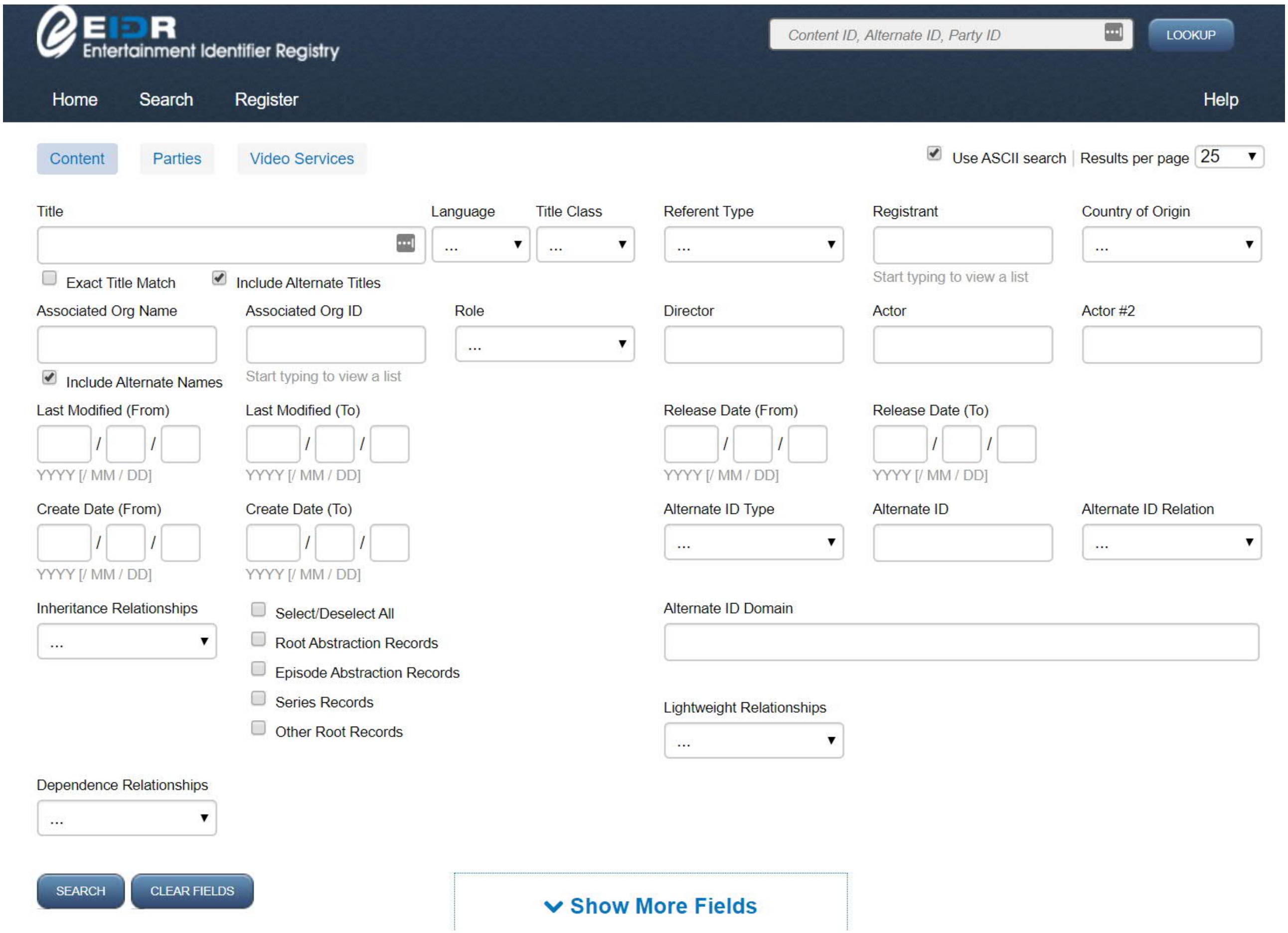
6.2.4 Data versioning

Data versioning is covered in section 7.4.1.

6.2.5 Keyword vocabularies

DiSSCo systems support full-text search over all data fields. Controlled vocabularies are used in conjunction with controlled field specifications to constrain the range of permissible field values.

For further study. Examples are needed of the different vocabularies for different sub-communities of DiSSCo. The picture below shows the EIDR search screen where some fields are free-form, and some are lists of controlled terms. ECOI needs something similar. Many of the controlled terms will derive from the object model, which still must be defined during DiSSCo Prepare.



6.2.6 Kernel information profiles

Specific PID Kernel Information profiles and object type definitions must be registered for the Digital Specimen object type and other object types in the well-known Kernel Information profile and Data Type registries.

For further study. It must still be decided which kernel information profile registry and data type registry to use. DiSSCo own, EOSC, other?

For further study: It must be determined whether the 15-attribute draft kernel information profile outlined in section 3 of the draft RDA Recommendation on Kernel Information profiles [RDA PID KI 2018] is enough for DiSSCo purposes, or whether a profile definition with additional attributes must be developed. There is a performance trade-off to be made between optimal retrieval and presentation of useful information at sub-second rates, and costly link following. The content of such profile really depends on the services that DiSSCo expects to operate directly in conjunction with the Handle service, as opposed to with the registry. These need to be defined.

Note 6: The expectation is that resolution of an NSId results (in the first instance) in data from the NSId Registry (nsidr.org) being returned to the user. This is likely to include a 'landing page' (i.e., a web page) containing human-readable information about the Digital Specimen and information serialized in machine-readable form (5.3.8).

For further study: The use of Digital Object Policies.

Note 7: On Digital Object Policies. Reconsider in conjunction with section 7.4.1. The draft RDA Recommendation on PID Kernel Information profiles [RDA PID KI 2018] explains the use of Digital Object Policies in specifying how to handle expected future changes to object contents. Changes to object contents imply changes to the object's Kernel Information record.

In DiSSCo, the expectation is that most Digital Specimens should have an objectLifeCycleType value of 'dynamic_irregular', meaning that the object content is likely to be updated at some point in the future, but it is not known when or if this will happen. Such changes are the result of formally defined processes operated for the determination and curation of specimens, and the result of processes operated by DiSSCo for the association of additional information to specimens.

Simple i.e., integer version numbers must be used. Each time the information content of a Digital Specimen is changed, the version attribute must be incremented by one and the dateModified and etag attributes must be updated.

Changing the digitalObjectLocation shall not result in change of the aforementioned version, etag and datemodified attributes.

Under some specific circumstances (see below) the objectLifeCycleType value can change to 'static', meaning that no future changes to the object content are expected or allowed. In this case, any future changes to the object content must result in the creation of a new independent, revised copy of the object. A link between the old and new objects must be made and expressed as Kernel Information attributes (“revisionOf” and/or “wasDerivedFrom”).

The specific circumstances in which a Digital Specimen is expected to undergo no further change to its information content (i.e., to become permanently static) are the following: <to be defined>.

6.3 Making data openly accessible

6.3.1 Policy for accessibility

Accessibility of data begins with their legal status. The scope of the data that must be openly accessible includes all that is produced, held or disseminated by virtue of the public task of DiSSCo Facilities and DiSSCo Hub, including use for research by those institutions themselves. Data that is not part of the public task or for which copyright is not held is excluded. Data produced or held as a result of activities funded by a non-public body is also excluded.

Note 8: The above statement is based on UK guidance for the cultural sector on the implementation of the Re-use of Public Sector Information Regulations 2015[16], introduced by European Directive 2013/37/EU (the ‘Amending Directive’). It is not possible to give an exhaustive treatment in the present document of the scope of data that should be accessible through DiSSCo. Readers are referred to relevant legal and data policy documents of DiSSCo and its member institutions.

The first step in access is findability (6.2). Access must therefore be made possible through freely accessible indexes that link to all relevant further information. This information must be open by default. DiSSCo Facilities and DiSSCo Hub holding such information must enable third parties to access, mine, exploit, reproduce and disseminate this data. Access and re-use may be restricted through appropriately access-controlled means to authorised users where it risks violating prevailing interests; such as national security and public safety, personal privacy, protection of endangered species and cultural resources, intellectual

[16] https://www.nationalarchives.gov.uk/documents/information-management/psi-implementation-guidance-cultural-sector.pdf.

property rights, etc. Restrictive declarations accompanying data and information that do not have a justification based on objective criteria in legislation are legally invalid and not permitted[17].

The DiSSCo policy with respect to access to data is therefore "as open as possible, as closed as legally necessary". 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[18] to adopt MIDS and MICS.

Access policies should be machine-readable and capable of being executed automatically as part of a data research procedure.

Exceptions to openness must be limited to reasons of national security, legal or regulatory compliance, sensitivity of collection information, or third-party rights. This policy reflects the dual aims of: i) maximising openness of data in the public interest, and ii) ensuring the application of legally required controls on access to sensitive information. Restricted data should not be delivered by DiSSCo Facilities to DiSSCo Hub.

Note 9: It is not possible to protect in one country against actions taken in another country that might accidentally reveal restricted information.

Note 10: Specific legislation applying to DiSSCo data includes the following:

- International multilateral environmental agreements (conventions), for example:
  - Convention on Biological Diversity (CBD), including the Nagoya protocol on Access and Benefit Sharing (ABS)[19];
  - Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES);
  - Convention on the Conservation of Migratory Species of Wild Animals (CMS);
  - IUCN Red List of Threatened Species;
- EC Regulations and Directives on:
  - General Data Protection Regulation (GDPR), Regulation (EU) 2016/679;
  - Open Data and Public Sector Information (PSI), Directive 2019/1024 (replacing the former Public Sector Information Directive 2013/37/EU);
  - Infrastructure for Spatial Information (INSPIRE), Directive 2007/2/EEC;
  - Conservation of natural habitats (Habitats), Directive 92/43/EEC;
  - Conservation of wild birds (Birds), Directive amended, 2009/147/EEC;
  - Invasive alien species (IAS), Regulation (EU) 1143/2014;
- Relevant national legislation.

Further study is needed to compile best practice recommendations and guidance for how DiSSCo should implement and demonstrate compliance with legal requirements of such legislation.

For further study. To be added, table(s) of what is openly accessible or not, and findable or not. For specimens and collections, the tables should be based on the defined information elements for MIDS and MICS. Location of specimens, place of origin of specimens (and by extension other information that might allow these to be determined) are examples of information elements that might not be either findable or accessible for specimens of rare (endangered) species or valuable gems.

[17] See: RDA-CODATA Legal Interoperability Interest Group, 2016. Legal Interoperability of Research Data: Principles and Implementation Guideline. doi: 10.5281/zenodo.162241; and Doldirina, C., Eisenstadt, A., Onsrud, H., and Uhlir, P. 2018. Legal Approaches for Open Access to Research Data. doi: 10.31228/osf.io/n7gfa.

[18] <Insert reference to DiSSCo open access policy when known>

[19] See https://www.cbd.int/abs/about/.

6.3.2 Tools for accessing data

FAIR principle A1: (Meta)data are retrievable by their identifier using a standardized communications protocol. FAIR principle A1.1: The protocol is open, free, and universally implementable. FAIR principle A1.2: The protocol allows for an authentication and authorization procedure, where necessary.

Data for individual digital objects are retrievable by the object's identifier (Handle) using the Digital Object Interface Protocol (DOIP) version 2.0 . Data is also retrievable through the REST (HTTP) API. DOIP and REST HTTP are open, free and universal protocols for information retrieval on the Web. A range of applications and services capable of exploiting these protocols are available to DiSSCo users. Appropriate authorization is necessary to retrieve data that is legally closed according to objective criteria.

6.3.3 Data retention, preservation and storage

FAIR principle A2: Metadata are accessible, even when the data are no longer available.

Data is retained for the lifetime of DiSSCo and its member organisations, which is currently expected to be for many decades. Data are stored in high-availability database servers of DiSSCo and its member organisations.

Many different preservation and storage systems and strategies exist for stewarding data associated with digitised natural science collections. For the DiSSCo community, needs vary on a spectrum from (at the one end) institutions with their own full in-house capability for all aspects of preservation and storage, to the opposite extreme of institutions having minimal levels of in-house capability with reliance on external/outsourced solutions. DiSSCo plays a flexible role by offering, on the one hand best practice guidance and standards for the management of data in-house through this DMP and its extension into policies, procedures and practices, through to offering concrete preservation and storage solutions 'as-a-service' to those institutions requiring specific services.

Preservation and storage solutions in DiSSCo can comprise, either individually or in combination:

- In-house repository systems, including Collection Management Systems (CMS) , Laboratory Information Management Systems (LIMS), Media/image Asset Management Systems (MAMS) and filestores;
- National-level facilities operated as a service by individual countries on behalf of their scientific communities;
- Generic international repositories, such as Zenodo, Dryad and Figshare; and,
- Thematic international repositories, such as the Global Biodiversity Information Facility (GBIF), Catalogue of Life (CoL), European Nucleotide Archive (ENA), Biodiversity Heritage Library (BHL), etc.

DiSSCo Hub exploits these capabilities as defined for each specific DiSSCo service, including for DiSSCo CMS-as-a-Service and DiSSCo MediaManagement-as-a-Service.

Data stores for the ECOI, ELViS and UCAS are the responsibility of DiSSCo Hub, with such data being stored in high-availability database servers of DiSSCo and its member organisations. These data stores are logically centralised but can be physically distributed and scalable, according to implementation. Interfaces between DiSSCo Facilities and these data stores must be low cost of initial entry and support continued operation.

Regardless of the specific preservation and storage strategy adopted, five functional areas must be addressed as part of good data management[20]: storage and location, fixity (permanence) and integrity, security, preservation of metadata and file formats. DiSSCo recommends the Digital Preservation Handbook[21] as a source of guidance and best practice on all aspects of digital preservation.

---

[20] Source: National Digital Stewardship Alliance, https://ndsa.org//activities/levels-of-digital-preservation/.

[21] Digital Preservation Handbook, 2nd Edition, http://handbook.dpconline.org/, Digital Preservation Coalition © 2015.

Preserving readability and retrievability of data, including image data over long periods of time is a protected characteristic (C5, section 4.5) of the DiSSCo infrastructure. DiSSCo Facilities and DiSSCo Hub must ensure that such data remains retrievable; for example, by providing appropriate services to retrieve and read data or by providing file format replacement services.

6.3.4 Data repositories: access and use restrictions

DiSSCo Hub maintains multiple open-access repositories and associated services for different categories of data e.g.:

- European Collection Objects Index (ECOI) for digital specimen and collection data;
- European Loans and Visits System (ELViS) for loans and visits transactions;
- Unified Curation and Annotation System (UCAS) for annotations and interpretations;
- Unified Knowledge Graph Index (UKGI) for links to supplementary data;
- Provenance and Traceability Index (PTI) for provenance data.

DiSSCo Hub maintains an open-access object type registry containing definitions of all the object types stored and processed within the DiSSCo infrastructure, including: Digital Specimen and Digital Collection objects, loan objects and visit objects, annotation objects and interpretation objects, link packet objects, and provenance objects, human and machine-readable license objects, and user identity objects.

Conditions for access to principal data categories must be defined by human and machine-readable license objects that implement the conditions of the DiSSCo data license types (6.5.1 below).

Identities of persons accessing restricted data must be ascertained in compliance with objectively justified restrictions of the DiSSCo policy on accessibility (6.3.1 above.)

Access to specific data items can be temporarily restricted in line with the DiSSCo embargo policy (6.5.3 below).

6.3.5 Multi-lingual support

Being able to find, access and use DiSSCo data via tools that operate in a specific user's native language is an essential means of achieving greater accessibility and re-use of data. To the extent practical, DiSSCo procedures and tools for data management should be available in the main European languages.

6.4 Making data interoperable

6.4.1 Policy for data interoperability

DiSSCo policy for data interoperability consists of four components, listed below and explained in sections 6.4.2 – 6.4.5:

1. Having the ability to exchange data between systems in common formats using standard protocols (syntactic interoperability);
2. Being in possession of a shared, congruent understanding of the context in which data exists and is exchanged i.e., attaching formal meaning to data through a process of interpretation and representing this with controlled vocabularies and relevant ontologies (semantic interoperability);
3. Agreeing upon and using common policies, principles and working procedures for digitisation and working with digital specimen and collection data across multiple organisations (cross-domain interoperability); and,
4. Having legal access to data, workflows and software, and their legal use and reuse across the domain (legal interoperability).

6.4.2 Exchangeable data

FAIR principle I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

DiSSCo uses JSON Schema for both internal and external representation of data and offers export to several popular formats such as CSV, Darwin Core (DwC) Archive, ABCD XML documents and RDF.

6.4.3 Vocabularies

FAIR principle I2: (meta)data use vocabularies that follow FAIR principles.

DiSSCo refers to open, external vocabularies for terms used, e.g.: OBO Foundry Biological Collections Ontology (BCO), Access to Biological Collections Data (ABCD) and its Extension For Geosciences (EFG), Darwin Core, Dublin Core, etc. Specifically, DiSSCo extends the BCO with terms relevant for digital specimens and digital collections and their relations to other concepts.

For further study: DiSSCo must build on and extend the concepts and relations in the Biological Collections Ontology (BCO), described here: 10.1371/journal.pone.0089606. Whether DiSSCo does this by expanding and extending the BCO itself, or by developing a separate ontology rooted in the BCO is for further study. The new concepts must be based on the new version of ABCD (https://abcd.tdwg.org/3.0/) and Darwin Core, taking into account also work aiming to "merge" ABCD and DwC (https://biss.pensoft.net/article/37491/). Related thoughts:

- Physical specimens, such as a preserved animal in a museum collection, or a herbarium specimen, or a fossil are examples of objects of class BCO:material_sample. As described in the above article, if a BCO:material sampling process is further carried out on a BCO:material sample, the resulting BCO:material sample is known colloquially as a subsample. Digitisation is one example of a BCO:material_sampling_process, and thus the images produced belong to class BCO:material_sample.
- Nevertheless, the BCO hasn't yet been extended to recognise digitisation as either a specific subclass of BCO:material_sampling_process or a subclass of the higher order OBI:planned_process.
- This Begs the question of what ontology exists or is needed to support natural science digitisation, and what would it best be derived from? Are there other digitisation ontologies in existence? If yes, and we use, then to be compatible with OBO Foundry ontologies such as BCO, they need to be able to trace their origin back to the Basic Formal Ontology (BFO).
- There are other digitisation ontologies developed: e.g., The FinBIF ontology (http://schema.laji.fi/) has been used to build the FinBIF portal, the Kotka CMS, and a field notebook service for citizen observers.
- See also definition of Unified Knowledge Graph in the glossary (section 14) and its associated note.
- OpenBioDiv-O, the ontology focusing on biological taxonomy [Senderov 2018] …… Also, https://doi.org/10.3390/publications7020038.

FAIR principle I3: (meta)data include qualified references to other (meta)data.

Each referenced external term is qualified by a resolvable context linking the term's properties to concepts in a relevant external ontology. i.e., JSON-LD context statements are used within JSON Schemas and Documents.

For further study: The specific formal vocabularies, relevant to various sub-communities within DiSSCo must be identified and listed. Gaps where new vocabularies are needed must be identified and filled. BCO, ENVO and PCO are probably all relevant. There's a long list of applicable vocabularies. For example, see Table 1 here: https://peerj.com/articles/1470. What are the key ones for DiSSCo?

- GeoNames for localities
- VIAF/Wikidata for collectors' names
- Others … …

6.4.4 Common policies, principles and working procedures

For further study: The common policies, principles and working procedures for mass digitisation, digitisation-on-demand and working with digital specimen and collection data across multiple organisations. Refer out to other relevant documents where these have an impact on data management.

6.4.5 Legal access to data

Legal interoperability can be achieved when the accumulated conditions of use for each and all the data are met, and when users can legally access and use each data without seeking authorization from data rights holders on a case-by-case basis. The ideal goal for legal interoperability is when data are positively identified as having no legal restrictions. See 6.5.2, Data licensing.

6.4.6 Building the Unified Knowledge Graph

DiSSCo aspires to a 'unified knowledge graph' for specimens in which digital representations in cyberspace of physical specimens in collections (i.e., Digital Specimens) sit as nodes in an interconnected graph of connections between specimens and connections from specimens to third-party sources of data/information related to those specimens. Collectively, this set of semantically interoperable objects substantially increases the potential for data re-use by ensuring that Digital Specimens and Collections are linked with relevant Supplementary Data to the maximum extent possible. Tools are planned to forage, suggest and create links.

Links between specimens/third-party sources can be of two types: i) actual links between a specimen and another specimen or piece of information directly derived from that first specimen e.g., a feather sample from a bird skin, a parasite found on a specific bird, a DNA sequence of it, a journal article referring to it; and ii) conceptual links i.e., samples/data were not taken or derived from the specific collected individual specimen but are indirectly related to it in some way e.g., this DNA sequence came from a bird of this same species but not from this specific catalogued individual, or this audio recording of the bird's song came from another bird like this one. Many examples of conceptual links are to the taxon concept the specimen represents rather than to the specimen itself. These are, nevertheless useful for many purposes. Both types of links can be expressed as values, such as a catalogue number, name or other text string, or they can be expressed more precisely and accurately as semantically rich links i.e., as a pointer to a definition or location of an existing verified value (e.g., links to GeoNames for textual locality information, links to VIAF/WikiData for collector names). In both cases, vocabularies (6.4.3) help to define and constrain the allowable values and characterize the relationship that the connection represents i.e., isPartOf, isRecordingOf, etc.

6.5 Increasing data re-use

6.5.1 Policy for reusability

The EU Directive for public sector information [Directive 2013/37/EU] states as a general principle that documents (including research data) to which the Directive applies must be re-usable for commercial or non-commercial purposes. DiSSCo data that is accessible without restriction must therefore be re-usable for all purposes.

6.5.2 Data licensing

FAIR Principle R1.1: (meta)data are released with a clear and accessible data usage license.

As far as DiSSCo data is covered by intellectual property rights, re-use must be made possible by using a copyright waiver such as CC0 or by applying an open access licence such as CC-BY. CC0 is the recommended choice and should be used in preference to CC-BY. Other licenses with more restrictive conditions must not be used.

License is a mandatory term in Digital Specimen and Digital Collection objects. Data retrieved and used by the users is subject to the license specified.

6.5.3 Embargo policy

As a rule, Digital Specimen data must be made publicly available as soon after digitisation, verification and curation as is practically possible. Publication of data does not preclude that the data can be modified and/or added to later.

Time embargoes on the release of research data may be justified by scientific needs, especially in order to verify results. However, such restrictions must be narrowly limited in time and have a justified basis.

6.5.4 Re-use by third-parties

Re-use by third parties is open by default, subject to any legal/regulatory restrictions based on objective critieria (6.3.1 above).

FAIR Principle R1: (meta)data are richly described with a plurality of accurate and relevant attributes.

To make re-use possible, each object contains a minimum of mandatory terms consistent with its formal object type definition, with the possibility to include optional additional terms and enrichments as necessary. In the case of Digital Specimen and Digital Collection object types, the minimum of mandatory terms corresponds to the object's classification as representing a specific level of digitization according to (respectively) the Minimum Information standard for Digital Specimens (MIDS) and the Minimum Information standard for Digital Collections (MICS).

FAIR Principle R1.2: (meta)data are associated with detailed provenance.

All data published/uploaded in DiSSCo is traceable to a registered DiSSCo user (the creator). Data describe the original collectors/authors of the published data.

FAIR Principle R1.3: (meta)data meet domain-relevant community standards.

DiSSCo is domain-specific for the natural sciences community, yet through adherence with relevant standard vocabularies, schemas and file formats from outside its domain, DiSSCo data is broadly usable by a wide cross-section of users across multiple domains.

6.5.5 Data life-cycle

The life-cycle of DiSSCo data is explained in detail in section 3. As general policy, data remains permanently re-usable, for the lifetime of the DiSSCo research infrastructure and beyond. See also section 7.5.3 on obsolescence. Deprecated, superseded and archived data must remain always eventually retrievable and reusable. This means readable. See protected characteristic C5 (4.5).

Due to the general data management principle that encourages the earliest and fullest publication of data (section 5.2.1), data managed by DiSSCo requires validation by experts. Data can change as new knowledge becomes available. This presents challenges for re-use and research reproducibility that are addressed by maintaining complete Provenance Data.

For further study. Requirements for persistently preserving user queries and results for reproducibility.

7 Identification of DiSSCo Data

7.1 Persistent identification of Digital Specimens and other objects

Each Digital Specimen or other digital object instance handled by the DiSSCo infrastructure must be unambiguously, universally and persistently identified.

The identifiers for Digital Specimens, Digital Collections, and other object classes shall be known as Natural Science Identifiers (NSId). These unique, persistent Handle System identifiers of the DOA, allocated during the digitisation process are location and service neutral and can be resolved by any available and well-known Handle service, such as http://hdl.handle.net/ or https://doi.org/. An NSId shall be assigned when the Digital Specimen or other digital object is first created.

Note 11: NSId is a distinct form of persistent identifier intended to integrate Digital Specimens and other data assets into the emerging global network of digital objects in the Internet. This includes, among other things all published journal articles identified by digital object identifiers (DOI).

Note 12: Other digital object classes to which NSIds can be assigned include collection and inventory objects as specified in Table 4, as well as individual images and sets of related images.

DMpr 11: Each Digital Specimen or other digital object instance handled by the DiSSCo infrastructure must be unambiguously, universally and persistently identified by a Natural Science Identifier (NSId), which shall be assigned when the object is first created.

NSId sits alongside other identifiers, such as institution specific ones and CETAF Stable Identifiers to identify digital collections, digital specimens, and other object classes within the European virtual collections that DiSSCo aims for. In the same way that Digital Object Identifiers (DOI) organise academic journal articles into a virtual collection of journal articles regardless of location (journal) or publisher, the effect of NSId is to virtualise natural science collections and associated services.

Note 13: NSId is an entirely separate form of identification from the collection-specific stable identifiers recommended by CETAF, which are intended primarily for unambiguously identifying physical specimens in collections at the level of the collection owning organisation.

DiSSCo operates a Europe-wide registry of digital specimens and other objects, and offers services based on resolution of NSIds. This registry is known as the European Collection Objects Index (ECOI) . For example, searching the registry to find which collections holds specimens of a specific kind, followed by resolving the NSId of any one from the list of results can automatically enter the user into the European Loans and Visits System (ELViS) to arrange a visit or a loan. Or it can initiate the request for hi-resolution imagery of that specimen.

7.2 Format of Natural Science Identifiers

Note 14: The present recommendation from the ICEDIG.eu project towards DiSSCo is that NSIds should be identifiers in the Handle System, like DOIs but with a specific DiSSCo top-level prefix and ability to mint its own handles. At the time of writing (Autumn 2019) discussions are ongoing about how to form a global consortium for identifying specimens/samples, including with IGSN and about becoming a Multi-Primary Administrator (MPA).

NSIds are Handles with the general form: "NN.prefix/suffix", where "NN." denotes a top-level 2-digit prefix, NN assigned to DiSSCo by the DONA Foundation; “prefix” denotes the second-level prefix assigned to DiSSCo Facilities by DiSSCo and "suffix" denotes the Handle suffix. Second-level prefix format is detailed in 7.2.1. Suffix format is detailed in 7.2.2.

NSIds are used for identifying many different object types but in general a distinction can be made (for which specific NSId variants are needed) between:

- Identifiers used to identify Digital Specimens and Digital Collections, both having identifiable physical counterparts; and,
- Identifiers used to identify transactions associated with specimens, such as interpretation and annotation events, and loans and visits requests.

7.2.1 Format of NSId prefixes

The prefix for identifiers identifying Digital Specimens and Digital Collections must indicate the registration agency responsible for minting and registering the identifier. This will normally be a lowercase alphabetic string that is either: i) the InstitutionCode allocated by the Global Registry of Scientific Collections[22], where that institution (DiSSCo Facility) is acting as the registration agency; or ii) the InstitutionCode “dissco” where DiSSCo Hub acts as the registration agency on behalf of one or more DiSSCo Facilities.

[22] https://www.gbif.org/grscicoll/.

The prefix for identifiers identifying transactions shall be a numeric prefix, allocated according to the transaction type as specified in Table 6.

Table 6: Prefixes for Transaction NSId types

Transaction type	Prefix
Interpretations	995
Annotations	996
Loans and visits	997
Other transactions	998

Note 15: Identifiers for each transaction type are minted and registered by the respective DiSSCo sub-systems dealing with each transaction type, each acting as their own minting and registration agent.

7.2.2 Format of NSId suffixes

Suffixes should be managed independently by each DiSSCo Facility and must be unique below the prefix to which they belong. The recommended suffix format for a Specimen NSId is a unique 8-character string obtained by base32 encoding a 40-bit positive integer (i.e., in the range $1 – 2^{39}-1$) as described in section 8 of [RFC4648]. Such suffixes are enough to identify more than 549 billion objects.

Note 16: Suffixes should conform to current recommended best practices i.e., they should:

- Remain 'opaque', meaning that it should not be possible to infer or interpret any meaning (especially concerning location) from the suffix itself;
- Work well in a web environment; for example, avoiding characters problematic in URLs;
- Be short and human-readable, meaning as short as possible but generally long enough to accommodate the expected number of objects to be curated;
- Be resistant to transcription errors, by avoiding use of easily confused characters such as 0 (zero) and O ('oh'), I ('eye') and l ('ell'), etc.;
- Be easy to generate.

7.3 NSId minting and registration

Each DiSSCo Facility shall be responsible for creating (minting) and managing their own NSIds in accordance with this policy, and for registering their own Digital Specimens with the DiSSCo Hub infrastructure.

Note 17: Registering a Digital Specimen or other object type causes the creation of an NSId Record. When a Digital Specimen is first registered, an NSId Record will be stored in the NSId Handle Service and an NSIdR Record will be stored in the NSId Registry (https://nsidr.org/). The NSId Handle Service and the NSId Registry together implement and act as the European Collection Objects Index.

DMpr 12: Each DiSSCo Facility shall be responsible for creating (minting) and managing their own NSIds in accordance with the DiSSCo policy for NSIds, and for registering their own Digital Specimens with the DiSSCo Hub infrastructure.

For an interim period and on a case-by-case basis, until contributing institutions can create/mint and manage their own NSIds, DiSSCo Hub infrastructure may create/mint and manage NSIds on behalf of specific DiSSCo Facilities as and when Digital Specimen information is published to DiSSCo Hub by the DiSSCo Facilities.

7.4 NSId resolution

DiSSCo operates one or more Local Handle Services to hold and resolve NSId Records corresponding to each Digital Specimen or other object type. Resolution of an NSId shall always return the current version of an object's content, as well as any interpretations and annotations associated with it.

DMpr 13: Resolution of an NSId shall always return the current version of an object's content, as well as any interpretations and annotations associated with it.

7.4.1 Controlling restricted data

NSId minting and registration and NSId resolution must not contravene the policies for findability (6.2.1) and accessibility (6.3.1) by accidentally making restricted information publicly findable and accessible. Services functionalities must not allow restricted data to be delivered to unauthorised users.

For further study. It is for further study as to how the Handle System can be used to support restrictive resolution. During the Handle minting and registration process, additional services can be used to check against CITES, IUCN Red List, etc. Such services can also be used when object re-identifications are made in a machine actionable way (e.g., object copying/cloning) to prevent that with a re-identification suddenly all object data becomes public. Restricted data in objects can be encrypted.

7.5 Mutability, versioning and obsolescence

7.5.1 Mutability of objects

The data content of Digital Specimens and other object types is not static, for a variety of reasons: New supplementary information about a specimen is found; new knowledge comes to light that leads to a reassessment of a previously classified specimen; images associated with a specimen might be found to be faulty (out of focus, wrong exposure, etc.); errors in transcribing label data have been made. All these and other reasons are reasonable cause to modify the data associated with a specimen. Similar considerations apply for other object types. Specifically, transaction related object types are living records of the status of a transaction, updated as a result of transaction events.

Each Digital Specimen object in the DiSSCo infrastructure has a single DiSSCo Facility as its current owner (ownedBy attribute). Approved and authorised experts, including the owner or their agent are permitted to change the Authoritative Data of a Digital Specimen (as explained in 5.2.3) whilst others can only use the specimen or add to or modify its Supplementary Data (and that also might have some access control). All changes, additions and deletions to contents of a Digital Specimen or Digital Collection object must be recorded as part of the object's history (Provenance Data), including details of who made the change.

Thus, the principle object types in DiSSCo are treated as mutable objects with access control and object history (provenance).

DMpr 14: The principle object types in DiSSCo (Digital Specimens, Digital Collections) are treated as mutable objects with access control and object history (provenance).

Note 18: Initially, DiSSCo will rely on Access Control Lists (ACL) for controlling access to objects but may eventually support digital signing of data and private/public keys for data guardians, along with providing digital wallets.

Note 19: Users without approval to modify specimen objects can make suggestions for modifications or additions by recording them as Annotations associated with the object. These are made through and recorded by the Unified Curation and Annotation Service (UCAS) and presented when the Digital Specimen is requested/viewed.

7.5.2 Versioning approach

Because of the mutability of objects, resolution of an object's identifier always returns the latest version of that object. Thus, users referring to or using a Digital Specimen are assured they are always receiving or working with the current/latest information about a specimen at the specific moment of use.

Nevertheless, users sometimes need to know the state of information about a specimen or collection at a moment in the past e.g., when a journal article is published, what was the available specimen data referred to (cited) by the article?

The DiSSCo data versioning approach relies on timestamped records of change (provenance data about actions performed on an object) in conjunction with a reconstruction service to provide a 'version' at a specific date and time in the past.

DMpr 15: Timestamped records of change (provenance data) must be kept, allowing reconstruction of a specific 'version' of a digital object at a date and time in the past.

Note 20: Such a reconstruction service is more powerful than it first appears. Not only can a specific Digital Specimen be reconstructed to provide its information content at a date/time in the past, but more general queries can be requested. These might include, for example: 'search using name X as it was applied in 1955'.

Note 21: See also section 12.5 on Data attribution and citation. A 'make data citable' service can store a timestamped snapshot of the relevant data, which can be retrieved via an assigned persistent identifier. The precise mechanism by which DiSSCo will implement this capability is for further study.

7.5.3 Object obsolescence, NSId errors and deletion, image file replacement

The general policy is that once created, Digital Specimens, Digital Collections, other object types and their associated identifiers exist permanently; for the lifetime of the DiSSCo research infrastructure and beyond.

NSId creation and object registration are tightly controlled procedures designed to ensure permanence and persistence. Thus, once an NSId string has been created and an object registered (usually both achieved in a single step) there is little room for change. The specific cases below must be managed

7.5.3.1 Object deletion due to extraordinary circumstances

Once created, an object and its identifier must not be deleted except in extraordinary circumstances. Extraordinary circumstances that justify object deletion can include:

- Objects that the object creator had no legal right to create.
- Improperly issued (published) data due to internal failures to follow due process.

The full range of extraordinary circumstances under which an object can be deleted is for further study.

The minimum action is that an identifier of a deleted object must be redirected to resolve to a 'tombstone' object that stands in for and indicates that the original object is no longer available.

7.5.3.2 Ordinary circumstances of object obsolescence and erroneous creation

Even under ordinary circumstances objects can become obsolete (i.e., because of one or more of the object is no longer maintained, used or is out of date); or objects and identifiers can be created in error.

When an object becomes obsolete it must be marked as such and its use must be deprecated. If appropriate, a link ("replacedBy") to any superseding object should be given. The superseding object may point to the obsolete object with a "replaces" relation. An obsolete object may be archived (but see 6.5.5 on re-use).

The following specific cases of erroneous creation must be managed:

1) If an object has been registered by mistake and this is not due to one of the extraordinary circumstances listed in 7.5.3.1 above, the object should be declared obsolete, as above. But see (3) below.
2) If an object has been registered with an NSId that is not the NSId by which it is referred to externally (for example, in a journal article) then the object may be re-registered with the correct identifier and the incorrect NSId must be aliased to the re-registered object.
3) If a wrong NSId has been created and that <u>NSId has not been distributed, communicated or used for linking</u>, then the object and the NSId must be deleted and the object re-registered correctly. If the NSId has already been distributed, etc. then the object must be declared obsolete and superseded by a correctly re-registered object. The wrong NSId must be marked as "replacedBy" the new object.

4) If a properly formed NSId is used incorrectly to identify the wrong object, and that NSId could correctly identify an object yet to be published, then the first object must be deleted and re-registered and the relevant correct object must be published.
5) Identification of data for temporary purposes (see 7.5.3.3).

7.5.3.3 Image file replacement

Potentially, an image file (object) can become obsolete due to evolution of image file formats, transformation of images and replacement by newer file formats. Such cases may be treated as ordinary circumstances of image obsolescence (7.5.3.2 above) and the image may be superseded by a newer image.

7.6 Institution codes and collection codes

The Global Registry of Scientific Collections GRSciColl[23] (formerly, GRBio) serves as an authoritative registry for identifying collection-holding institutions by institutionCode and collections by collectionCode. These codes must be used in the publishing of data as Digital Specimens and Collections. Other object types (e.g., Organisation and Collection) may also use these codes.

7.7 Identification of people

Management of person identifiers, including living and deceased persons/collectors. Living persons of any role can obtain ORCIDs and these can and should be supported by DiSSCo. However, for already deceased persons, such as collectors from more than 100 years ago, ORCID will not assign identifiers. Thus, another scheme will be necessary for these, such as the ISNI scheme. However, it can be done in several ways; it just needs someone to hold the definitive authority file. Even NSIds can be used. Nevertheless, ORCID will be important in the future, so we may expect that a person could have more than one person-identifier.

For further study. ISNI/VIAF/ORCID for people.

7.8 Identification of people and organisations

Similar considerations must be made for organisational identifiers, although here the solution is more obvious: ISNI (www.isni.org) provides ISO 27729 Identifiers and has already 700,000 organisations in its database. Now strengthening and opening database and providing portal/api for additions to enhance records. But ROR.org looks like the new game in town, with the weight of Crossref behind it. You can try it here: https://ror.org/search (but not working yet from a mobile device).

For further study. There is a cost to using ORCID and Crossref services that will have to be borne by DiSSCo, either centrally or by institution depending on the partnership arrangement reached with each organisation and the operational mechanism that will apply.

7.9 Identification of data for temporary purposes

Sometimes, data needs to be identified unambiguously but for temporary rather than persistent purposes. The definition of temporary in this context means a short period of time (perhaps of hours, days, or weeks, up to several months) where there is a need to transiently and unambiguously identify (temporary) data for management, administrative, research or other purposes but where there is no expectation of that data being further used beyond the immediate current need, formally referred to or cited by others i.e., in published collections, journal articles, etc.

Under these circumstances, data may be identified temporarily using 'temporary NSIds' having the special prefix "NN.999". Requests for identifiers under this prefix must be justified and made via the DiSSCo Hub minting service. NSIds having this prefix can disappear at any time and thus must not be relied upon in cases where persistence is essential. Note that this is a special case of object deletion (7.5.3).

---

[23] https://www.gbif.org/grscicoll.

If it subsequently becomes clear that data initially identified as needing temporary identification should become persistent, then such data must be properly identified and published through the normal registration process.

7.10 Authenticity and status of replicas and copies

7.10.1 Signatures for authenticity

Being able to determine the authenticity of digital specimen and collection data is an essential aspect of generating trust and confidence in DiSSCo data. Authentic DiSSCo digital specimen and collection data is data that has not been altered when compared to the original 'reference data' held in the CMS of collection-holding organisations.

Cryptographic hashes are an easy way to verify the integrity of data. A SHA-256 hashing function yields an almost-unique 32-byte signature that can be used to determine whether modifications have been made to the block of data with which it is associated.

Each Digital Specimen and Digital Collection should be created with a 32-byte SHA-256 signature, and this signature should accompany the specimen data when it is communicated from one system to another. When such data is adapted, remixed, or transformed in some way, performing the hash function over the same data elements must yield the accompanying signature value for the data to be considered as authentic.

7.10.2 Replicas as trusted duplicates

Replicas are precise duplicates of the original data, each with the same signature. Replicas are linked such that when a change is made to the original data, it is possible for all replicas to be found and updated. The mutable status of replicas means that a change to one replica propagates both to the original and to all other replicas. This allows, for example changes to the original data in the CMS of a DiSSCo Facility to be propagated to the authentic replica in the ECOI, and vice versa.

Each replica must be created with its own NSId that becomes linked in the Handle System to the NSId of the original object. Resolving one NSId must reveal all available replicas.

The precise circumstances under which replicas are needed are for further study.

7.10.3 The lower status of copies

Copies, on the other hand are not linked. Changes made to the original or to a copy cannot be propagated to other copies because the existence of these may not be known. Thus, copies can diverge from the original and from one another as they are modified. Over time, copies cannot be relied upon as an authentic resemblance of the original object (or of other copies) because the original object is mutable (and may therefore have changed) and there is no linkage between the original and any copies of it.

8 Data service management and service level agreements

Data service management covers the entire spectrum of services provided by DiSSCo Hub to its community of users that includes researchers, collections staff, non-governmental organisations, students and citizens, as well as members of the Press/media. DiSSCo does not offer its services under formal Service Level Agreements (SLA) because to do so would admit that non-compliance with stated levels of service is open to enforcement by legal means and compensation, which is not appropriate when end-users are not paying for the services they use. DiSSCo prefers an approach based upon openness, transparency, standards and best practice whereby trust is earned through demonstrated availability and reliability of service.

By the same principle, DiSSCo does not enter into and enforce SLAs with those service providers that DiSSCo itself depends upon, except in cases where financial payment is made for services provided.

DiSSCo undertakes its service management responsibilities in compliance with the XYZ framework (see next paragraph) for IT service management.

For further study. There is a choice of two applicable lightweight IT service management frameworks: FitSM (https://www.fitsm.eu/) and VeriSM (https://verism.global/). VeriSM is newer and more organizationally oriented whereas FitSM is more IT oriented. FitSM is already established at the EOSC level and DiSSCo can make use of existing resources and training provided by EGI/EOSC. An important consideration in applying any ITSM framework will the extent to which certain activities are carried out centrally by DiSSCo Hub due to lack of capacities and competencies in the DiSSCo Facilities, and how this will this change over time. There is a case for suggesting that the VeriSM activities of 'define, produce, provide, respond' are carried out centrally for DiSSCo Hub services. Doing more centrally, on behalf of the DiSSCo Facilities can increase the value of DiSSCo to its member institutions. On the other hand, an important part of the DiSSCo mission is to increase the capabilities at the institutions as well. Thus, 'define, produce, provide, respond' can be used to deliver upskilling and increased capability into the DiSSCo Facilities.

9 Data quality and minimum information standards

For further study. Describe how DiSSCo uses emerging data quality assessment frameworks, processes and standards as the means of delivering data fit for purpose.

For further study. This is where the detailed mandatory / optional element tables for MIDS and MICS appear until we have a reference to a self-standing external document.

10 Data security

10.1 Back-up, recovery and service continuity

Back-up and recovery planning and service continuity planning are essential protections against data and/or service loss due to calamities that can include accidental deletion, equipment failure, fire, flood or other disaster. Back-up, recovery and continuity plans must be maintained and regularly tested.

10.1.1 For DiSSCo Hub

DiSSCo Hub shall take steps to secure data against the possibility of loss and to ensure service continuity. In the case of data loss, two cases must be covered: i) loss due to disasters such as equipment failure, fire, flood, etc., and ii) loss due to deletion/corruption of data as consequence of unintended action. The latter is considered separately below.

Plans shall be documented and maintained in a DiSSCo Disaster Recovery Plan, considering the anticipated risks and procedures for each service and each database. The expectation shall be that equipment will fail and recovery will be needed. Thus, the disaster recovery plan shall be tested monthly.

An approach based on daily off-site backups of key databases to be recovered, combined with automated deployment procedures for those databases and for key services is adequate. This avoids the need to maintain standby machines synchronised with live services yet ensures prompt and efficient redeployment and restart on fallback machinery when needed.

10.1.2 For DiSSCo Facilities

Each DiSSCo Facility is strongly recommended to maintain its own back-up and disaster recovery plan in line with its institutional requirements.

10.1.3 Unintended data deletion

The likelihood of unintended data deletion is low due to the privileged nature of data deletion operations (section 7.5.3). Nevertheless, such cases do occur, and such events must be managed to return deleted data to its prior state where possible. Important data that must be recoverable on request include:

- Digital Specimen and Digital Collection objects
- ELViS loans and visits transactions
- UCAS annotations and interpretations.

The precise mechanism by which this is achieved can vary from service to service.

10.2 Physical data security

Protected characteristic C9 Securability (authentication, authorization, accounting, auditing) is implemented through DiSSCo's Identity and Access Management (IAM) sub-system.

Authentication and authorization controls to ensure appropriate protection of data (protected characteristic C4) must be applied to all users and other persons working with DiSSCo data, in line with the DiSSCo User Access Policy and other requirements for identity and access management. This is to ensure not only compliance with necessary legal and regulatory requirements but also to ensure unauthorized access and tampering. Identity and Access Management shall include appropriate controls on physical access to equipment, data and services, as well as electronic access.

IAM remains compatible with national and international (European) identify federation mechanisms (i.e., eduGAIN) and EOSC federated identity services; offering single sign-on (SSO) and single logout (SLO) and authorised access capabilities. Alternative mechanism(s) providing equivalent levels of authentication and authorization compliance are sustained for users outside of the European research and education sector.

For further study: Governance arrangements. How do users become authorised? It should be delegated to DiSSCo Facilities for their employees and done centrally by DiSSCo Hub for all others. Against agreed criteria and proof of identity when necessary. Must include collecting explicit consent for lawful processing of their personal data in conjunction the management and use of DiSSCo data, including public display of certain personal information elements as appropriate; for example, their name as a collector of a specimen or annotator of specimens. Such information can include the optional provision of contact details for public display e.g., as part of a user profile record/page.

10.3 Certification

DiSSCo aims to achieve the highest levels of certification for all areas of operations, including for data storage and security e.g., CoreTrustSeal.

10.4 Compliance with GDPR (security)

Consent in relation to the General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 is dealt with in 12.1 below. The present section is concerned with data management principles deriving from Article 32 Security of data processing.

For further study.

11 Data provenance

Data provenance is concerned with providing a traceable record about data, its origins and the processing actions that have been applied to it. DiSSCo provenance data records the history of Content Data, Supplementary Data, Annotations and Interpretations.

Data associated with published studies should be verifiable and fully traceable from production to publication, and later republication.

It must be possible to trace the provenance of DiSSCo data backwards from where it is used to its point of origin. It should be possible to reproduce the process by which it ended up in its present state. It should also be possible to trace forwards to determine how data is being used.

Provenance information must be readable both by humans and by machines. Tools and libraries implementing the W3C PROV specifications [W3C PROV 2013 [24], such as those listed under the openProvenance initiative [openProvenance 2018] should be employed throughout to support automated provenance generation and tracking. This leads to the potential for provenance graphs to be automatically traversed to understand origins and dependencies.

12 Ethical and legal aspects

12.1 Compliance with GDPR (lawfulness of processing)

Security in relation to the General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 is dealt with in 10.4 above. The present section is concerned with data management principles deriving from Article 6 Lawfulness of processing.

The only basis for lawful processing of personal data by DiSSCo itself and DiSSCo users is explicit consent by the data subject[25].

Names of researchers and collectors on labels, in Content Data, in publications, and employed by institutions are legitimate public information because that is the normal expectation of the sector. Data revealing personal details beyond names must be more carefully protected. Nevertheless, compliance with the General Data Protection Directive (GDPR) requires that DiSSCo provides clarity about personal information collected: where personal information is collected and the reasons for doing so; what DiSSCo uses personal data for and the legal basis in each case. The DiSSCo Privacy Policy sets this out and is additional to the privacy policies of the DiSSCo member institutions.

For further study. Privacy policy to be written. A good structure example is here:
https://www.eurostar.com/uk-en/privacy-policy.

It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection[26]. Therefore, data subjects should be allowed to give their consent to areas of scientific research in the scope of DiSSCo when that research is carried out in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose.

12.2 Compliance with INSPIRE

For further study. Refer to relevant annexes / articles of the Directive. Specific rules in relation to Metadata, Data Specifications, Network Services, Data and Service Sharing and Monitoring and Reporting.

12.3 Subsidiary procedures applicable at national or other level

For further study.

12.4 Outsourcing digitisation to subcontractors, transcribers, etc.

All or parts of a process or project of digitisation can be outsourced to subcontractors, which can the subject of a commercial arrangement or voluntary work. In such cases, and especially in a commercial arrangement

---

[24] For background and tutorial, see: Missier, P., Belhajjame, K., Cheney, J., 2013. The W3C PROV family of specifications for modelling provenance metadata. In: Proceedings of the 16th International Conference on Extending Database Technology - EDBT '13. ACM Press, New York, New York, pp. 773. https://doi.org/10.1145/2452376.2452478.

[25] General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, article 6(a), recital 32.

[26] General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, recital 33. In the DiSSCo context this can mean at the moment a collector's name is assigned to a specimen or the moment at which an annotation, etc. is made and attributed to specific person.

the work should be the subject of an outsourcing agreement specifying the work to be done, applicable quality criteria, cost, timescale, etc.

Outsourcing agreements should address important data management principles, such as guarding against data loss and ensuring compliance with relevant legal aspects to ensure the usability, integrity and protection of data generated from outsourced work. Especially important is to include provisions covering transfer of rights from the subcontractor to the public domain and to respect the personal privacy of those carrying out the work.

12.5 Data attribution and citation

Even for data in the public domain i.e., subject to a copyright waiver such as CC0 there is a moral obligation to correctly attribute sources of data, for example to organisations and/or individual persons. For data subject to a licence such as CC-BY there is a license obligation to attribute and credit appropriately.

The RDA-TDWG Attribution Metadata Recommendation [RDA ATTRIB 2019] recommends the use of three information elements: the person performing the curatorial action, the action they perform, and the digital or physical object they are curating as the means for attributing work

Attributions include citations of formally published works describing the data (so-called 'data papers') that are becoming increasingly common, as well as, for example taxonomic treatments and heritage literature.

The RDA Recommendation on Data Citation of Evolving Data [RDA CITE 2015] recommends storing data in a versioned and timestamped manner and identifying citable data sets by storing and assigning persistent identifiers to timestamped queries that can be re-executed against the timestamped data store. This is covered in section 7.5.2.

13 Other data management issues

13.1 Software maintenance and sustainability

For further study. Two initiatives dealing with this topic are https://www.softwareheritage.org/ and https://www.software.ac.uk/. Similar sustainability issues can apply to workflows as well (we can treat "workflow as code").

14 Glossary of terms and abbreviations

Terms and abbreviations used in the present document have the meanings given below.

Annotation: An assertion made on or about the Digital Specimen, such as determination of the species and comments. One of the main data types managed by DiSSCo.

Authentic Image of Record ("lossless image"): The original, uncompressed image(s) produced by the digitisation process and retained/preserved as the categorical absolute reference image(s) of the physical specimen. Note that the authentic images of record are not usually the ones made available for on-line work due to their often-large file sizes.

Authoritative Data: Data that is authoritative about a specimen or collection. See definitions of data and supplementary data. One of the main data types managed by DiSSCo.

Basic Record: A database typically in the institution's CMS containing a limited set of information about the specimen; for example, physical specimen identifier and barcode details, and details of the collection it belongs to. (Cf. regular record and extended record).

Collection Digitisation Dashboard (CDD): A system that collects and presents reliable, complete and up-to-date information on the taxonomic and geographic scope of collections as well as the degree and level of digitisation achieved and remaining.

Collection Management System (CMS): A system (typically a database) for recording and organising information about the objects in a museum or other collection.

CMS-as-a-Service (CMSaaS): A CMS provided as a service by an infrastructure provider/operator for the benefit of its community of users.

Common Services: Services provided, for example by the European Open Science Cloud (EOSC) for the benefit of the scientific community at large. Such services can include general and high-performance computer processing services, services for discovery and collaboration, and services for authentication and authorization of users.

Content Data: An alternative term for data when it is necessary to differentiate from other kinds of data. See also the definition of Data. One of the main data types managed by DiSSCo.

Data (a.k.a. Content Data, Authoritative Data): Data relating directly to describing collections and physical specimens, such as (in the latter case) images of those specimens, information from specimen labels (such as scientific name, location where collected, date collected, collector name, etc.), or measurements and other analyses of specimens. One of the main data types managed by DiSSCo.

Note 22: An essential characteristic of this data is that it is authoritative about a specimen or collection. That is, the information that this data represents has been determined by the scientists and curators at the owning institution and it is they alone that hold authority to make changes to it as knowledge and understanding about specimens and collections evolves. When clarity is need, the term 'Authoritative Data' is sometimes used. Contrast with the definition of supplementary data.

Data Acquisition: A data lifecycle phase concerned with the principal activity of digitising physical specimens by dedicated digitisation lines/factories.

Data Curation: A data lifecycle phase concerned with curating digital specimen and collection data.

Data Publishing: A data lifecycle phase concerned with making curated data accessible to DiSSCo users and publicly to parties external to the DiSSCo infrastructure, as well as directly to other services within DiSSCo.

Data Processing: A data lifecycle phase concerned with further processing of data after it has been curated and published.

Data Use: A data lifecycle phase where the research community (represented by the scientist/researcher role) can exploit the digital and physical collections for science and can design digital experiments and analyses acting on the published and processed data.

Digital Collection: A digital representation (surrogate) corresponding to a natural science collection. Cf. Digital Specimen.

Digital Collection type (DC): A collection of property definitions about a Digital Collection, the structure of which conforms to the requirements of the openDS specification [openDS <date>].

Digital Object (DO): A meaningful entity of an application domain that is exchanged between and processed in different information systems. “Meaningful” implies that the content of the object has value for and can be interpreted by a human and/or a machine. Digital objects have content, such as data and are defined by a type, such as Digital Specimen type or Collection type. Each object is uniquely and ambiguously identified by a persistent identifier (pid).

Note 23: This definition is a practical definition for DiSSCo data management purposes. It has been refined from [Wittenburg 2019b], incorporating elements of the formal technical definition provided by the DONA Foundation (https://www.dona.net/) as custodian of the Digital Object Architecture: "A sequence of bits, or a set of sequences of bits, incorporating a work or portion of a work or other information in which a party has rights or interests, or in which there is value, each of the sequences being structured in a way that is interpretable by one or more of the computational facilities, and having as essential elements an associated unique persistent identifier (pid) and a type." [DOIP 2.0 2018]

The concept of a digital object is the same as the notions of a digital object defined by the Society of American Archivists and the Research Data Alliance, and the same as the notion of digital entity defined in ITU-T Recommendation X.1255. A specific characteristic of digital objects is that they can possess methods that can be invoked upon their contents.

Digital Object Architecture (DOA): A logical extension of the Internet architecture supporting digital information management more generally than just conveying units of information from one place in the Internet to another. [Kahn 2006].

Digital Object Interface Protocol (DOIP): One of two standard communication protocols (the other being the Identifier Resolution Protocol) supporting the Digital Object Architecture, that specifies a standard way for software clients (applications and services) to interact with digital objects. [DOIP V2.0 2018]

Digital Specimen: A digital representation (surrogate) corresponding to a physical specimen in a natural science collection. Cf. Digital Collection.

Digital Specimen object type (DS): A collection of property definitions about a Digital Specimen, the structure of which conforms to the requirements of the openDS specification [openDS <date>].

Digitisation: The process of making physical objects digitally available.

Digitisation Line/Factory: The facilities (premises, personnel, equipment (hardware, software)), processes and procedures necessary for large-scale, mass digitisation.

DiSSCo: See Distributed System of Scientific Collections.

DiSSCo Hub: The infrastructure of integrating services, information technology components (hardware and software), human resources, organisational activities, governance, financial and legal arrangements that collectively have the effect of unifying natural science collections through a holistic approach towards digitisation of and access to the data bound up in those collections. Cf. definition of DiSSCo Facility(ies).

DiSSCo Digital Object Repository (DDOR): A repository of instances of various digital object types managed by the DiSSCo infrastructure.

Note 24: It is likely that digital objects managed by the DiSSCo infrastructure can and will be widely distributed across different repository systems, of which the DDOR is just one example.

DiSSCo Facility(ies): The geographically distributed collection-holding organisation(s) (i.e., natural science/history collection(s)) and related third-party organisations that deliver data and expertise to the DiSSCo Hub infrastructure, and which can be accessed by users via the DiSSCo Hub infrastructure. Cf. definition of DiSSCo Hub.

Distributed System of Scientific Collections (DiSSCo): A pan-European Research Infrastructure mobilising, unifying and delivering bio and geo-diversity digital information to scientific communities.

European Collection Objects Index (ECOI): A searchable, electronic index of catalogued objects (typically, specimens and collections) held by and discoverable (findable, accessible) in natural science collections of the DiSSCo collection-holding partners.

European Loans and Visits System (ELViS): A unified pan-European system for managing loans and visits access to any collection for any authorised user under a consistent access policy (for restrictions, responsibilities, reporting, etc.)

Extended Record: A database record, typically in the institution's CMS containing a comprehensive set of elements describing the specimen; for example: all label data, annotations and determination history. (Cf. basic record and regular record).

External Sources: Sources providing additional data, held externally to the DiSSCo infrastructure from outside the DiSSCo Facilities that can be linked to collection objects/specimens. They can also provide specialist data and/or processing services. DiSSCo can serve external sources with high quality data delivery services for DiSSCo originated data.

FAIR: A set of four foundational principles (Findability, Accessibility, Interoperability, and Reusability) serving to guide data producers and publishers towards good data management and stewardship. See Wilkinson et al. 2016 (doi: 10.1038/sdata.2016.18)

FAIRness: A characteristic exhibited by an infrastructure (component) when it maintains compliance with the principles of FAIR.

Handle System: An implementation of the Identifier and Resolution component of the Digital Object Architecture (DOA).

Identifier Resolution Protocol (IRP) (a.k.a. the Handle System Protocol): One of two standard communication protocols (the other being the Digital Object Interface Protocol) supporting the Digital Object Architecture, that specifies a standard way for creating, updating, deleting, and resolving digital object identifiers (or other handles, such as the Natural Science Identifier).

Interpretations: Interpretations are the application of expertise, based on facts at hand to define more precisely the exact meaning of ill-defined text describing specimens. One of the main data types managed by DiSSCo.

Kernel Information: Meta-information about a digital object, stored at the DiSSCo Local Handle Service and thus accessible by a resolver and necessary for resolving the object's identifier; also allowing smart programmatic decisions accomplished through inspection of the object's PID record alone. (Cf. definition of PID kernel information in RDA Recommendation on PID Kernel Information [RDA PID KI 2018]).

Kernel Information Profile: Kernel Information Profiles are registered schemas for PID Kernel Information. Multiple kernel information profiles are needed to cover the various digital object types specified by the openDS specification [openDS <date>].

Long-term Image Archive (repository): A storage location for data that is not accessed very often. A long-term 'deep archive' is typically based on magnetic tape storage and has slow response times for access. Typically, in the context of digitised natural history collections, it is where the 'authentic images of record' are preserved – the 'long-term image archive', according to an institutional or other digital preservation policy.

Metadata: Metadata is additional data that establishes a context for the data it describes i.e., it is data about data. Data and Metadata are treated as the same by DiSSCo i.e., data can be metadata when necessary and vice versa.

MIDS (Minimum Information about a Digital Specimen): The minimum information standard for digital specimens specifies the mandatory and optional information elements that must be present in a digital specimen at various levels of digitisation.

MICS (Minimum Information about a Digital Collection): The minimum information standard for digital collections specifies the mandatory and optional information elements that must be present in a digital collection at various levels of digitisation.

Natural Science Identifier (NSId): A universal, unique persistent identifier for digitised natural science specimens (i.e., Digital Specimens) and other associated object types.

NSId Record (a.k.a. NSId Handle Record) : The minimum set of information about a digital specimen, in association with its NSId that permits smart programmatic decisions to be accomplished through inspection of the record alone, without the need for costly unpacking and following of links. The NSId Record is the combination of a persistent identifier and the handle (kernel) information defined by the digital specimen kernel information profile. NSId Records are maintained by the DiSSCo Local Handle Service.

NSIdR Record (a.k.a. NSId Registry Record) : A specific kind of record defining the comprehensive attributes of a Digital Specimen, a Collection, etc. stored in and directly accessible from the NSId Registry (nsidr.org). One or more NSIdR Records are returned to a user following a search/query for specimens having specified characteristics.

NSId Registry: A registry of groupings of digital objects (that can be distributed across one or more repositories, including the DiSSCo Digital Object Repository) that can be searched and browsed to find collections, specimens, etc. with specified characteristics.

Open-access repository: An open-access repository or open archive is a digital platform that holds research output and provides free, immediate and permanent access to research results for anyone to use, download and distribute. (Source: Wikipedia)

Persistent Identifier (PID): A persistent identifier is a string (functioning as a symbol) that identifies a digital object. The identifier can be persistently resolved (digitally actionable) to meaningful information about the identified digital object. In the case of DiSSCo the Natural Science Identifiers (NSId) are the principal persistent identifiers used.

Provenance Data: Data providing a traceable record about other data, its origins and the processing actions applied to that. One of the main data types managed by DiSSCo.

Regular Record: A database record, typically in the institution's CMS containing a partial set of information describing the specimen; for example, physical specimen identifier, barcode details, taxon name, collection data and place, collector, etc. (Cf. basic record and extended record).

Role: A role in the DiSSCo community is a prescribing behaviour that can be performed any number of times concurrently or successively.

Note 25: Scientist/researcher and curator are examples of roles performed in the context of DiSSCo. The major DiSSCo roles and their behaviours/responsibilities are defined in Table 1 and sections 3.4 – 3.8.

Serialization: The process of translating data object structures into a format that can be stored (e.g., in a file) and/or transmitted between one computer/software system and another. When the serialized bitstream is reconstructed according to the serialization format used, the original data object structure is obtained. JSON, JSON-LD, XML and RDF/XML are common serialization formats.

Supplementary Data: Other content data about a specimen, additional to the authoritative data that contributes to an overall understanding of the specimen. Supplementary data can be generated by specimen owners and/or by third-parties and can include biodiversity literature, DNA sequence data, chemical composition data, acoustic recordings, and other information relating to specific specimens and collections. One of the main data types managed by DiSSCo.

Unified Curation and Annotation System (UCAS): A system enabling direct contributions to the curation and improvement of natural science data, with advanced annotation services that including recording of annotation history and management of object annotations.

Unified Knowledge Graph (UKG): The DiSSCo Unified Knowledge Graph (UKG) acquires and integrates data about Digital Specimens into the DiSSCo Ontology and applies a reasoner to derive new knowledge about those entities and their relations. The reasoner can be human or a machine.

Note 26: This definition implies that an ontology (named the DiSSCo Ontology) is needed to formalise the interpretation framework for the data held in the UKG i.e., that ontological terms (meaning) and relations are attached to and encoded alongside the data, rendering the data as comprehensible information. The reasoner, whether it be human, or machine actioned derives the new (implicit) knowledge that the graph must also hold. The knowledge graph approach supports regular updates to information caused by the acquisition of new data, and growth in the interpretative framework (i.e., the DiSSCo ontology) itself as knowledge is derived and understanding changes. The definition and accompanying note are based on concepts described by [Ehrlinger 2017] and [Stichbury 2017].

15 References

[DOIP 2.0 2018] DONA Foundation, 2018. Digital Object Interface Protocol Specification, version 2.0, November 2018. https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf.

[ECMA-404] ECMA-404. The JSON data interchange syntax. 2nd edition, December 2017. http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf. For additional information, see also: https://www.json.org/.

[ENVRI 2017] ENVRIplus deliverable D6.1. 2017. D6.1 A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations. http://www.envriplus.eu/wp-content/uploads/2015/08/D6.1-A-system-design-for-data-identifier-and-citation-services-for-environmental-RIs.pdf. Accessed: 2019-07-17.

[Kahn 2006] Kahn, R. and Wilensky, R., 2006. A framework for distributed digital object services. *International Journal on Digital Libraries*, *6*(2), pp.115-123. https://doi.org/10.1007/s00799-005-0128-x.

[Kallinikos 2010] Kallinikos, J., Aaltonen, A., and Marton, A. (2010) A theory of digital objects. First Monday, 15 (6) http://ear.accc.uic.edu/ojs/index.h/fm/article/view/3033/2564. Accessed 16 July 2019.

[Lannom 2020] Lannom, L., Koureas, D., and Hardisty, A.R. 2020. FAIR data and services in biodiversity science and geoscience. Data Intelligence 2, 122–130. doi: 10.1162/dint_a_00034.

[Mons 2017] Mons, B., Neylon, C.D., Velterop, J., Dumontier, M., da Silva Santos, L.O.B., and Wilkinson, M. 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use, vol. 37, no. 1, pp. 49-56, 2017. https://doi.org/10.3233/ISU-170824

[Mora 2011] Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B., Worm, B. 2011. How Many Species Are There on Earth and in the Ocean?. PLOS Biology 9(8): e1001127. https://doi.org/10.1371/journal.pbio.1001127.

[OA 2017] Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, version 3.2, 21 March 2017.

https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

[openDS <date>] Standard for digital specimen and other natural science digital object types. To be written.

[RDA ATTRIB 2019] Thessen, A.E., Woodburn, M., and Koureas, D. 2019. RDA/TDWG Attribution Metadata Working Group: Final Recommendations. February 2019. doi: 10.15497/RDA00029.

[RDA CITE 2015] Rauber A, Asmi A, van Uytvanck D, and Pröll S. RDA Recommendation on Data Citation of Evolving Data. October 2015. doi: 10.15497/RDA00016.

[RDA PID KI 2018] Weigel T, Plale B, Parsons M, Zhou G, Luo Y, Schwardmann U, Quick R, Hellström M, Kurakawa K. RDA Recommendation on PID Kernel Information. https://www.rd-alliance.org/sites/default/files/RDA%20Recommendation%20on%20PID%20Kernel%20Information.pdf.

[RFC4648] The Internet Society. 2006. The Base16, Base32, and Base64 Data Encodings. Proposed standard RFC 4648. url: https://tools.ietf.org/html/rfc4648. (accessed 15 April 2019).

[Senderov 2018] Senderov, V., Simov, K., Franz, N., Stoev, P., Catapano, T., Agosti, D., Sautter, G., Morris, R.A. and Penev, L. 2018. OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 20189:5. https://doi.org/10.1186/s13326-017-0174-5.

[Stichbury 2017] Stichbury, J., 2017. WTF is a knowledge graph? [WWW Document]. URL https://hackernoon.com/wtf-is-a-knowledge-graph-a16603a1a25f

[W3C PROV 2013] Groth P and Moreau L (2013) PROV-Overview An Overview of the PROV Family of Documents. W3C Working Group Note 30 April 2013. https://www.w3.org/TR/prov-overview/.

[Wilkinson 2016] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3. https://doi.org/10.1038/sdata.2016.18

[Wittenburg 2019a] Wittenburg, P., Strawn, G., Mons, B., Boninho, L., Schultes, E. 2019. Digital Objects as Drivers towards Convergence in Data Infrastructures. doi: 10.23728/b2share.b605d85809ca45679b110719b6c6cb11.

[Wittenburg 2019b] Wittenburg, P. 2019. From persistent identifiers to digital objects to make data science more efficient. Data Intelligence 1, 6-21. doi: 10.1162/dint_a_00004.

Appendix A: User stories for DiSSCo

This appendix lists the user stories upon which the design of DiSSCo data management is based.

A.1 <first category of stories>

To be included later.

A.2 <second category of stories>

To be included later.

Appendix B: Information flows for NSId resolution

Figure 5 illustrates a typical information flow involved in the activity of resolving an NSId and retrieving data about a Digital Specimen.

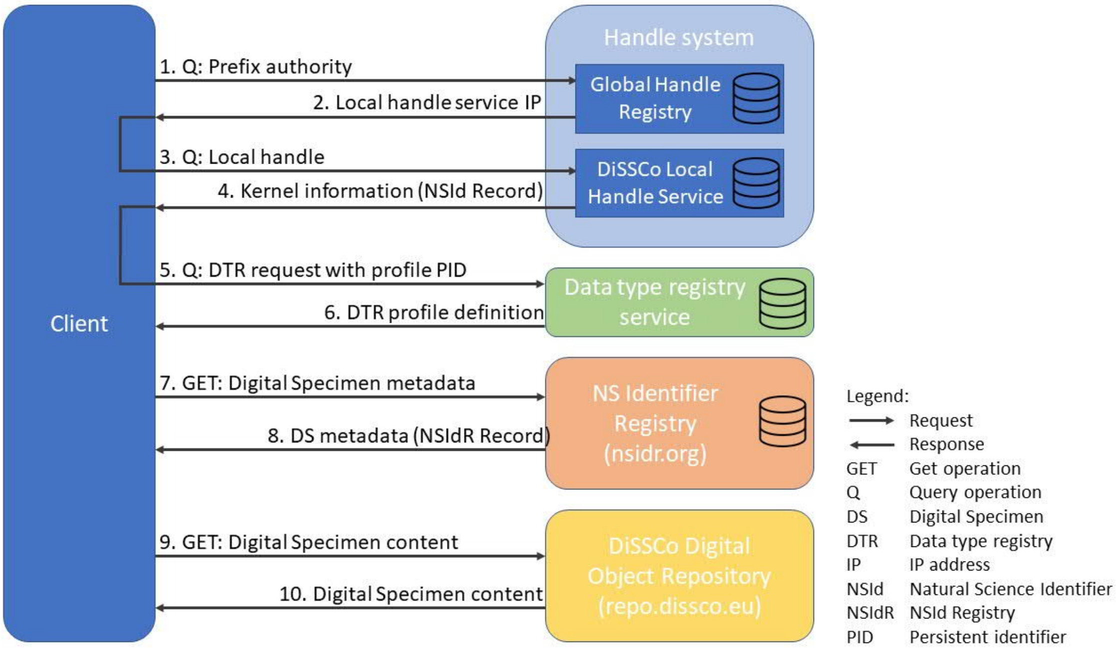


Figure 5: Resolving NSId and retrieving Digital Specimen data

The steps are as follows:

Step 1/2: The Global Handle Registry is queried to determine (from the NSId prefix) the Local Handle Service to be used to completely resolve the NSId.

Step 3/4: The DiSSCo Local Handle Service resolves the NSId and returns kernel information (the NSId Record) about the specimen of interest. The NSId Record is the minimum set of information about a Digital Specimen that permits simple but smart decisions to be made without the need for costly unpacking and following of links. However, it cannot be interpreted without an understanding of its structure.

Step 5/6: A request must be made to the data type registry for the NSId Record profile definition so that the NSId Record can be understood and interpreted. Keeping the profile definition out of the client (i.e., not fixing it in the client software) allows for the structure of the NSId Record to evolve if necessary.

Step 7/8: A request is made to the NSId Registry for the NSIdR Record for the specimen. This record contains the comprehensive data for the Digital Specimen and permits the following step to retrieve the required content.

Step 9/10: The required content from the Digital Specimen object (such as images, for example) is retrieved from the digital object repository. This can be from the DiSSCo Digital Object Repository (as illustrated) and/or any other relevant trusted repository.

Appendix C: Data Flow Diagrams for DiSSCo

C.1 Introduction

This appendix contains the data flow diagrams for the DiSSCo infrastructure.

C.1.2 Explanation of the data flows

As noted by the iDigBio project, there is already a broad range of digitisation workflow implementations throughout the collections' community, with no consensus workflow that fits all situations, even within a single preservation type (flat sheets, pinned specimens, specimens in jars, etc.). We do not expect that all institutions can harmonise every digitisation task, nor that they should want to do so. Thus, the Data Flow Diagrams (DFD) in this appendix capture the main data flows most likely to exist when considering a wide range of orderly, comprehensive digitization tasks such as those proposed by iDigBio[27]. We based our analysis initially on the iDigBio modular approach and extended from there based on community contributions and DiSSCo specific requirements.

The DFDs in this appendix support the two main variants of digitisation workflow i.e.:

1. Data entry preceding image capture
   (object -> transcribe data -> make image -> store image -> publish digital specimen)
2. Image capture preceding data entry.
   (object -> make image -> store image -> transcribe data -> publish digital specimen)

The purpose of the DFDs is principally to help DiSSCo system architects and engineers to understand the main data flows that must be managed (according to the requirements of the present Data Management Plan) and that influence the DiSSCo design. Thus, the DFDs only show the 'normal' data flows and not data flows arising from error or other situations. For example, where poor quality of images dictates that a specimen must be re-imaged, the data flows associated with this remedial activity are not shown. Similarly, data flows associated with operational matters of the infrastructure intended to secure the robustness and reliability of the infrastructure (for example, temporary backup protection for digital images covering the time between imaging and deposition in a trustworthy repository) are also not shown.

Two levels of DFD are given:

1. Top-level DFD (section C.2 below) illustrating the broad arrangement of the two main subsystems of DiSSCo (Digitisation line/factory and DiSSCo data management) in relation to the main external entities with which DiSSCo interacts; and,
2. Second-level DFDs (section C.3 below) illustrating the main activities of digitisation (section C.3.1) and management/use of digital specimens (section C.3.2).

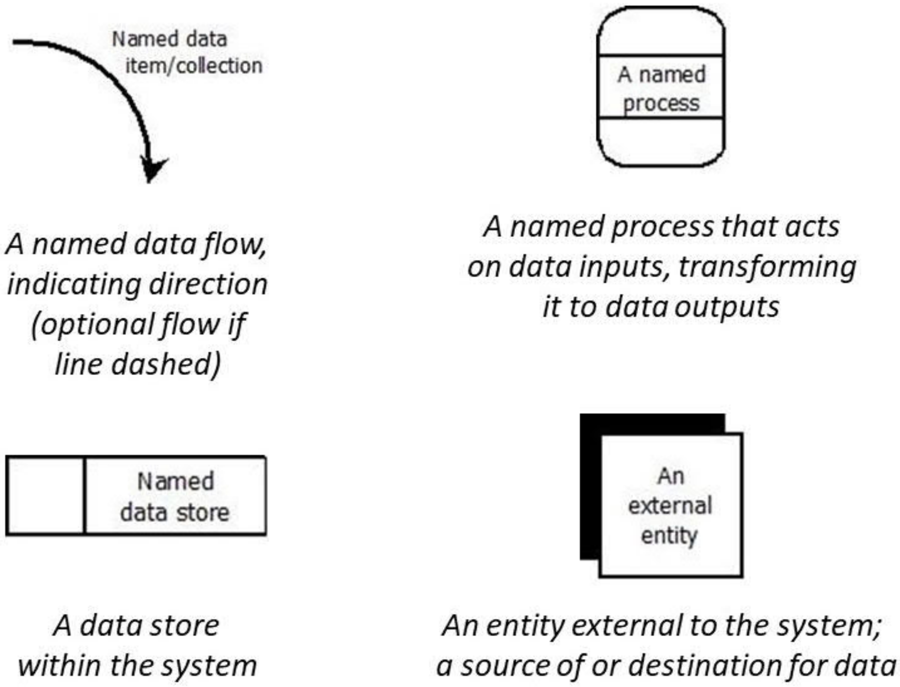
Note 27: A third level of DFD is possible but too detailed for the present Data Management Plan.

For further study. Detailed data flows between DiSSCo Facilities and DiSSCo Hub, as depicted by the 'Digitised specimens' flow in Figure C.1.

[27] https://www.idigbio.org/content/workflow-modules-and-task-lists; for flat sheets and packets of plants, algae, fungi, for example see doi: 10.3732/apps.1500065.

C.1.2 Symbols

Data flow diagrams use four symbols, as illustrated with their meanings below.



C.2 Top-level Data Flow Diagram

The top-level DFD (Figure C.1) illustrates the two main aspects (processes) of DiSSCo, namely digitisation and managing digital specimen data, and the relations between these and four entities external to the infrastructure. The external entities are the key roles of curator and scientist; the governance entity prioritising digitisation activities and multiple, various 3rd-party information sources/sinks.

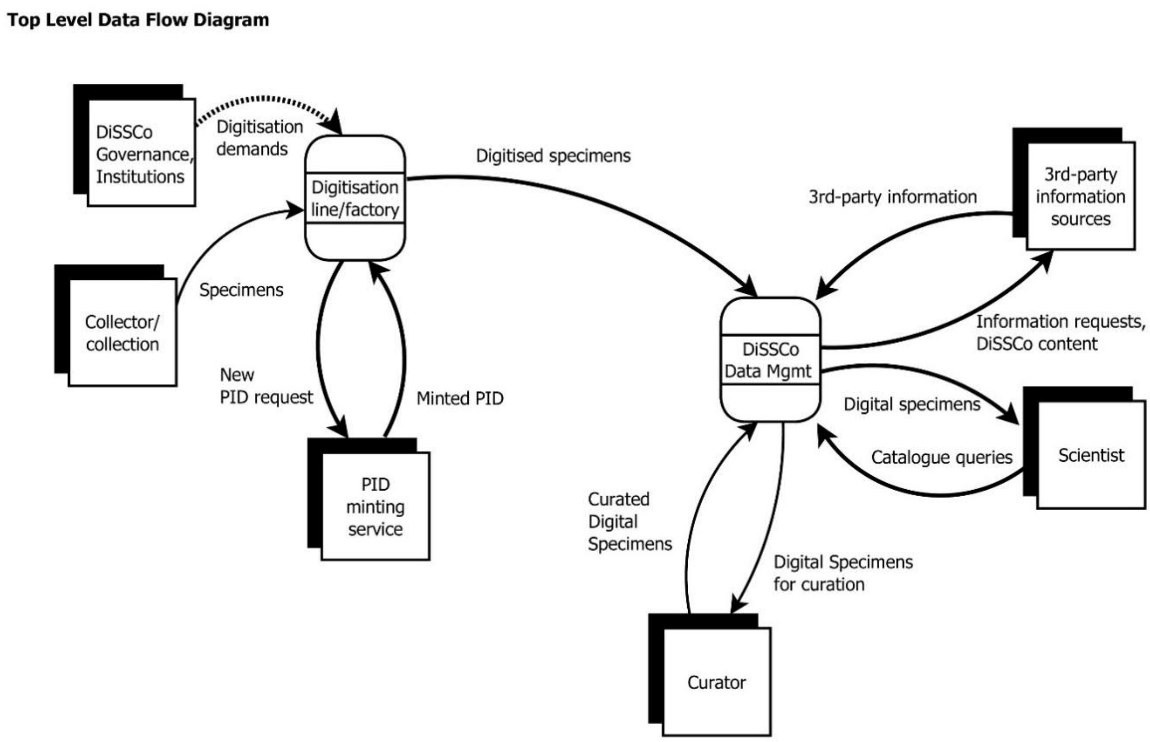


Figure C.1: Top-level DFD for DiSSCo

C.3 Second-level Data Flow Diagrams

C.3.1 Digitisation line/factory

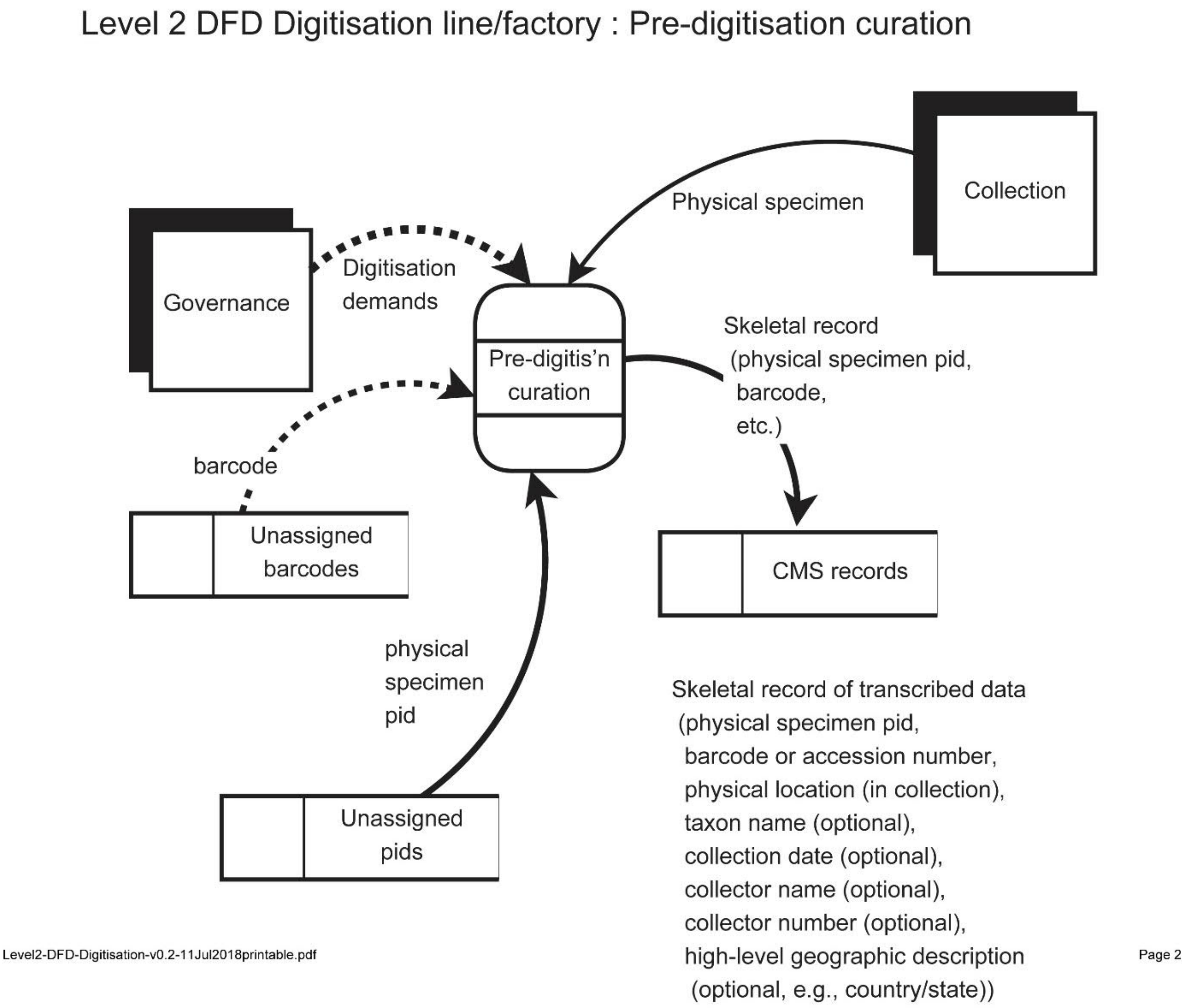


Figure C.2: Level 2 DFD Digitisation line/factory – pre-digitisation curation

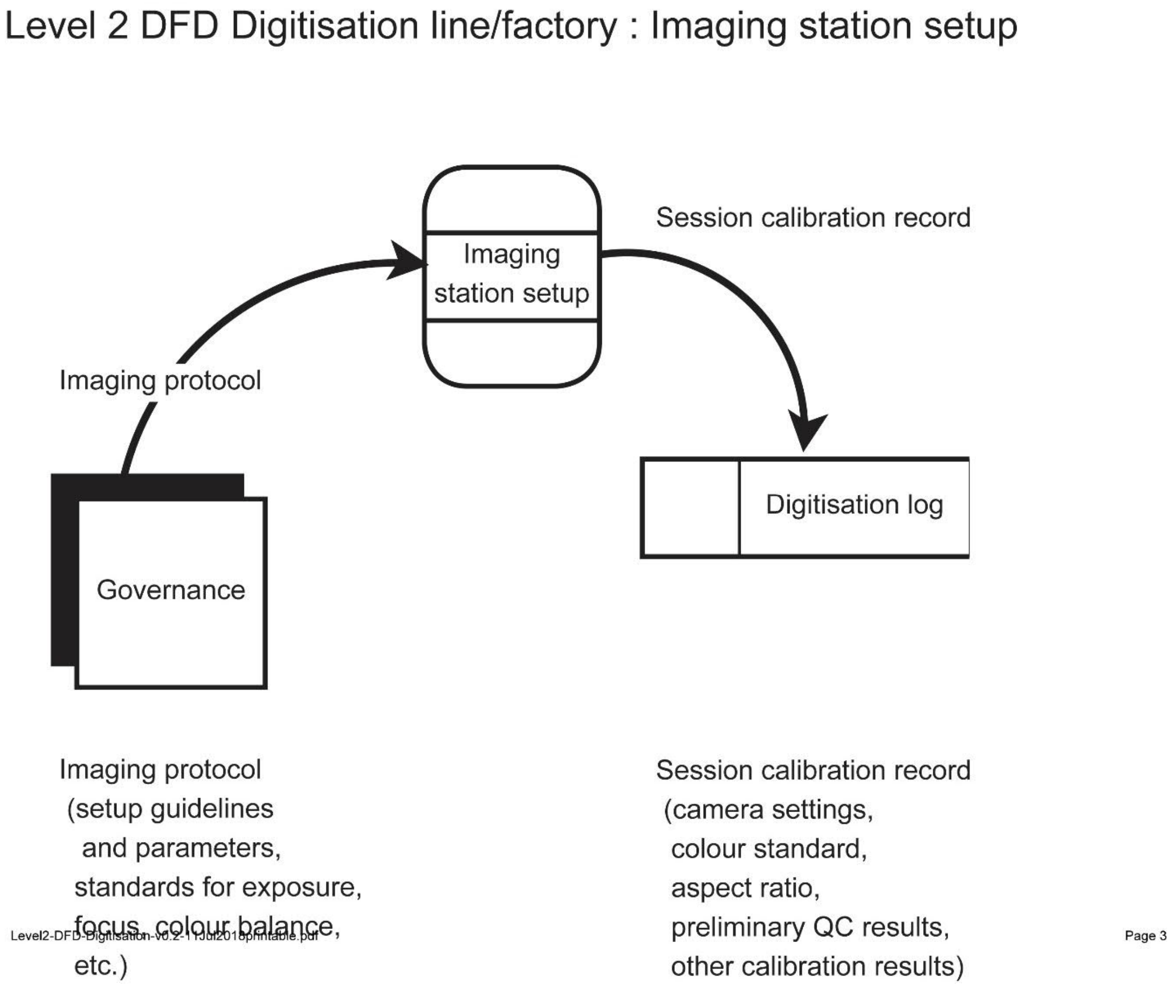


Figure C.3: Level 2 DFD Digitisation line/factory – imaging station setup

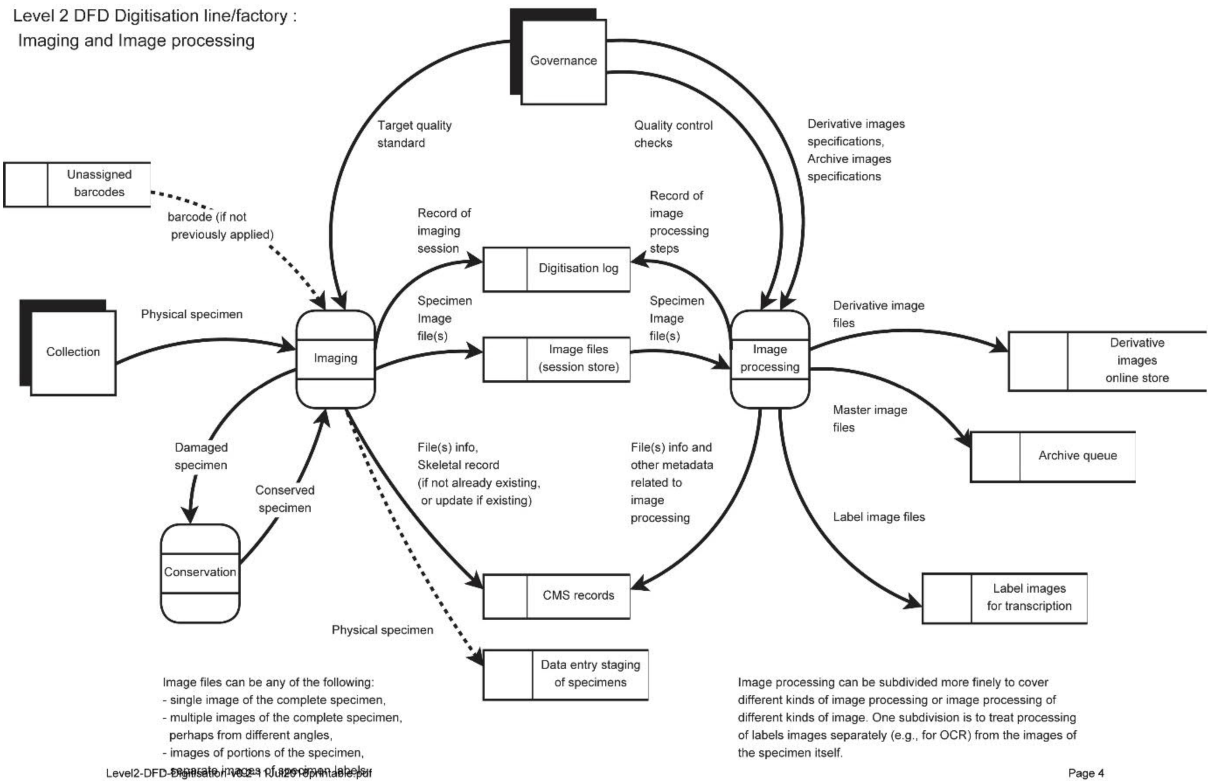


Figure C.4: Level 2 DFD Digitisation line/factory – imaging and image processing

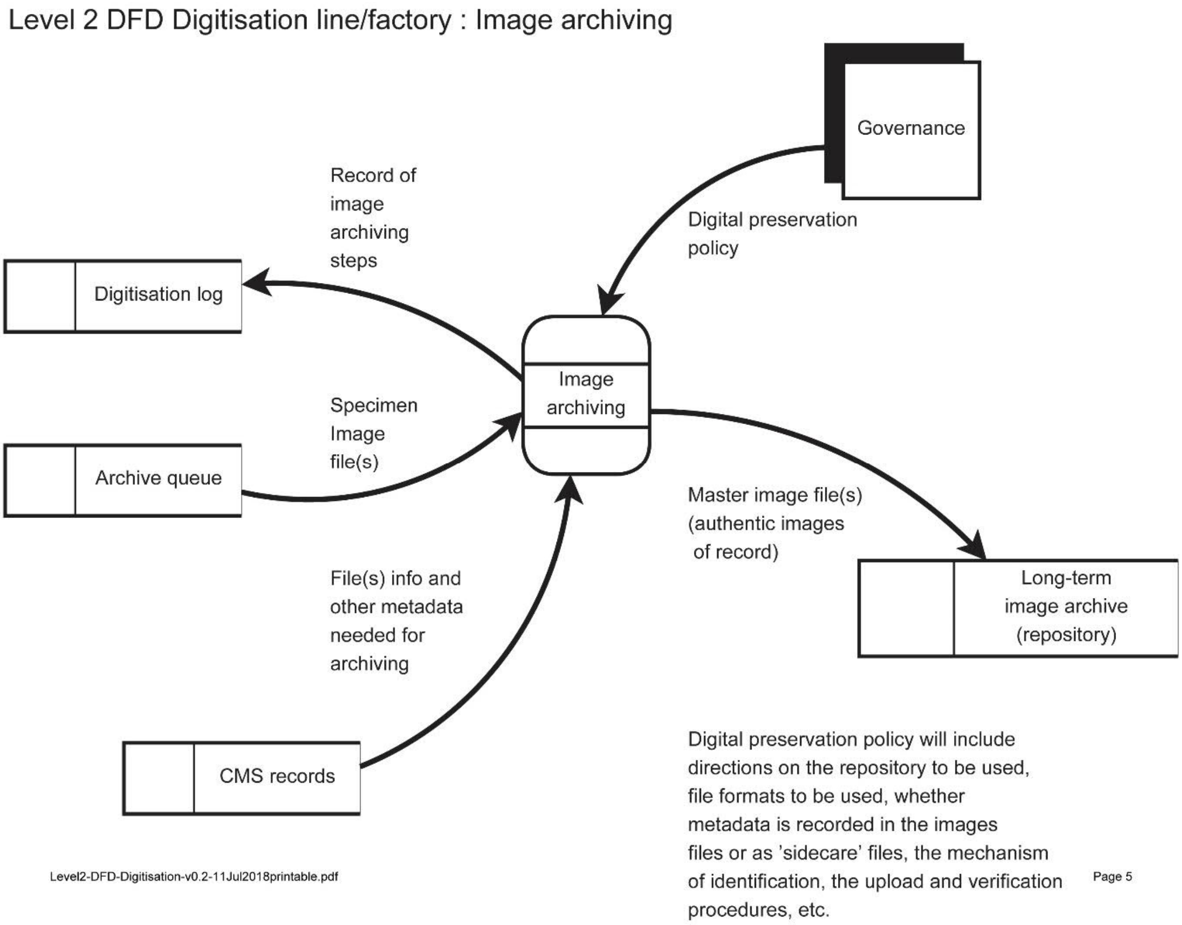


Figure C.5: Level 2 DFD Digitisation line/factory – image archiving

Note: After image processing there is an activity/process of image archiving and/or digital asset management. Typically, such a process is concerned with preparing images for long-term preservation by, for example converting the image to a format that is sustainable for long-term preservation, adding preservation metadata, etc. For DiSSCo design, it is sufficient to know the references and location of the archive images for any given specimen and how to obtain them when necessary. This information must be embedded in the DSO for the specimen.

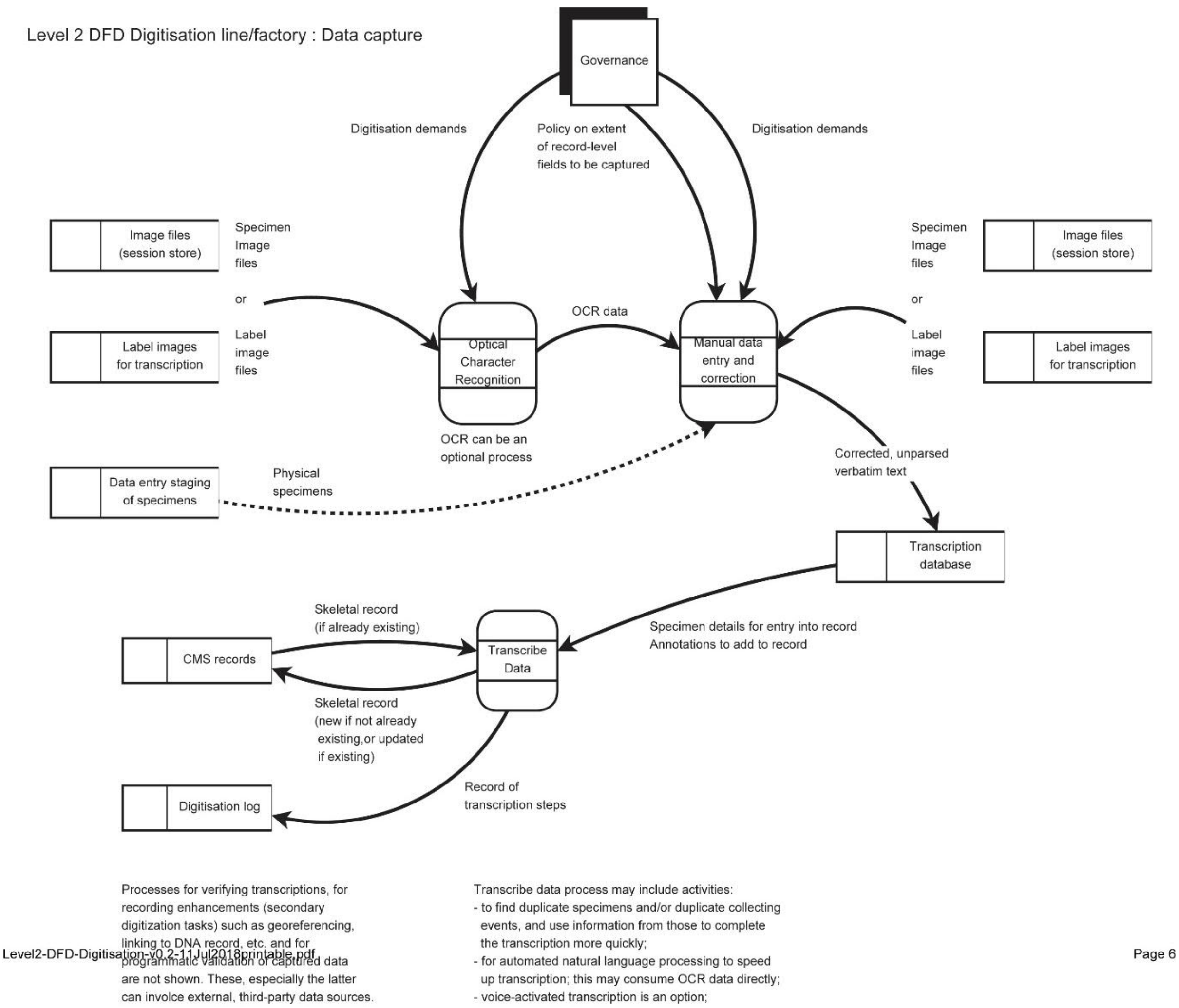


Figure C.6: Level 2 DFD Digitisation line/factory – data capture

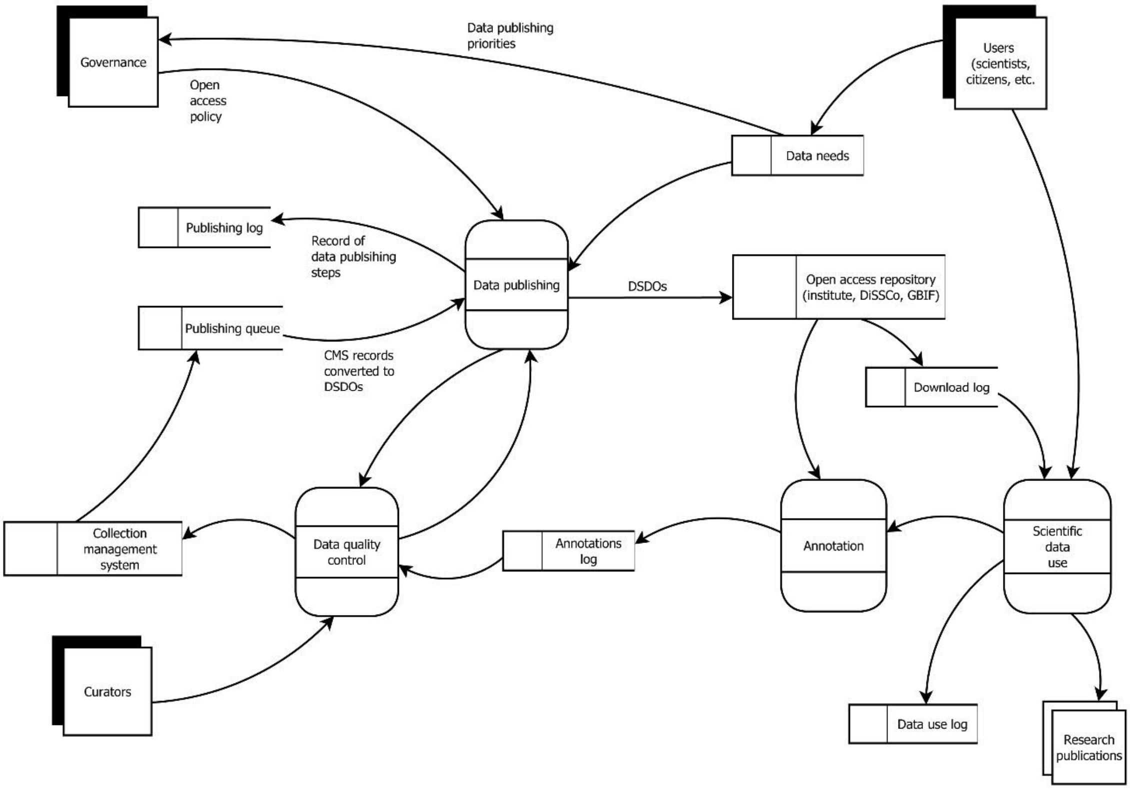


Figure C.7: Level 2 DFD Digitisation line/factory – data publishing

C.3.2 DiSSCo Data management

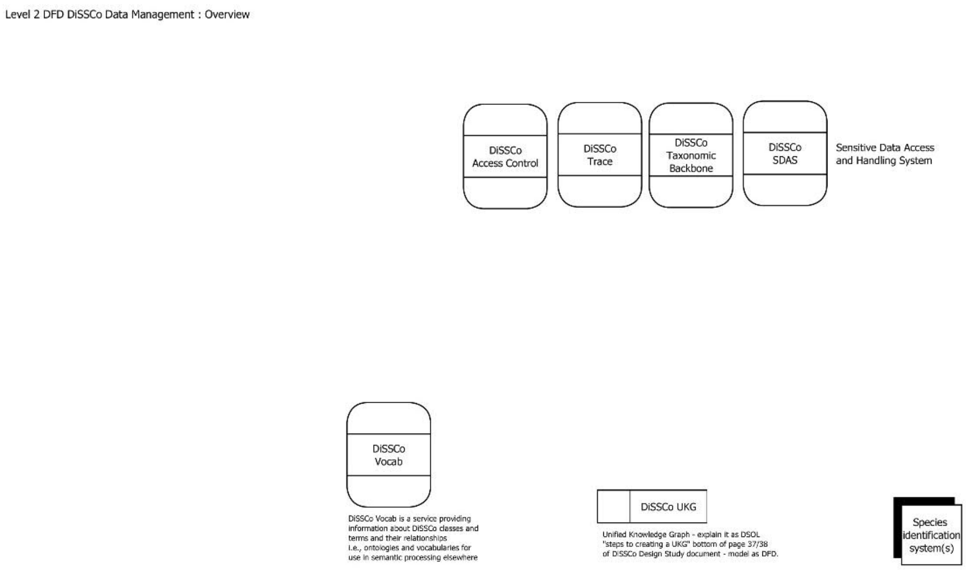


Figure C.8: Level 2 DFD Data management – overview

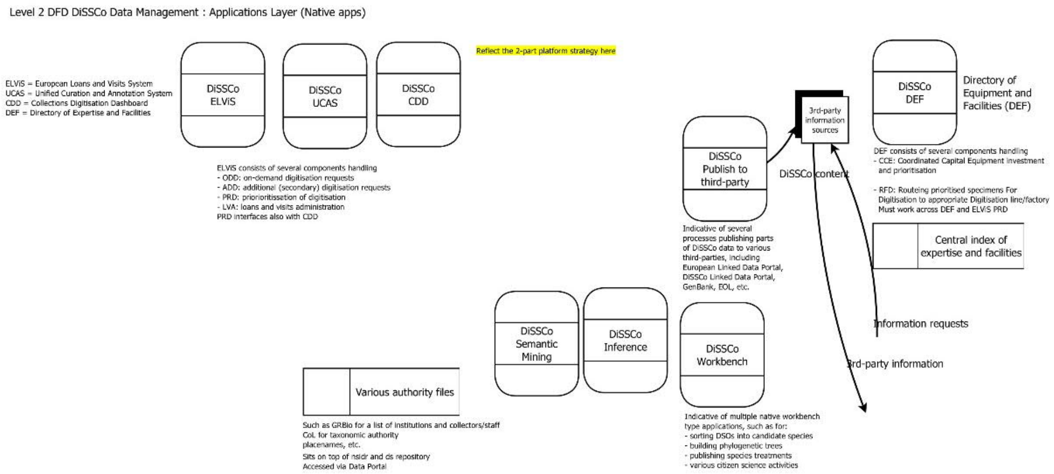


Figure C.9: Level 2 DFD Data management – applications layer (native apps)

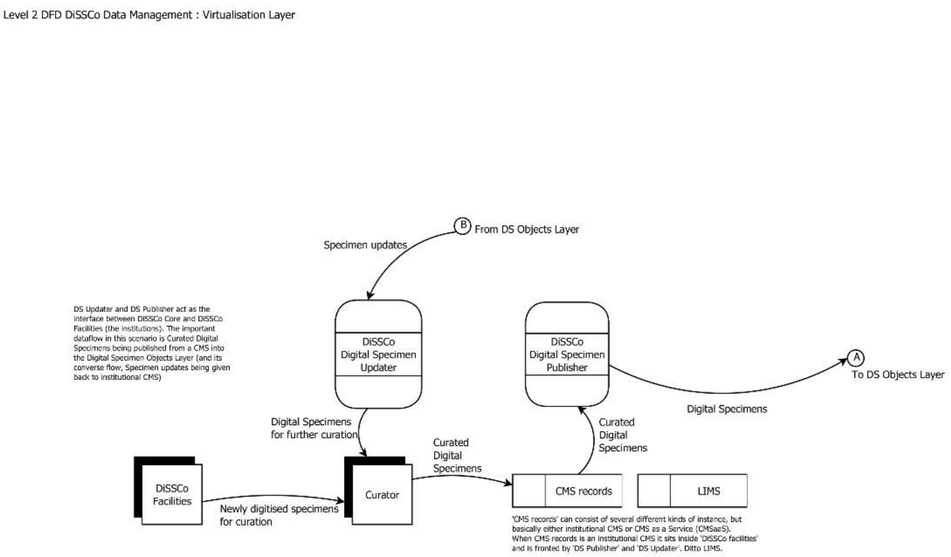


Figure C.10: Level 2 DFD Data management – virtualisation layer

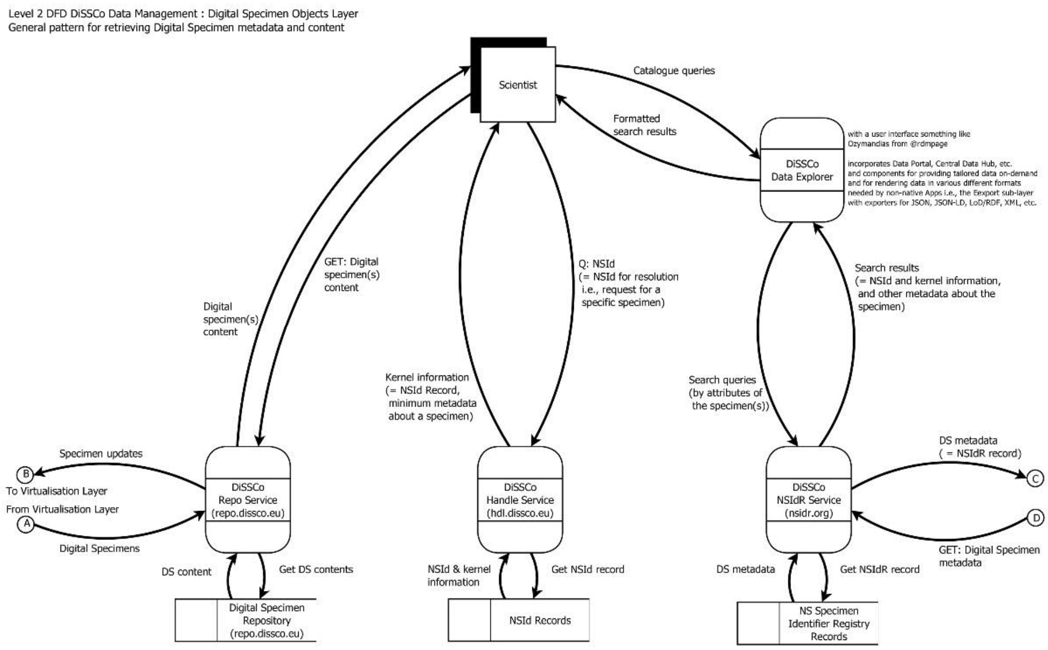


Figure C.11: Level 2 DFD Data management – digital specimen objects layer

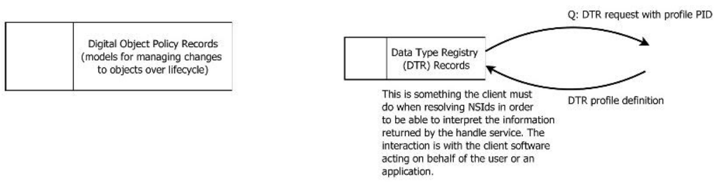


Figure C.12: Level 2 DFD Data management – digital specimen objects layer (continued)

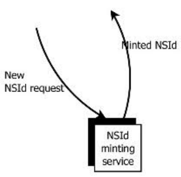


Figure C.13: Level 2 DFD Data management – minting

Appendix D: Estimates of expected volumes of data

D.1 Introduction

This appendix contains estimates of the volumes of data expected for each of the data categories identified in section 5.2.

D.2 Estimates

For further study. To be added by DiSSCo Prepare project.

Indicate the volumes of the image and non-image data. In the latter case, metrics about the expected size of the knowledge graph can help to inform various decisions. Knowledge graphs have 15-20 specific links to other data. Together with rates of digitisation it becomes possible to estimate the numbers and growth.

Appendix E: DiSSCo implementation of FAIR principles

E.1 Introduction

This appendix contains a summary statement of DiSSCo's implementation of the FAIR Guiding principles [Wilkinson 2016, Mons 2017]. Refer to section 6 (page 25) for full details.

The FAIR principles are to be Findable, Accessible, Interoperable and Reusable, and together these constitute one of the protected characteristics of the DiSSCo infrastructure – 'FAIRness' (section4.3, page 12).

E.2 To be Findable

FAIR principle F1: (meta)data are assigned a globally unique and persistent identifier.

- A handle is issued to each object published in or by DiSSCo, allowing the object data to be found regardless of its location.

FAIR principle F2: data are described with rich metadata (defined by R1 below).

- DiSSCo does not specifically distinguish between data and metadata, allowing most data fields to be used also as metadata for discovery.
- Digital specimen and digital collection data mainly take the form of 'named_attribute : value' pairs, with attribute names and definitions being sourced from appropriate standard vocabularies and schema where possible.

FAIR principle F3: metadata clearly and explicitly include the identifier of the data it describes.

- The NSId (Handle) is a top-level and mandatory field in the data of each DiSSCo object type and is used wherever an object or reference to an object appears.

FAIR principle F4: (meta)data are registered or indexed in a searchable resource.

- Data of each object type is indexed and searchable directly in the relevant DiSSCo index (the European Collection Objects Index, European Loans and Visits, and Unified Curation and Annotation System) thus allowing relevant identifiers (handles) to be found when these are unknown.
- Kernel information of each object is sent to the relevant Local Handle Service servers during handle registration.

E.3 To be Accessible

FAIR principle A1: (meta)data are retrievable by their identifier using a standardized communications protocol.

- Data for individual objects are retrievable by the object's handle using the Digital Object Interface Protocol (DOIP) version 2.0. Data is also retrievable through the REST (HTTP) API.

FAIR principle A1.1: the protocol is open, free, and universally implementable.

- DOIP and REST HTTP are open, free and universal protocols for information retrieval on the Web.

FAIR principle A1.2: the protocol allows for an authentication and authorization procedure, where necessary.

- Data are publicly accessible under a regime of 'as open as possible, as closed as legally necessary'. They are licensed under public domain. Appropriate authorization is necessary to retrieve data that is legally closed according to objective criteria.

FAIR principle A2: metadata are accessible, even when the data are no longer available.

- Data is retained for the lifetime of DiSSCo and its member organisations, which is currently expected to be for many decades.
- Data are stored in high-availability database servers of DiSSCo and its member organisations.

E.4 To be Interoperable

FAIR principle I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- DiSSCo uses JSON Schema as both internal and external representation of data and offers export to several popular formats such as Darwin Core Archive, ABCD XML documents and RDF.

FAIR principle I2: (meta)data use vocabularies that follow FAIR principles.

- DiSSCo refers to open, external vocabularies for terms used, e.g.: OBO Foundry Biological Collections Ontology, ABCD + EFG, Darwin Core, Dublin Core, etc.

FAIR principle I3: (meta)data include qualified references to other (meta)data.

- Each referenced external term is qualified by a resolvable context linking the term's properties to concepts in a relevant external ontology. i.e., JSON-LD context statements are using within JSON Schemas and Documents.

E.5 To be Reusable

FAIR principle R1: (meta)data are richly described with a plurality of accurate and relevant attributes.

- Each object contains a minimum of mandatory terms consistent with its formal object type definition, with the possibility to include optional additional terms and enrichments as necessary.
- In the case of Digital Specimen and Digital Collection object types, the minimum of mandatory terms corresponds to the object's classification as representing a specific level of digitization according to (respectively) the Minimum Information standard for Digital Specimens (MIDS) and the Minimum Information standard for Digital Collections (MICS).

FAIR principle R1.1: (meta)data are released with a clear and accessible data usage license.

- License is a mandatory term in Digital Specimen and Digital Collection objects and normally specifies either: i) there are no rights reserved by the copyright holder i.e., CC0; or ii) the copyright holder requires attribution i.e., CC-BY.
- Data retrieved and used by the users is subject to the license specified.

FAIR principle R1.2: (meta)data are associated with detailed provenance.

- All data published/uploaded in DiSSCo is traceable to a registered DiSSCo user (the creator).

- Data describe the original collectors/authors of the published data.

FAIR principle R1.3: (meta)data meet domain-relevant community standards.

- DiSSCo is domain-specific for the natural sciences community, yet through adherence with relevant standard vocabularies, schemas and file formats from outside its domain, DiSSCo data is broadly usable by a wide cross-section of users across multiple domains.

Appendix F: Standards applicable for data management

F.1 Introduction

Here are listed in summary the standards (technical specification, policies, etc.) that are expected to be implemented and complied with in support of implementing the DiSSCo Data Management Plan (the present document). In the list that follows, cross-reference is made to other sections of the DMP where further details are given.

F.2 List of standards

For further study. To be added.

Appendix G: DMP Compliance Checklist

G.1 Introduction

This appendix contains a checklist that DiSSCo constituents can use to assess their achievement of best practice data management in accordance with the requirements of the present data management plan document.

G.2 Checklist

For further study. To be added.

END.

Index

Page numbers in bold indicate the glossary entry for a term definition.

E

F

END.