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 D8.1 / [R] / [PU]
Work package 8 / Lead: CU
Delivery date M27
DOI: *To be assigned by Pensoft.*

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Conceptual design blueprint for the DiSSCo digitization infrastructure

DELIVERABLE D8.1
v0.22, final draft for submission to EC and RIO journal article conversion

Alex Hardisty (Cardiff University)
Hannu Saarenmaa (University of Helsinki),
Eva Alonso (DiSSCo CSO)
Ana Casino (CETAF)
Mathias Dillen (Botanic Garden, Meise)
Karsten Gödderz (CETAF)
Hilary Goodson (Natural History Musuem, London)
Quentin Groom (Botanic Garden, Meise)
Helen Hardy (Natural History Musuem, London)
Dimitris Koureas (Naturalis Biodiversity Center)
Abraham Nieva de la Hidalga (Cardiff University)
Deborah Paul (Florida State University)
Veljo Runnel (University of Tartu)
Xavier Vermeersch (CETAF)
Myriam van Walsum (Picturae)
Luc Willemse (Naturalis Biodiversity Center)

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
31 March 2020, v0.22	**As listed on cover page.**	**Final draft for submission to EC and conversion to RIO journal article.**

Blank page

Table of Contents

List of Tables

List of Figures

Blank page

Executive summary

DiSSCo, the Distributed System of Scientific Collections, is a pan-European Research Infrastructure (RI) mobilising, unifying bio- and geo-diversity information connected to the specimens held in natural science collections and delivering it to scientific communities and beyond. Bringing together 120 institutions across 21 countries and combining earlier investments in data interoperability practices with technological advancements in digitisation, cloud services and semantic linking, DiSSCo makes the data from natural science collections available as one virtual data cloud, connected with data emerging from new techniques and not already linked to specimens. 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.), to name but a few; and will lead to a wide range of end-user services that begins with finding, accessing, using and improving data. DiSSCo will deliver the diagnostic information required for novel approaches and new services that will transform the landscape of what is possible in ways that are hard to imagine today.

With approximately 1.5 billion objects to be digitised, bringing natural science collections to the information age is expected to result in many tens of petabytes of new data over the next decades, used on average by 5,000 – 15,000 unique users every day. This requires new skills, clear policies and robust procedures and new technologies 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, realising the core principles of 'findable, accessible, interoperable and reusable' (FAIR).

Synthesised from results of the ICEDIG project ("Innovation and Consolidation for Large Scale Digitisation of Natural Heritage", EU Horizon 2020 grant agreement No. 777483) the DiSSCo Conceptual Design Blueprint covers the organisational arrangements, processes and practices, the architecture, tools and technologies, culture, skills and capacity building and governance and business model proposals for constructing the digitisation infrastructure of DiSSCo. In this context, the digitisation infrastructure of DiSSCo must be interpreted as that infrastructure (machinery, processing, procedures, personnel, organisation) offering Europe-wide capabilities for mass digitisation and digitisation-on-demand, and for the subsequent management (i.e., curation, publication, processing) and use of the resulting data. The blueprint constitutes the essential background needed to continue work to raise the overall maturity of the DiSSCo Programme across multiple dimensions (organisational, technical, scientific, data, financial) to achieve readiness to begin construction.

Today, collection digitisation efforts have reached most collection-holding institutions across Europe. Much of the leadership and many of the people involved in digitisation and working with digital collections wish to take steps forward and expand the efforts to benefit further from the already noticeable positive effects. The collective results of examining technical, financial, policy and governance aspects show the way forward to operating a large distributed initiative i.e., the Distributed System of Scientific Collections (DiSSCo) for natural science collections across Europe. Ample examples, opportunities and need for innovation and consolidation for large scale digitisation of natural heritage have been described. The blueprint makes one hundred and four (104) recommendations to be considered by other elements of the DiSSCo Programme of linked projects (i.e., SYNTHESYS+, COST MOBILISE, DiSSCo Prepare, and others to follow) and the DiSSCo Programme leadership as the journey towards organisational, technical, scientific, data and financial readiness continues.

Nevertheless, significant obstacles must be overcome as a matter of priority if DiSSCo is to move beyond its Design and Preparatory Phases during 2024. Specifically, these include:

Organisational:

- Strengthen common purpose by adopting a common framework for policy harmonisation and capacity enhancement across broad areas, especially in respect of digitisation strategy and

prioritisation, digitisation processes and techniques, data and digital media publication and open access, protection of and access to sensitive data, and administration of access and benefit sharing.

- Pursue the joint ventures and other relationships necessary to the successful delivery of the DiSSCo mission, especially ventures with GBIF and other international and regional digitisation and data aggregation organisations, in the context of infrastructure policy frameworks, such as EOSC. Proceed with the explicit aim of avoiding divergences of approach in global natural science collections data management and research.

Technical:

- Adopt and enhance the DiSSCo Digital Specimen Architecture and, specifically as a matter of urgency, establish the persistent identifier scheme to be used by DiSSCo and (ideally) other comparable regional initiatives.
- Establish (software) engineering development and (infrastructure) operations team and direction essential to the delivery of services and functionalities expected from DiSSCo such that earnest engineering can lead to an early start of DiSSCo operations.

Scientific:

- Establish a common digital research agenda leveraging Digital (extended) Specimens as anchoring points for all specimen-associated and -derived information, demonstrating to research institutions and policy/decision-makers the new possibilities, opportunities and value of participating in the DiSSCo research infrastructure.

Data:

- Adopt the FAIR Digital Object Framework and the International Image Interoperability Framework as the low entropy means to achieving uniform access to rich data (image and non-image) that is findable, accessible, interoperable and reusable (FAIR).
- Develop and promote best practice approaches towards achieving the best digitisation results in terms of quality (best, according to agreed minimum information and other specifications), time (highest throughput, fast), and cost (lowest, minimal per specimen).

Financial

- Broaden attractiveness (i.e., improve bankability) of DiSSCo as an infrastructure to invest in.
- Plan for finding ways to bridge the funding gap to avoid disruptions in the critical funding path that risks interrupting core operations; especially when the gap opens between the end of preparations and beginning of implementation due to unsolved political difficulties.

Strategically, it is vital to balance the multiple factors addressed by the blueprint against one another to achieve the desired goals of the DiSSCo programme. Decisions cannot be taken on one aspect alone without considering other aspects, and here the various governance structures of DiSSCo (General Assembly, advisory boards, and stakeholder forums) play a critical role over the coming years.

List of recommendations towards DiSSCo

The ICEDIG project consortium (see Acknowledgements section below) makes 102 recommendations to the DiSSCo Coordination and Support Office, the DiSSCo General Assembly and to projects of the DiSSCo programme. These are listed below. For each, a page number indicates where the recommendation is made and discussed in the body of text.

If you are reading this document on-screen, either in its original WORD document format or as its PDF version, hovering over and clicking a recommendation in the list below will take you to the correct place in the present document. In the PDF version, Alt-LeftArrow will then take you back to the previous view (i.e., this list of recommendations). In the WORD version, use the Headings section of the Navigation pane (Ctrl+F) and collapse the major headings (right click, collapse all) so you can see the 'List of recommendations' heading to click on to return you to this list.

Acknowledgements

The work reported and the recommendations made in the present document are the result of a concerted effort over the period January 2018 – February 2020 by numerous experts participating in and contributing to the ICEDIG project. This has led to many interim 'milestone' reports and deliverables that form the basis of the present blueprint and recommendations for DiSSCo.

This project, ICEDIG – "Innovation and consolidation for large scale digitization of natural heritage" has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777483. The ICEDIG consortium consists of 12 partners from 7 countries:

Finnish Museum of Natural History – LUOMUS, Helsinki, FI

Naturalis Biodiversity Center, Leiden, NL

National Museum of Natural History, Paris, FR

Botanic Garden Meise, Meise, BE

University of Tartu Natural History Museum and Botanical Garden, Tartu, EE

National Computing Center for Higher Education, Montpellier, FR

Natural History Museum, London, UK

Cardiff University, Cardiff, UK

Consortium of European Taxonomic Facilities, Brussels, BE

Picturae, Heerhugowaard, NL

Royal Botanic Gardens, Kew, Kew, UK

Plazi, Bern, CH

Blank page

1 Introduction

1.1 Background

Strategically, it is vital to balance multiple factors – technical and engineering, organisational and political, financial and legal, and operational and governance – against one another to achieve the desired goals of the DiSSCo programme. Decisions cannot be taken on one aspect alone without considering other aspects.

DiSSCo, the Distributed System of Scientific Collections, is a pan-European Research Infrastructure (RI) mobilising, unifying and delivering bio- and geo-diversity information connected to the specimens held in natural science collections to scientific communities and beyond. Bringing together 120 institutions across 21 countries and combining earlier investments in data interoperability practices with technological advancements in digitization, cloud services and semantic linking, DiSSCo makes the data from natural science collections available as one virtual data cloud in association with a wide range of end-user services that begins with finding data, accessing data, using data and improving data. New services, made possible by the expanding variety and volume of data, coupled with open access to that data and new ways of connecting and manipulating it, will transform the landscape of what is possible in ways that are hard to imagine today. In addition to serving as the framework for interpreting and validating species data, DiSSCo connects historical collection data – traditionally connected using scientific species names as primary identifiers – with data emerging from new techniques and not already linked to species names. 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.) to name but a few. 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.

With approximately 1.5 billion objects to be digitized, bringing natural science collections to the information age is expected to result in 90 petabytes of new data over the next 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 beginning of this section concerned with realising the core principles of 'findable, accessible, interoperable and reusable' (FAIR).

The present document, the DiSSCo Conceptual Design Blueprint is based on the results of the ICEDIG project work. It has been synthesised from the essential outcomes and conclusions of that project. These are mostly (but not all) reported in other deliverable documents. The present document constitutes the essential background needed by the DiSSCo Prepare project to carry out its work to raise the overall maturity of the DiSSCo Programme across multiple dimensions (organisational, technical, scientific, data, financial) to achieve readiness to begin construction.

1.2 Scope

The present document is a ‘blueprint’ report covering the technological design innovations, organisational consolidation, partnership, governance and business model proposals for constructing the digitization infrastructure of DiSSCo. In this context, the digitization infrastructure of DiSSCo must be interpreted as that infrastructure (machinery, processing, procedures, personnel, organisation) offering Europe-wide capabilities for mass digitization and digitization-on-demand, and for the subsequent management (i.e., curation, publication, processing) and use of the resulting data.

By mass digitization, we mean digitizing entire collections or their major distinct parts at industrial scale (i.e., millions of objects annually at low cost (e.g., < c.€0.50 per item), characterised by improved workflows, technological and procedural frameworks based on automation (both hardware and software) and enrichment (link-building). This is critical within DiSSCo to mobilise the data from collections as rapidly as possible, so that these data can be more easily found and used; and can act as an anchor or 'keyring' for other data.

Alongside mass digitization, DiSSCo also recognises the need for digitization-on-demand; by which we mean processes, procedures and technologies that respond to and support a request for digital data about specimens that have not yet been through a mass digitization process or project-based digitization workflow. Such requests may sometimes be supported by prioritisation within a mass digitization workflow, or they can help institutions to refine and pilot new mass workflows. But such requests can also demand more data than the 'normal workflow' to meet specific needs of the users in each case or to deal with digitization of the most complex specimens and preservation types. Digitization-on-demand is a critical channel for digital access as envisaged for this research infrastructure, to allow truly global access to and use of collections. Mass digitization and digitization-on-demand are treated in detail in section 3.

1.3 Structure of the document

The present document emphasises that the considerations in the following sections, and recommendations in section 7 are a balance of multiple factors in play (Figure 1). These are not solely technical and engineering but also include organisation and socio-political factors, financial and legal factors and operational and governance factors.

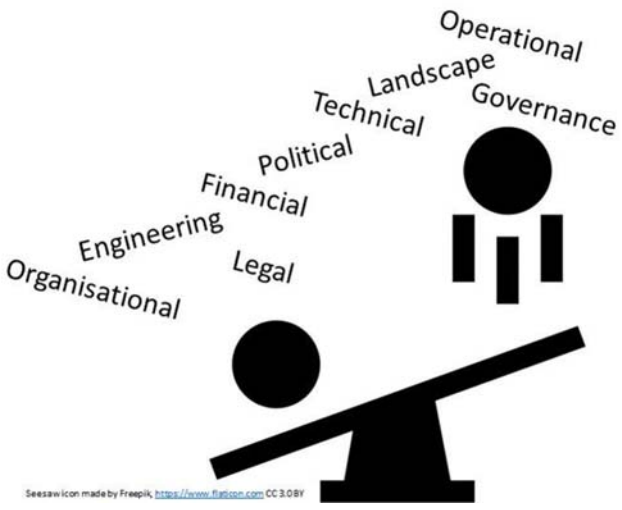


Figure 1: Multiple factors balance to achieve goals

Practically, this means focussing on the arrangements, processes and practices that need to be in place to achieve mass digitization, the architecture, tools and technologies to make digitization and management of the resulting data least-cost, and the culture, skills and capacity-building essential to operating in an efficient and sustainable manner. The present document is structured this way, and the recommendations coming from this practical focus contribute towards increasing overall implementation readiness[1] across four of the five dimensions of the infrastructure implementation i.e., data, technological, financial and organisation – only the scientific dimension is not of immediate concern for the present deliverable.

[1] Implementation readiness is measured by the Implementation Readiness Level (IRL), defined as the measure of the ability of the organisation (DiSSCo) to embark on specific implementation actions (construction project) based on clear, actionable guidelines with minimum risk and across the scientific, data, financial, technological and organisational dimensions of the infrastructure implementation. DiSSCo Prepare will act as the main vehicle through which DiSSCo RI will raise its overall maturity and set itself in a position to implement its construction programme by i) improving the overall IRL, and ii) delivering the DiSSCo Construction Masterplan. Meeting these two high-level objectives will ensure that DiSSCo embarks on its construction phase with minimal but well-understood risks, and a clear and detailed construction plan that leads to the commencement of operations of the infrastructure by 2025 (as currently projected).

1.4 Conventions in the present document

1.1.1 Key terms

The terms below are used with the meanings given:

Innovation: The alteration of what is established by the introduction of new elements or forms; a change made in the nature or fashion of anything; something newly introduced; a novel practice, method, etc.

Consolidation: The action of making solid, or of forming into a solid or compact mass (also figuratively); solidification; combination into a single body, or coherent whole; combination, unification.

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 digitization of and access to the data bound up in those collections.

DiSSCo Facility: 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.

A special kind of DiSSCo Facility could be a DiSSCo Centre of Excellence (DCE), specialised in one or more of researching, innovating, developing and operating/performing techniques and/or processes of digitization or other related facets, and disseminating information on same.

For the meaning of other terms and abbreviations, see the glossary (section 9, page 128).

1.1.2 Recommendations of the report

Main recommendations of the report are contained in double-bordered boxes, like so:

Recommendation N: The words of the recommendation.

Supplementary information not integral to the recommendations of the present report but of interest to the reader is contained in single-bordered boxes, like so:

Supplementary information. The words of the supplementary information.

1.1.3 References to other documents

References to other documents include references to ICEDIG project deliverable documents of the form "ICEDIG project deliverable Dx.y". These are referred to via footnotes and are listed in full in section 8.1.

References to other documents are of the form [firstauthor year]. These are listed in section 8.2.

2 The DiSSCo research infrastructure

2.1 Rationale for a Distributed System of Scientific Collections

2.1.1 Digitizing natural science (biological, geological) collections

Digital transformation of society affects all areas of human activity, and science is at the forefront of this development: "*Digital science means a radical transformation of the nature of science and innovation due to the integration of ICT in the research process and the internet culture of openness and sharing*". Emphasis on sharing the results and data of publicly funded research has led to the concept of "Open

Science" [EC 2013, EC 2015]. At the European level we are witnessing the coming together of open science with the digital single market as a movement towards European strategic autonomy for the data economy. This leads eventually to consolidation in research, public sector and commercial sectors of the economy.

Natural science collections are an integral part of the global natural and cultural capital. They include 2-3 billion animal, plant, fossil, rock, mineral, and meteorite specimens. The European collections account for 55% of the natural sciences collections globally and represent 80% of the world's bio- and geo-diversity. Data derived from these collections underpin countless innovations, including tens of thousands of scholarly publications and official reports annually (used to support legislative and regulatory processes relating to health, food, security, sustainability and environmental change); inventions and products critical to our bio-economy; databases, maps and descriptions of scientific observations; instructional material for students, as well as educational material for the public. Natural science collections, which exist in all the world's countries, are some of the oldest Research Infrastructures (RI). Their collections have always been open for all scientists and form the hard core of biodiversity science that studies the existence of life on earth and geoscience that studies the earth itself. Life sciences pioneered open access through establishing the GenBank in 1982. In 2001, the OECD Megascience Forum established the Global Biodiversity Information Facility (GBIF) to share biodiversity data freely and openly. Initially, GBIF focussed on digitization of biological collections, but is now an important source of any kind of primary biodiversity observation data.

Digitizing biological collections, which have been gathered for more than 250 years, is a gargantuan task. Europe's collections are home to about one and a half billion specimens, of which about 10% have been digitally catalogued. For the remaining collections, only the physical objects exist, containing – besides the biological material itself – highly useful biodiversity data attached to each specimen on paper labels. Only about 1-2% of the objects have been imaged. User demand for the available data is high. For example, in 2016, Royal Botanic Gardens, Kew's herbarium catalogue (which at that time had over 900,000 specimen records available) was visited over 125,000 times, with more than 960,000 page-views.

2.1.2 Accelerating beyond the current situation

The current situation hinders modern science, as contemporary research requires access to data digitally to address some of the biggest challenges of our time e.g., [Suarez 2004, McCulloch 2013]. For instance, we need to understand the impact of global change and biodiversity loss, and the effect of climate change on ecosystems. Large datasets are being compiled to model and analyse these effects. There is a lot of modern observation data available, but only digitized collections provide the baseline over time for which organisms existed where and when. This provides an auditable basis for modelling their future distributions under different scenarios. Biodiversity loss and extinctions, and the associated loss of ecosystem function, are leading to a situation where a substantial part of the world's biodiversity and critical ecosystem services are lost even before we understand their value. We need to accelerate the discovery of species, which can only be done by accelerating scientific cooperation, data sharing, and more effective use of biological collections.

One part of the solution is already in place. The e-infrastructure for accessing biodiversity data already exists through GBIF, as well as through other related initiatives such as the Catalogue of Life (CoL), the Biodiversity Heritage Library (BHL), the Encyclopedia of Life (EOL), the International Barcode of Life (iBOL), and the Biodiversity Information Standards organisation (TDWG). These global organisations each tackle their own special data types and services. For instance, GBIF gives harmonised access to over 50,000 datasets with more than 1.3 billion biodiversity records[2], of which about 167 million records represent specimens from museum collections worldwide. However, the current e-infrastructures are not covering

[2] As of January 2020.

the entire field, since many kinds of data (including trait data, ecological observation data, and geological data) are not connected.

2.1.3 Disconnected infrastructure

There is a serious disconnect between the e-infrastructure and the physical infrastructure. Museums do have keen interest in digitizing their collections [Ang 2013, Balke 2013]. However, the sheer magnitude of the task of digitizing collections is daunting. For example, with traditional methods, working one specimen at a time, one person can image and completely digitize the data associated with 50 specimens in a working day, with a basic cost of about €5 per specimen. It would thus take 100,000 person years to digitize one billion specimens, with a cost of €5 billion. Organising such a task has been beyond the capacity of museums. What is needed is the transformation of a dispersed and fragmented access model to an integrated data-driven RI that will bring the natural science collections into the information age. This new RI will unify access to the European collections and facilitate innovations and consolidations that streamline digitization of physical collections in an industrial fashion and scale. This new RI appears as the Distributed System of Scientific Collections (DiSSCo; www.dissco.eu) on the ESFRI 2018 Roadmap [ESFRI 2018].

2.1.4 Industrialising digitization

There is evidence that industrialising digitization of natural science collections is achievable [Beaman 2012, Blagoderov 2012, Heerlien 2013, Tegelberg 2014, Rogers 2016]. The work demands and costs can be brought down at least by factors of 5-10 compared to manual methods [Oever 2012]. Efficient workflows can be designed when all material is first imaged and followed by the subsequent transcription of data from the images [Lehtonen 2011, Mantle 2012, Nelson 2012]. Large scale imaging, however, requires both technical innovation and dedicated environments. However, the latter are not always straightforward to build within museums. It requires a new mode of operation for collections managers, as the physical work of digitization and related steps, such as receiving and imaging of endowed collections, would need to be performed by dedicated infrastructures, even factories [Tegelberg 2012]. The non-physical aspects (i.e., transcribing data) need more automation [Tulig 2012, Barber 2013, Drinkwater 2014] and any continuing manual activities may need to be distributed amongst a globally organised workforce. The storage of petabytes of image data, analysis of images of millions of specimens, and complementing their metadata with other details requires massive computing power and the mining of data from big repositories.

Pioneered by the Herbarium of Muséum National d'histoire Naturelle (MNHN Paris), outsourced digitization activities have been undertaken in several innovative programmes, including those of Digitarium in Finland, Plantentuin Meise in Belgium, Naturalis in the Netherlands, and across France (ReColNat project) and Norway. Further major initiatives are being proposed, for instance in Germany, Sweden, and the UK, and similar approaches are also used in the USA for mass digitization, for example by the Smithsonian Museum of Natural History. Both inhouse and outsourcing approaches at scale have now been implemented and tried across several projects.

Total spending in known European projects is in the range of €10-20 million annually. In the USA, the Advancing the Digitization of Biological Collections (ADBC) program receives $10 million annually from the National Science Foundation for nationally distributed digitization of non-federal collections. Collections in the USA write grants to compete for these digitization funds. iDigBio receives about a fifth of those yearly funds to aggregate the digitized data in a central infrastructure. In addition, iDigBio works collaboratively with the worldwide community to build capacity development for digitization and data use, foster research use of these data, develop enhanced education and outreach materials using collections data, and lead efforts to increase inclusivity in museum collections. The majority rely on major public investments. A problem is that these projects are funded in the short-term. Sustainability can only be achieved when the results are anchored in long-lived RIs, and the new ways of work are internalised by stakeholders. These initiatives will need a common infrastructure to effectively share their

technologies, data, and experiences, and jointly build the most demanding functions. This will further accelerate the process and bring in more resources from more countries, when the task becomes tangible. Furthermore, it is crucial to develop strong links between physical and virtual access programmes. Complementarity between physical and virtual access at European level enhances the overall capacity of researchers to discover and retrieve relevant information.

However, in most countries there are no systematic mass digitization programmes. There is lack of funding, lack of skills, workflows able to cope only with low throughput, and lack of suitable ICT systems. This makes the unit cost of digitization too high for rapid mass digitization. The existence of a large research infrastructure that tackles digitization could change this. In other words, to be viable, digitization requires large volumes to become more affordable.

2.1.5 Understanding digitization

It's evident there are many understandings of what is meant by digitization, beyond the fundamental definition of being the process of converting analog information about physical specimens to digital format (which includes electronic text, images and other representations). Does it mean creating a database record in a collection management system (CMS)? Does it mean producing one or more photographs of a specimen, and if so, with what characteristics? Does it include transcription of label data into database records? Does it include more than just making the 'what, where, when and who' of the gathering event available as digital data? Different digitization initiatives can (and typically do) have varying aims and scope, leading to quite different outcomes in the characteristics of the resulting data sets. This can lead to different uses and benefits and therefore different assessments of usefulness (fitness for purpose) and cost-effectiveness. Small scale, deep digitization can certainly be cost-effective (i.e. offer good value for money in relation to project aims), but it may not be affordable at scale.

Additionally, as digitization proceeds and more varied uses are made of digital specimen data, we are beginning to find out more about what makes such data most useful. This itself is likely to affect what we consider essential data to digitize over time.

2.1.6 International landscape and DiSSCo positioning

DiSSCo data management and stewardship will take place in the context of a wide and extraordinarily varied and complex landscape of data generation and capture, management and use initiatives that extend from the local practices of individual researchers/groups, through institutional and national activities to extensive European and other regional initiatives that collectively make up the global landscape of biodiversity, ecological, and geo-environmental informatics. The range of stakeholders is broad, with many missions and interests, sometimes overlapping. Nevertheless, at present (late 2019) there is hardly any coordinated management of digital specimen and collection data outside of the historical and traditional institutions in all countries charged with caring for collections of physical natural science specimens i.e., the natural science collection-holding institutions such as national and local museums.

The activities of the Global Biodiversity Information Facility (GBIF), whose main mission is the mobilisation of primary biodiversity data are of great significance for DiSSCo. Here, DiSSCo eventually acts as the key stakeholder representing European collections that contribute primary occurrence data to GBIF based on specimens preserved in collections[3]. DiSSCo also acts as the European partner in developing a global

[3] Identifiable in GBIF data as dwc:basisOfRecord equals PreservedSpecimen, FossilSpecimen, LivingSpecimen, MaterialSample, etc.

commons for biodiversity information[4] in conjunction with the Alliance for Biodiversity Knowledge[5]. In the longer term, DiSSCo should take the responsibility to deliver data to GBIF on behalf of the DiSSCo collection-holding institutions, to avoid that institutions must publish separately to GBIF and DiSSCo.

Regional initiatives taking place elsewhere in the world are also of significance. These include:

- iDigBio/ADBC program[6] funded by the USA's National Science Foundation, for advancing the digitization of north American biodiversity collections. Already underway for several years, this initiative is acting out a ten-year plan for digitizing, imaging and mobilising collections data. In the process, much experience and wealth of technical knowledge, practice and procedures has been accumulated that is valuable for DiSSCo to draw upon as DiSSCo sets out its own digitization and data management plan objectives.
- National Specimen Information Infrastructure (NSII)[7] of China, underway since 2003; and,
- Australia's digitization of its national research collections (NRCA Digital)[8].

The importance of these initiatives extends beyond exploiting their experience and best practice. Global level alliances and coordination, especially global coordination around persistent identification of Digital Specimens and Digital Collections is essential and beneficial; for example, in working towards:

- A global biodiversity [Devictor 2016];
- Fulfilling commitments and obligations towards the Nagoya Protocol on Access and Benefit Sharing;
- Implementing the 'extended specimen' concept described by [Webster 2018] and [Lendemer 2019]; and,
- For managing 'Next Generation Collections' [Schindel 2018].

2.2 Innovations and consolidations identified by ICEDIG

Against the foregoing background, the general objective of the ICEDIG project has been to lay the ground work for such an organisation, namely DiSSCo, firstly by identifying and recommending the technological innovations that will be needed to efficiently digitize one and a half billion collection objects (including subsequent management of their data) in a foreseeable time, such as the next 30 years; and secondly to identify organisational consolidations needed to perform this task. When this has been achieved, the natural science community will be a fully enabled player in digital society, and the most fundamental scientific data on the diversity of our planet will be freely and openly available for all.

Table 1 – Table 4 below highlight the innovations and consolidations the various tasks of the ICEDIG project have identified and recommended in their detailed work in each of four areas of consideration. These areas are each considered in detail respectively in sections 3 – 6 below and are:

1) Arrangements, processes and practices;
2) Architecture, tools and technologies;
3) Culture, skills and capacity building; and,
4) Governance and business model.

In Table 1 – Table 4, the first column contains a brief description of the recommended innovation or consolidation. The column headed "I/C" indicates which is proposed. The column head "A/W" indicates

[4] A global 'biodiversity commons' is the notion of a defined community of use for some defined information space, with guaranteed free and unhindered access to data and information for that community in that information space (http://www.dlib.org/dlib/june02/moritz/06moritz.html).
[5] https://www.biodiversityinformatics.org/
[6] https://www.idigbio.org/.
[7] http://nsii.org.cn/2017/home.php.
[8] https://www.csiro.au/en/Research/Collections/Digitization.

with “A” whether the proposed innovation/consolidation has already been made/adopted, or with “W” whether it still needs to be further worked upon, put in place, actioned, etc. (for example, by the DiSSCo Prepare project). The next two columns respectively indicate the responsible work package in the ICEDIG project that proposed the item, and the work package in the forthcoming DiSSCo Prepare project that is expected to take up the item. The last comment provides for relevant remarks.

Table 1: Innovation and consolidation in arrangements, processes and practices

Description of the recommended innovation or consolidation	I/C	A/W	ICEDIG WP	DiSSCo Prepare WPs	Comment
DiSSCo related Policy Analysis	C	A	7		
Related research infrastructure landscape	C	A	7		
Alignment of related research infrastructure	C	W	7		
Procedures for developing standards and their review and adoption	I	W	4		
Principles for engaging citizen scientists	C	W	5		
Human transcription of specimen label data remains the most accurate way to obtain structured data for each specimen, though automated systems show potential in large volumes and for certain types of data	C	A	4		
Collection preparation for digitization is an essential element in the digitization process, though it is easily overlooked and underestimated	C	A	3		
Keep digitization processes as lean as possible. Accommodating all exceptions and other collection preservation tasks hinders the rate of digitization	C	A	3		
Use imaging of labels to later enrich MIDS level from 0-1 to 2-3 and data validation	C	A	3		See table 2 for entomology case.
DiSSCo should keep an overview of ongoing development and innovations in digitization workflows and technology (knowledge base)	I	W	1/9		
Collection Digitization Dashboard (CDD) is an extremely useful tool for high-level decision-making processes. It facilitates catalyzing categories and different levels of granularity per user community	I		2	8	
DiSSCo should develop a template to record cost and efficiency of digitization workflows to allow real comparisons	I	W	8		

Table 2: Innovation and consolidation in architecture, tools and technologies

Innovation/consolidation	I/C	A/W	ICEDIG WP	DiSSCo Prepare WPs	Comment
Adopt Digital Object Architecture (DOA) as basis for DiSSCo data management	I	A	6		
Nine characteristics to be protected throughout DiSSCo lifetime	C	A	6		
Store annotations separately	I	A	6		Same principle can be applied to other object types/transactions of DiSSCo, such as ELViS transactions.
Use Handles as persistent identifiers	I	W	6		There is still an open issue as to whether these should be NSId, IGSN or something else.
Separate authoritative and supplementary information about a specimen	I	A	6		Notion that trusted/approved experts outside of owning institution can modify authoritative information.
Root the definition of DiSSCo / openDS object model in Biological Collections Ontology (BCO)	C	W	6		Implies extending BCO
Base the definition of DiSSCo / openDS object model on ABCD 3.0 with EFG, ensuring alignment and compatibility with DwC	C	W	6		Open issue is how to deal with alternate terms used by ABCD and DwC for the same concept. openDS specification should contain a mappings annex. Take note of ABCD/DwC convergence work proposed in TDWG.
Support versions of Digital Specimen objects, with accompanying annotations and interpretations of verbatim data	I	W	6		
Pipeline for automated text digitization and entity recognition	C	W	4		
Review of automated georeferencing methods	C	W	4	5	
Recommendations for collection management systems	C	W	4	6	
Benchmark dataset of herbarium specimens	I	A	4	5	
Interoperability analysis of published specimen data	C	A	4	5	
Recommendations for improved standards for specimen data transcription	C	W	4	5	
Efficient semi-automated systems for entomology data and image capture	I	W	3		In development but need further improvement to digitize the bulk of NH specimens. See D3.5.

Table 3: Innovation and consolidation in culture, skills and capacity building

Innovation/consolidation	I/C	A/W	ICEDIG WP	DiSSCo Prepare WPs	Comment
Comparison of text digitization by humans	I	W	4		
Training in data management, such as the FAIR principles will help researchers take good decisions					
Citizen scientists can play and important role in the creation of data about collections.	C	W	2, 5		
Functional Units approach towards skills to more flexibly manage them by reorganising the competencies in respect to the existing capacities per institution	I	W	8	3	See MS48, MS49
Recurrent training is urgently needed throughout the entire data cycle and digitization process to allow effective implementation	C		8	2	

Table 4: Innovation and consolidation in governance and business model

Innovation/consolidation	I/C	A/W	ICEDIG WP	DiSSCo Prepare WPs	Comment
Assessment of electronic marketplaces for transcription	C	W	4		
Assessment of large-scale outsourced transcription	C	A	4		

2.3 Overall approach and direction

2.3.1 Digitization, Digital Specimens and Digital Collections

Digitization is the process of making data about physical objects digitally available, and the output of that process in the DiSSCo context 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 core information about the physical things it represents. These identified objects are amenable to processing and to transport from one information system to another. A persistent link must be maintained from the Digital Specimen to the physical specimen it represents, and which acts as a voucher[9] of the data making up the Digital Specimen. This link is the identifier of the physical specimen, together with the institution code and collection code to which it belongs; because physical specimen identifiers are not always unique. These Digital Specimen and Digital Collection objects are the principal data types 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) assigned when the object is first created. Each DiSSCo Facility must be responsible for creating (minting)

[9] A voucher specimen is a preserved, representative sample of a natural object class used for identification and as supporting evidence of information learned during the research process. It serves as a verifiable and permanent record because it preserves as much of the physical remains as possible.

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 must always return the current version of an object's content, as well as any interpretations and annotations associated with it.

Digital Specimen and Digital Collection objects are examples of the fundamental unit – a digital object – that is manipulated and managed by Digital Object Architecture (DOA), a powerful yet simple extension of the existing Internet. Such objects are treated as mutable objects with access control and object history (provenance) meaning they can be updated as new knowledge becomes available. Provenance data must be generated and preserved by operations acting upon DiSSCo data objects. Timestamped records of change 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 (ECOI). Other services, such as Collection Digitization Dashboard (CDD), the European Loans and Visits System (ELViS) and the European Curation and Annotation System (ECAS) build on and work alongside the ECOI service to provide a portfolio of DiSSCo services.

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.

2.3.2 The FAIR Guiding Principles

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, the principles are becoming widely adopted. Scientists increasingly rely upon computational and data infrastructure capabilities and capacities to assist them with their science. Collection-holding institutions increasingly must (will have to) rely upon computational and data infrastructure capabilities to assist them with modern-day collections management for best return on investment and to meet staff needs. The FAIR principles are a key element contributing towards responsible data stewardship and thus they are an essential consideration for DiSSCo data infrastructure.

Whilst adhering to FAIR principles, DiSSCo data management principles 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 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 followed [Lannom 2020, Wittenburg 2019b]. In DOA the core concept is the 'digital object'. Digital objects (of which in DiSSCo, Digital Specimen and Digital Collection objects are principal types) combine with FAIR principles in a framework of guidelines and principles surrounding 'FAIR Digital Objects' (FDO) – the FAIR Digital Object Framework (FDOF) [FDOF 2020]. For DiSSCo this is the primary route to 'FAIRness', one of DiSSCo's protected characteristics (i.e., enduring compliance with the FAIR Guiding Principles).

2.3.3 Minimum Information about a Digital Specimen (MIDS)

An important concern in digitization is how much detail to digitize from each physical specimen. While a photographic image can be made quickly, transcribing and interpreting all the details from labels, enriching the data with external information, and making specific measurements of the specimen take more time and resources. The idea of 'Minimum Information about a Digital Specimen' (MIDS) has been

conceived to capture and structure this complexity. In the present document we repeatedly refer to the MIDS levels. Briefly, they are:

0. MIDS-0: **catalogue** level information, unique identifiers, etc.; images if available;
1. MIDS-1: **basic** information derived from the collection, such as a (higher) taxon and geography;
2. MIDS-2: **regular** information of what (valid name), where, when, by whom, and how; as derived from specimen labels; and,
3. MIDS-3: **extended** information, enriched with external sources, not directly available from labels.

At the time of writing (January 2020) the MIDS concept is still under development. While the definitive specification and published article about MIDS is not available, the reader is advised to refer to the discussion of the subject in the Open Access Implementation Guidelines for DiSSCo[10] and in the provisional Data Management Plan for the DiSSCo Infrastructure[11].

A related concept, Minimum Information about a Collection (MICS) capturing and structuring an approach to describing collections has also been proposed and is under study. The suggested MICS levels are:

1. MICS-0: **Overview** information about a collection and the organisation holding it; and,
2. MICS-1: **Inventory** information describing the collection in its entirety.

2.4 The provisional Data Management Plan for DiSSCo infrastructure

The provisional Data Management Plan for DiSSCo infrastructure (hereinafter, 'the DMP') [DMP 2019] 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.

The intention in the present document is not to reproduce/duplicate all the information, data management principles and requirements documented in the DMP. Readers are referred to the original document for that [DMP 2019]. Here, the present document highlights (as necessary) specific elements having their basis in the DMP, for which further attention and recommendations might be made to other projects within the DiSSCo Programme.

3 Arrangements, processes and practices

Presently (2020), the task ahead of us, of mobilizing data from natural science collections is still enormous. Ninety percent (90%) of the collections still need to be mobilized. It is imperative for DiSSCo as it is in fact for all stakeholders and the community at large to tackle the backlog as quickly as possible, balancing this with a continued focus on impact and work to understand what makes data most useable and impactful. Arrangements, processes and practices, including organisational consolidation and use of selective partnerships must be initiated to efficiently digitize one and a half billion collection objects in a foreseeable time, at reasonable cost and to make this data publicly accessible. In addition, however, DiSSCo will need to recognise that digitization impact – the critical metric of success - depends on user needs; the aims of specific projects and research; and the research readiness of the data. Mass digitization workflows are still being developed and piloted for some collection types, and others may never be suitable for an 'industrial' approach (e.g., the very largest specimens). For these reasons, while focusing on mass digitization, DiSSCo will recognise and support the need for balanced portfolios of digitization activities that include digitization-on-demand capabilities driven and prioritised by researcher needs.

[10] ICEDIG project deliverable D6.5, doi: 10.5281/zenodo.3465285.
[11] ICEDIG project deliverable D6.6 doi: 10.5281/zenodo.3532937.

3.1 Role and development of a common digital research agenda

The need for greater alignment between digitization and biodiversity research agendas has been highlighted by both the Alliance for Biodiversity Knowledge (ABK) and in the Global Biodiversity Informatics Outlooks (GBIO). ABK specifically names one of its ambitions as working "with other research communities and infrastructures to achieve interoperability with earth observations, social science data and other resources." Stakeholders at GBIC2 called for better coordination mechanisms between research infrastructures and coordination between scientific communities with different but parallel expertise. A clear international digital research agenda for collections will be important in determining precise criteria for both mass digitization (3.9 below) and digitization-on-demand (3.9.3 below).

Recommendation 1: It is important to set a clear international digital research agenda that can serve as a guideline within DiSSCo to determine what to prioritize in terms of digitization of collections and specimens in more detail.

The coordination is required both internally within DiSSCo and externally with other research infrastructures. Research priorities and policies related to collections, ICT and data management vary widely across DiSSCo institutions and have a significant impact on how institutions can collaborate and facilitate a collective agenda. The first step in overcoming policy barriers that limit research alignment is identifying these internal policies and potential sticking points. ICEDIG task 7.2 took the first step towards documenting the existence and subject of policies across institutions (3.2 below). This lays the initial foundations for assessing where potential policy gaps, overlaps and barriers need to be anticipated and overcome in developing a common research agenda – and policy framework that supports it - within DiSSCo institutions. This will be facilitated in part by a new DiSSCo service providing a new central index service (marketplace) of expertise and facilities that lists the sources and availability of taxonomic and other scientific expertise, equipment and facilities across the DiSSCo collection-holding institutions.

External coordination between DiSSCo and other research infrastructures is also necessary for expanding the impact of bio/geodiversity research through collaboration with parallel areas of expertise such as environmental sciences, digital humanities, molecular biology and many others. A clear understanding of the existing RI landscape will also ensure there are no duplication of efforts but rather provides the opportunity to take advantage of existing work and potential efficiencies for achieving biodiversity research objectives.

ICEDIG task 7.3 was a major first step in documenting existing research infrastructures and their potential contributions to the research priorities of DiSSCo. In collaboration with GBIF and the ABK, the development of a Research Infrastructure Database has been started, containing a list of 59 research infrastructures (RIs) like or adjacent to DiSSCo [ICEDIG MS46]. To better assess the areas of potential international convergence, this list is being converted to a database with a visual presentation for a representative sample of the RIs[12] that incorporates the details of each RI including taxonomic coverage, funding sources and timelines, services, projects and users. Systematically coding and combining this data will allow for querying RIs based specific user needs and interests. The RI Database, once compiled, will provide insight into the types of services offered by RIs, ranging from field sites to ICT and hardware, and where potential opportunities lie for convergence and collaboration.

Recommendation 2: DiSSCo should promote that a global initiative, such as the Alliance for Biodiversity Knowledge or GBIF adopt and sustain the RI Database, ensuring that it is open, accessible and usable as a tool for identifying potential synergies and opportunities for collaboration with other research infrastructures.

[12] https://icedig.eu/content/research-infrastructure-introduction.

3.2 Common policy elements

Multiple national and European policy and legal issues can be identified as essentially affecting the successful delivery of DiSSCo. By identifying broad policy areas relevant to digitizing natural science collections and the proposed activities of DiSSCo (Table 5), there is potential to consolidate and harmonise. This will simplify development and implementation of DiSSCo services across participating institutions.

Table 5: Policy areas relevant to proposed activities of DiSSCo

1. Access and Benefit Sharing (ABS)
2. Data and digital media publication
3. FAIR / Open Data / Open Access
4. Freedom of information (FOI)
5. Intellectual Property Rights (IPR)
6. Data Standards
7. Personal data
8. Protection of sensitive collections data
9. Public Sector Information
10. Responsible Research and Innovation (RRI)
11. Cloud services and storage
12. Information risk management
13. Information security
14. Collections access and information
15. Collections care, development and scope
16. Digitization strategy and prioritisation

Each topic in Table 5 has been further analysed and broken down to reveal structured detail on how these topics have been described and implemented across the institutions surveyed (ICEDIG project deliverable D7.1). This structure supported a quantitative comparison of interpretations and implementation.

In some cases, such as freedom of information, intellectual property rights, personal data, public sector information, etc., policy areas are covered by national or European legislation, providing for a base level of harmonisation across all institutions. However, there is considerable variation in practical implementation of these policies, with much detail not sufficiently described or left to local interpretation. In comparison to internal, institution-specific policies, externally mandated policy is often generic, technical and not well-tailored to collection-holding institutions. Critically, the ICEDIG survey has revealed major gaps in policy coverage across all institutions, with some relying largely on national legislation that has been locally interpreted but arguably too abstracted from the details of practical implementation to provide comparison with other institutional policies.

Although policies might have been reported as existing, the amount of detail and formality represented by the available documentation varies widely. Furthermore, language barriers create significant potential for misinterpretation during the analysis, as policy documents were submitted in their original languages of Finnish, Estonian, English, French, German and Dutch. The remainder reported that their policies were either in progress, partially complete, developed external to the organisation, or were not in existence.

‘Collections access and information’ and ‘Collections care, development and scope’ have internal coverage in all institutions, which is likely to reflect that physical collections management policy is generally more mature than other elements of digital collections policy. Generic data and ICT policies such as ‘Information security’, ‘Personal data’ and ‘Intellectual property rights’ are also well represented; the legal and/or regulatory implication of these is likely to have been a driver for policy development in these areas. ‘Data standards’, as a more community-driven subject, appears to be less well embedded. Most institutions do not have a formal policy for ‘Cloud services and storage’ or ‘Public Sector Information’.

Another point to consider when reviewing policy coverage is the extent of policy that is followed with little or no internal documentation i.e., with reliance on external documents. For example, if an institution is a member of CETAF there are policies they will be party to (such as those relating to Stable Identifiers and Passports, for example) that are not specifically transcribed into local (institutional) policies. Similarly, when it comes to access and benefit sharing, institutions that are part of the European Union are obliged

to follow the Nagoya Protocol, but this may not be stated as a written policy and in some areas (e.g., Digital Sequence Information) subject to significant variance in interpretation. Other examples include GDPR regulations and the CITES Convention.

Not all institutions are able to share their policies externally. Only one third of institutions stated that all their policy documentation submitted could be shared, whereas 60% of one institution's documentation was not sharable. This is important to consider when creating common policy around digitization as DiSSCo moves toward a common research agenda.

In sum, the ICEDIG policy analysis faced significant challenges around collecting and interpreting policies from ICEDIG partner institutions. This was due to several factors, including difficulties in obtaining relevant policies, differing levels of policy making it harder to map policies to categories, and policy within different legal frameworks and language barriers that created the potential for misinterpretation. To address these challenges within DiSSCo, solutions should consider:

- Proposals on how to introduce and streamline relevant policy towards a common DiSSCo agenda;
- Establishing a knowledge base on how policies are organised within a collection-holding institution;
- Establishing where responsibility lies in ensuring the correct policies are in place and adhered to; and,
- Establishing who has authority over enabling policy change within collection-holding institutions.

Such actions must be framed in the context of developing a common policy framework that addresses the gaps in provision currently experienced across DiSSCo institutions, as well as supporting local variation in how these policies are implemented.

Recommendation 3: Actions on common policy elements across DiSSCo institutions must be taken in the context of a common framework of policy definition and implementation that recognises the organisation of policies and responsibility for implementation at local level, and authorization for change within collection-holding institutions.

3.3 Participation of citizen science

In natural science collection-based research, citizen scientists are most often engaged in transcribing specimen labels. The use of public, web-based transcription platforms, of which there are several available as open-source in the domain[13], allows crowdsourcing and the mobilisation of citizen scientists to assist in the digitization process. Transcription by volunteers is expected to continue in coming years, even increasing and diversifying in the tasks they undertake – although it is likely there are limits to the usefulness of crowdsourcing as a mass transcription tool, with data quality and cost needing continual attention. Increasing automation may lead to new forms of participation by experts and wider citizens, for example 'human in the loop' approaches to dealing with exceptions from automated data extraction.

In general, there is no single best platform to recommend nor any need for DiSSCo to seek to build a single, universal DiSSCo volunteer platform. The community engaged around existing platforms or instances of platforms such as the DoeDat implementation of DigiVol by Meise Botanic Garden[14] is often their strongest asset. Thus, it is beneficial to exploit those platforms better, making use of the variety of their features, languages, scientific interests and gamification mechanisms they each have as their strengths.

[13] For example, DigiVol, Les Herbonautes, Zooniverse, Notes from Nature. See ICEDIG project deliverable D5.1 (section 'evaluation of existing volunteer transcription systems, milestone MS26'); as well as BioSPEX from the USA.
[14] https://www.doedat.be/.

Language specificity is one of the strongest of such attractions, and with multiple platforms operating in the European landscape can open citizen science to institutions with no present transcription capacity.

Integrating the diversity of platforms into digitization workflows is made possible by implementing a common data exchange protocol[15]. Such a common data protocol must operate between digitization lines and transcription platforms, from transcription platforms to collection management systems (CMS) and from transcription platforms to dashboards.

Recommendation 4: DiSSCo should exploit the diversity of available citizen science platforms (e.g., for specimen label transcription), taking advantage of their individual strengths (surrounding community interest, language specificity, etc.) as appropriate and should encourage such platforms to implement the data exchange format and protocol for transcription platforms (doi: 10.5281/zenodo.2598413), as well as supporting this format/protocol in digitization lines, workflows and collection management systems.

As noted above, a strength of existing transcription platforms is their surrounding community of volunteers engaged in collection transcription. Often though, the fact of citizen scientists' involvement in data enrichment processes is hidden in the final versions of datasets. Giving proper credit to these people for their contribution is morally right and essential for motivating future contributions, as well as being ethically right for the curation and maintenance of collections (i.e., establishing provenance of data). Furthermore, integrating and exposing citizen science activity through DiSSCo dashboards could be a powerful incentive for increased mobilisation of volunteers in the future[16].

In the future, it is advisable to ensure that attribution details are preserved in transcription datasets, transferred to collection management and systems and published publicly when such data is published. Darwin Core and GBIF Metadata Profile standards allow to some degree to describe also volunteer involvement in data collection and enrichment. A new recommendation for the representation of attribution metadata recently published by the Research Data Alliance[17] will make this easier to achieve.

Recommendation 5: The involvement of citizen scientists in DiSSCo data work and activities must be properly acknowledged and attributed, for example using Research Data Alliance recommendations for the representation of attribution metadata (doi: 10.15497/RDA00029).

For properly engaging the public with collections and collection digitization, deep understanding of the relationships between collections, particularly digital collections; formal and informal education; museum-related citizen science; and the skills and knowledge that these can advance is needed. This is a complex area that ICEDIG has only touched upon briefly[18]. Nevertheless, it seems clear that while national and cultural differences can have an impact on citizen science and education, the similarities – for example in what approaches are likely to engage people – are more important than the differences. Thus, a few general business model principles can help to guide future projects that engage citizens with collections. Time spent in considering when is it worthwhile (cost beneficial) to embark on a citizen science activity, on precise definition of the contributors/audience for the activity, on how to engage the contributors, how to raise the skills and knowledge levels of participants and how to sustain their interest for the

[15] Such a protocol has been developed. See ICEDIG project deliverable D5.1 (section 'Specification of data exchange format for transcription platforms, milestone MS28'; also, doi: 10.5281/zenodo.2598413).

[16] Noting, of course the need to deal appropriately with the personal data of volunteer transcribers in the context of the General Data Protection Regulation (GDPR). This is covered in the DiSSCo Data Management Plan, section 12 [DMP 2019].

[17] See doi: 10.15497/RDA00029 for the attribution metadata recommendation itself and doi: 10.5334/dsj-2019-054 for an explanatory article.

[18] See ICEDIG project deliverable D5.3, doi: 10.5281/zenodo.3364541 where this is explored from practical perspectives, leading to five general business model principles.

duration of the work – these can all help to maximise the opportunities for collections to be successfully used in citizen science and education. This can increase the engagement, skills and knowledge of citizens, whilst at the same time achieving objectives of the collection-holding institutions themselves around digitization, curation, maintenance and use of their collections.

Recommendation 6: Recognising the likely future increase in citizen science involvement with natural science collections, DiSSCo should further develop a package of business model principles and guidance that collection-holding institutions can use to design and manage citizen science engagements and activities to their collections.

3.4 Organisation and partnering choices

3.4.1 Types of partnering

Strictly speaking, the word 'alliance' should be used instead of partnering or partnership because it's the more accurate and wide-ranging term conferring the idea of being united for a common purpose or for mutual benefit. Alliances match parties' strength to strength and balance control with collaboration. They increase the capacity and capability of each of the involved parties without necessarily asking the parties to relinquish control from one to the other. From the perspective of DiSSCo needs (and discussed in each of the following three sub-sections 3.4.2 – 3.4.4) three main forms of alliance exist: customer-supplier relationships, strategic alliances and stakeholder investments.

Compared to these three kinds of alliance, true partnerships are much more about participation, pairing and merger of individual interests. Although they are concerned with collaboration, separate control is not retained. They not relevant for DiSSCo at present.

3.4.2 Customer-supplier relationships

Customer-supplier relationships are relationships in which customers receive goods and services and suppliers receive monetary payments or other considerations. Such relationships are usually more tactical than strategic, although they often are important contributors towards some higher strategic goal. They can be straightforward and often involve a tendering process leading to contractual arrangements that usually include a service level agreement (SLA). Customer-supplier relationships can occur on many levels, from institutional through national and regional to pan-European. Choice is normally governed by the available suppliers having the correct solution to meet requirements at a value for money price. DiSSCo and its member institutions are subject to the requirements of EU legislation on public procurement[19], or its national equivalent in the case of non-Member States.

DiSSCo has the need for many kinds of customer-supplier relationships, including the following:

- Gaining access to and exploiting vast amounts of storage capability in the form of third-party trusted repositories;
- Speeding up and cutting the cost of digitization processes and procedures; and
- Operating a persistent identifier minting and resolution mechanism;

3.4.3 Joint ventures

Joint ventures, in which the parties/partners/beneficiaries commit resources to jointly pursue common goals can be either operational or strategic, depending on their purpose. Normally they are the latter, being essential to enhancing the value of each individual party and because the parties separately cannot achieve the desired goals on their own. Strategic joint ventures are normally long-term arrangements. When more than two parties contribute, this is a consortium.

[19] https://ec.europa.eu/growth/single-market/public-procurement/rules-implementation_en.

Strategic alliances include development of interfaces that do not necessarily include provision of services or products from either side, but which can lead to mutual benefit for both parties. Strategic alliances can include technology, knowhow, and skills transfer. They are governed by bespoke agreements (MoU, consortium agreement, etc.) setting out the ambitions and obligations of each party in the context of the shared goals.

DiSSCo has the need for many kinds of joint venture, including the following:

- Creating bidirectional (resolvable) links between natural science specimens and DNA sequences;
- Creating bidirectional (resolvable) links between natural science specimens and relevant literature;
- Guaranteeing availability of Digital Specimen information for the next 100 years;
- Contributing to and receiving from the European Open Science Cloud;
- Achieving interoperability with cognate research infrastructures (biodiversity science, ecology, environmental sciences, social-economic, molecular, chemical, etc.) at both European and global levels.

3.4.4 Stakeholder investment

There is a third kind of alliance – stakeholder investment – whereby one party makes an investment in a second party with a view to, for example ensuring the sustainability and longevity of the second party, being crucial to the first party's operations, or influencing the behaviour of the second party in a direction more favourable to the first party. The second party benefits by having its resources and future sustainability improved because of the investment. A topical example of this kind of alliance might be the situation whereby DiSSCo becomes a member of the International DOI Foundation or the DONA Foundation, or where investments are made for sustainable software development. This kind of alliance is considered further in the sub-section on shared liabilities (6.4.5 below).

3.4.5 Organising alliances

The landscape within which DiSSCo is positioned is complex, and its mission and needs encompass:

- Collections science;
- Related fields (applied research fields);
- Interdisciplinary fields; and,
- Decision making/policy informing/public engagement.

On top of this, the broader landscape encompasses other related facilities, resources and services used by the scientific community to conduct research such as instruments, archives or structures for scientific information, computer-assisted tools and ICT infrastructure, such as cloud computing environments and Internet communications.

The question then is: How to organise alliances that allow DiSSCo to operate – to deliver what is mandated to the best of its ability and to deliver that better through alliances than DiSSCo could do alone?

Several Roundtables[20] gathering focused expertise from outside the ICEDIG/DiSSCo contributed to the overall level of knowledge about where and how to organise alliances:

1) Collection Digitization Dashboard: see 3.10.5;
2) Analogue 2 Digital: One of the most time-consuming steps within the digitization process, i.e. the extraction of label information and the different methods available to do so;
3) Future of warehousing and use of robotics: see 4.6.6. Also, use of robotics in 3D scanning;
4) Partnership Frameworks for Distributed Research Infrastructures, to share experiences, learn from more mature initiatives and identify possible best practices to follow;

[20] See ICEDIG project deliverable D9.3, doi: 10.5281/zenodo.3632535.

5) Museums specimen and molecular data linkage; and,
6) Cultural Heritage Synergies, establishing digital needs and requirements of humanities researchers.

Seeking out complementary infrastructures, commercial organisations/industries and other initiatives, and identifying where joint efforts could exist; building on existing investments; and leveraging cross-infrastructures collaboration is all important work that must be continued within the overall DISSCo programme.

Technological innovation, efficient deployment, harmonized and collaborative infrastructure development will be critical. Potential innovators working in closely related fields such as optics, robotics, artificial intelligence, geo-localisation, imaging, lab instruments, data storage and many others are to be further engaged as necessary e.g., for joint development of collective technology and infrastructure that would each on its own require investment and competence beyond the capacity of a single RI, or with innovative or already existing technical capabilities such as Artificial Intelligence, or connective infrastructure such as Access and Authentication Infrastructure.

Organising alliances must occur in the framework of the common digital research agenda (3.1) to enhance identification of optimal potential partners, for example where third party data services can potentially augment the DiSSCo offering and produce mutual research benefit e.g., ELIXIR life sciences services. Quantitative assessment and due diligence must be carried out on relevant activities, services or components of organisations and infrastructures that can inform decisions on alliance making as the DiSSCo blueprint is further developed.

Recommendation 7: DiSSCo should establish criteria and procedures for assessment and due diligence of activities, services and components of relevance to any potential alliance that keep in mind the common digital research agenda of DiSSCo.

Convergence of the ESFRI RIs in the common landscape is necessary and should be continued. The different state of RIs and their timing do not help in this endeavour as different facets, like needs and solutions, develop at different speeds. This also applies to the relationship between RIs and the e-infrastructure providers. Nevertheless, RIs will have to invest efforts to identify interfaces with common lines of production from where then common services could emerge. This needs commitment from other RIs – the new ESFRIs, established ESFRIs, and ERICs – to continue this dialogue. Such a dialogue began with Roundtable 5 and will be continued.

Recommendation 8: DiSSCo should continue dialogue with representatives of complementary research infrastructures to ensure convergence towards a common approach, especially in the context of the European Open Science Cloud.

Building better and more effective coordination mechanisms for international cooperation across natural science collections globally is essential to overcome the challenges of an increasingly crowded and complex landscape. Specifically, duplication of effort must be avoided but much more importantly: divergence in approach to enabling global collections-based science must be avoided at all costs.

Recommendation 9: Working proactively with international partners DiSSCo should aim to avoid divergence in technical approach to the support of global collections-based science.

3.4.6 Identified strategic opportunities for DiSSCo

When it comes to joint venture, challenges arise from the distributed nature of the RI. DiSSCo will exist in a complicated landscape of: i) European and international obligations and initiatives (2.1.6) ; ii) legal

constraints and regulatory implications (3.8); and iii) national interests. Each of these must be successfully navigated.

Alliances must be forged appropriately, along three main lines:

a. Within DiSSCo with the national nodes;
b. With other thematic players – i.e., other RIs and their technical and strategical interfaces, and global partners (e.g., via GBIF, iDigBio) that either serve or incorporate DiSSCo's mission; and,
c. With the foundational e-Infrastructure providers contributing to the European Open Science Cloud (EOSC).

DiSSCo already has several strategic opportunities available to it that must be further developed during the Preparatory Phase for mutual benefit during the construction and operation phases.

One big strategic advantage for DiSSCo is its fundamental alliance with CETAF, the Consortium of European Taxonomic Facilities[21]. This was established at an early stage as part of the work to bring DiSSCO onto the ESFRI 2018 Roadmap. It roots the RI firmly in its own community with executive level commitment from the major natural science institutions of Europe.

Another advantage available to DiSSCo is the already well-established network of GBIF national nodes and GBIF itself as a global initiative with the goal to mobilise the world's primary biodiversity data. Note however, that this also represents an area of overlapping interests and thus potential conflict.

Digitization initiatives around the world (see 2.1.6) such as ADBC/iDigBio, NRCA Digital and National Specimen Information Infrastructure, each with similar aims to DiSSCo have much to offer in the way of mutual support and development in pursuit of seeking global solutions to common aims for digitized and extended natural science collections.

Recommendation 10: DiSSCo should further clarify, develop and nurture the joint ventures (strategic alliances) that will be important to its plans and operations, including with CETAF, GBIF, iDigBio (or equivalent), EOSC, etc.

3.5 DiSSCo Centres of Excellence

A DiSSCo Centre of Excellence (DCE) is a designated DiSSCo Facility that specialises in one or more of researching, innovating, developing and operating/performing techniques and/or processes of digitization or other related facets, and disseminating information on same.

There are several ways by which DCE can be established, funded and governed, including for example by combining private sector technology, innovation and training with publicly funded digitization projects (i.e., private/public partnerships) involving the previously mentioned organisations such as Bioshare, Dinarda, and Picturae. (See also 3.9.5 below).

Thematic DCE (considered in detail in the [ICEDIG MS45] report) concentrate services focussed on specific collections constrained by characteristics such as object type, taxonomy and geographic regions, and the related digitization workflows and domain expertise. Such specialisms can potentially have influence on factors like funding models, legislative and legal requirements, availability of facilities and logistics that differ from the more generic model. There are also regional contexts to be considered, where the fit between services and organisational levels may be influenced by patterns of local and national funding,

[21] https://cetaf.org/.

institutional expertise and regional differences in collections management practices. This having been said, Centres of Excellence should follow a principle of harmonising differences in practices across thematic, geographic and community boundaries where possible and beneficial.

3.6 Role of the private sector and options for public procurement

The private sector has a potential role to play in several areas of DiSSCo operations, being mainly product or service supply and (where appropriate) maintenance and/or training in the following areas:

- Specialist digitization equipment, such as scanners, cameras and other imaging technologies, conveyor machinery and other automation, including associated specialised software;
- Bespoke and/or outsourced digitization services (digitization factory); and,
- Data storage services.

There are opportunities, for example in automated text digitization (4.6.2) for partnerships with global companies such as Google.

A further possibility arises when commercial companies use and benefit from the 'free' data, agreeing perhaps to give funding in return. Among geological collections in the USA, this is an opportunity that is just beginning to be tapped.

According to the [ICEDIG MS45] report, purely commercial entities are unlikely to be able to provide enough assurances on several fronts (breadth of provision, conflict of interest, intellectual property, sustained very long-term data management) to qualify to become a DiSSCo Centre of Excellence (DCE). However, commercial entities are expected and should be strongly encouraged to take a role in DiSSCo service provision, by entering into cooperative public/private partnership (PPP) with DiSSCo Centres of Excellence and/or institutional stakeholders. As well as offering leading-edge services, a DiSSCo Centre of Excellence must be able to act as a neutral broker between DiSSCo stakeholders without perception of possible conflicts of interest or technology bias. Thus, having multiple PPP across several suppliers is beneficial. Such DCEs are likely to offer expertise, support and even training either free or at-cost to users who would not otherwise be able to take those up so, again having multiple PPP is beneficial.

3.7 Open science provider partners

3.7.1 Generic data storage and computation services

Through the ICEDIG project, DiSSCo has worked closely with open providers of data storage and archival services at national level[22], at European level (EUDAT[23]) and internationally (Zenodo[24]) on the potential adaptations such services should make to enable the long-term storage of large-scale digitized biodiversity data (meaning, primarily image data). There is no 'one size fits all' solution and the expectations are that DiSSCo and its member institutions will exploit different kinds of storage and computational solutions, including those mentioned, HPC solutions and solutions from cloud service providers, according to the different scenarios and needs in different parts of the DiSSCo infrastructure. This has been covered in-depth in the provisional data management plan for the DiSSCo infrastructure [DMP 2019]. Wherever possible, such services should be procured from providers having the ambitions and aims of open science at the heart of their mission statements.

Recommendation 11: DiSSCo should exploit generic services for data storage and computation where possible and procure from service providers having the ambitions and aims of open science at the heart of their mission.

[22] ICEDIG project deliverable D6.4, doi: 10.5281/zenodo.3469490.
[23] ICEDIG project deliverable D6.2, doi: 10.5281/zenodo.3364533.
[24] ICEDIG project deliverable D6.3, doi: 10.5281/zenodo.3346782.

3.7.2 EOSC and the FAIR Digital Object Framework

DiSSCo has welcomed and fully endorsed[25] the European Open Science Cloud (EOSC) Declaration, recognising the mission-critical role of EOSC towards an open science and open innovation research landscape. DiSSCo has acknowledged the operation of EOSC as foundational for the successful delivery and provision of the DiSSCo Research Infrastructure (RI) services. To this end, DiSSCo participates actively (especially through the Research Data Alliance Group of European Data Experts, GEDE[26]) in areas of critical importance; for example, the newly emerging FAIR Digital Object Framework (FDOF) and its accompanying declaration[27], and in ensuring that DiSSCo requirements for persistent identifiers can be met by any new European Persistent Identifier Service (EUPS).

Recommendation 12: DiSSCo must adopt the FAIR Digital Object Framework (FDOF) and its realisation through Digital Specimen, Digital Collection and other relevant digital object types as the basis for complying with the FAIR Guiding Principles for natural sciences data management, and as the means of delivering FAIR compliant natural sciences data into the European Open Science Cloud (EOSC).

Note: At the time of writing (January 2020), several institutions active in the DiSSCo programme planning are committing towards a proposal for a new EC-funded project under work programme item H2020-INFRAEOSC-03-2020[28], on increasing the service offer of the EOSC portal by proposing digital infrastructure based on the FDOF that equally serves humans and machines. Here, DiSSCo serves as not only one scientific testbed, but also a strategic partner in developing elements of the FDO core system. This will lead eventually to close strategic integration between DiSSCo and EOSC for the long-term.

3.8 Legal and regulatory implications

3.8.1 Implications for establishment of research infrastructures

The legal establishment and financing of a supranational research infrastructure (RI) such as DiSSCo can be lengthy, complex and difficult, taking several years to implement. The applicable laws are: community law, the law of the Member State of the statutory seat, and law of the Member States where operations are carried out. Several formalisms are possible and applicable at different stages of an RI's lifetime.

3.8.1.1 Early stage arrangements

It's possible during the early stages of an RI's life (such as for investigation, planning and organisation) to be adequately served by a simple consortium agreement, or an arrangement based on a Memorandum of Understanding (MoU) signed by a few institutions. Indeed, this is the case currently for DiSSCo whereby stakeholders have committed via an MoU to support and contribute to the development of DiSSCo's mission and goals. MoUs are not legally binding but carry a degree of seriousness of intent and have mutual respect from the signatories. Often, MoUs are the first step towards a more permanent agreement with a legally binding basis.

An MoU acting as a framework for cooperation can be further strengthened when accompanied by a set of statutes. Such statutes can, to all intents and purposes be nearly identical to those that would exist in

[25] https://www.dissco.eu/dissco-endorses-the-eosc-declaration/.
[26] https://github.com/GEDE-RDA-Europe/GEDE.
[27] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects.
[28] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/infraeosc-03-2020. "Integration and consolidation of the existing pan-European access mechanism to public research infrastructures and commercial services through the EOSC Portal".

conjunction with more formal arrangements (see below), and as such can help to provide a smooth transition when necessary.

In addition, specific funded projects during the early stages of DiSSCo's life, such as ICEDIG, SYNTHESYS+ and DiSSCo Prepare are normally organised based on formal consortium agreements between the beneficiaries, who are not necessarily all signatories to the MoU but nevertheless share some common interests and motivations.

3.8.1.2 More formal arrangements

On the other hand, creation of the final, long-term collaboration over many years expected of a supranational RI most likely requires a negotiated intergovernmental agreement. Being backed by governmental/ministerial decisions, negotiated agreements of this type have a significance and weight beyond other formalisms. Although these arrangements can take a long time to achieve, they can be stable and enduring once established.

To support such agreements in respect of European RIs, the European Commission, responding to requests from EU Member States and the scientific community, has proposed a community legal framework - a European Research Infrastructure Consortium (ERIC) - adapted to the needs of new European-level research facilities and of distributed infrastructures as well. An ERIC is an entity with a legal status recognised in all EU Member States. It meets the needs for recognition of the European identity on a non-economic basis, has a flexible internal structure to accommodate diverse types of infrastructures, and provides some privileges and exemptions (most notably, with respect to Value Added Tax (VAT)). Under certain conditions, an ERIC can also include non-EU partner countries.

Another approach is a public benefit foundation, established to carry out the mission of DiSSCo. However, with no members or shareholders, and no formal recognition in European law, such a foundation must be established in a single country. A foundation of this kind must be certain that it is guaranteed not only a reliable income stream for the long-term (e.g., by contributions from its supporting organisations to carry out their mandated activities) but also that the ongoing and long-term support of the stakeholders is certain. It is not always a reliable arrangement.

3.8.1.3 Importance of correct legal status

The choice of the correct legal status is also important from funding and commercial perspectives. A minimal solution such as the MoU approach described above that permits achieving the scientific goals with the smallest possible legal, administrative and financial complications, whilst administratively preferable, could underestimate the need for an adequate legal foundation. For example, with MoUs and other non-legal forms, a group of institutions cannot enter collectively into commercial contracts and other agreements with third parties (such as product or service suppliers). This must be done (and the risk assumed) by one of the institutions alone. Similarly, funding agencies often have formal eligibility requirements regarding the legal standing of entities that submit proposals or are the recipients of grants. Whether this is a single institution acting on behalf of a consortium (as, for example in European framework funding for research) or whether it is with an entity that is a representation and constitution of the institutions collectively, has significant administrative, taxation, employment and financial regulatory implications.

Recommendation 13: DiSSCo should propose the legal form required to achieve its aims and objectives and to administer and support its operations, keeping in mind the need for long-term viability and stability, the need to be able to enter into legal agreements with third-parties, and the need to assume responsibility for and mitigate risks and liabilities.

3.8.2 Implications for data management practices

Many of the 16 policy areas studied by ICEDIG (see Table 5, page 34) have implications from the data management perspective when it comes to complying with legal and regulatory requirements. In some areas, such as: i) personal data, where compliance with the General Data Protection Regulation (GDPR)[29] is mandatory; and ii) data and digital media publication, where the INSPIRE Directive[30] applies for spatial information (of which natural sciences data is kind), there are specific data management requirements that must be met and practices/procedures that must be put in place.

Some examples of specific legislation applying directly to DiSSCo activities with implications on data management practices:

- International multilateral environmental agreements (conventions), for example:
  - Convention on Biological Diversity (CBD), including the Nagoya protocol on Access and Benefit Sharing (ABS)[31];
  - 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 guidance for how DiSSCo and its member institutions should implement and demonstrate compliance with legal requirements of such legislation.

In other areas, there is more discretion and flexibility. Legislation and regulations may not apply directly (Freedom of Information and Intellectual Property Rights are two example areas) but taking the appropriate legislation into account when designing data management practices and systems can make compliance much easier to achieve and police.

In both cases, it is sensible that DiSSCo and its member institutions adopt, as far as possible common approaches to data management to comply with such requirements. This is a topic for the data management plan of DiSSCo that is marked for further study in the present provisional data management plan [DMP 2019] prepared by the ICEDIG project. Further work will be needed, for example in the DiSSCo Prepare project to document the specific data management requirements arising from each specific piece of current legislation. This is part of achieving the legal readiness of DiSSCo.

Recommendation 14: For each broad policy area affecting DiSSCo activities and directly covered by mandatory legal and regulatory considerations for data management, DiSSCo must list the legislation and regulations that apply at national and/or European level and say how DiSSCo and its member institutions will comply with each of the mandatory requirements (for example, by indicating specific clauses in the DiSSCo Data Management Plan). The broad policy areas are: i) Access and Benefit

[29] General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. Latest consolidated version (as of 21st January 2020) here: http://data.europa.eu/eli/reg/2016/679/2016-05-04.
[30] Directive 2007/2/EC, INSPIRE. Latest consolidated version (as of 21st January 2020) here: http://data.europa.eu/eli/dir/2007/2/2019-06-26.
[31] See https://www.cbd.int/abs/about/.

Sharing (ABS); ii) Data and digital media publication; iii) FAIR / Open Data / Open Access; iv) Freedom of information (FOI); v) Intellectual Property Rights (IPR); vi) Data Standards; vii) Personal data; viii) Protection of sensitive collections data; ix) Public Sector Information; x) Responsible Research and Innovation (RRI); xi) Cloud services and storage; xii) Information risk management; xiii) Information security; xiv) Collections access and information; xv) Collections care, development and scope; and xvi) Digitization strategy and prioritisation.

Recommendation 15: For each of the broad policy areas mentioned in recommendation 14 affecting DiSSCo activities and affected by legislation indirectly, DiSSCo should state what practices it will adopt to make compliance easier to achieve and police.

Specific attention will need to be given to the topic of moving sensitive data about collections, specimens, personnel (e.g., collectors) and places across international borders in cases where DiSSCo collection-holding institutions are not located in Member State countries belonging to the European Union. In such cases, there are additional considerations, such as (for example) whether the European Commission has determined that the country in question provides an adequate level of data protection[32]. The effect of Brexit, 31st January 2020 and new arrangements agreed during the subsequent transition period must also be considered in the case of UK participation to DiSSCo. A definition of 'sensitive data' should be made to assist this and it might be necessary to develop and deploy a DiSSCo compliance and moderation service (COMOS). Note, however that complete control of sensitive data, especially across multiple jurisdictions is a daunting challenge and so the best reasonable effort should be made in order to balance protective control with usability (which means convenience) and cost.

Recommendation 16: DiSSCo must give specific attention (perhaps by implementing a 'compliance and moderation service') to the rules governing the movement of sensitive data across international borders i.e., between European Union Member States and third countries (including defining specifically what is meant by 'sensitive data' in the context of the legislation affecting DiSSCo operations).

3.8.3 Open research data

In 2012, the European Commission published a first recommendation, updated in 2018[33] on access to and preservation of scientific information that encourages all EU Member States to put publicly funded research results in the public domain in order to strengthen science and the knowledge-based economy. EC Recommendations do not carry the same weight as other types of EU legislation, but the expectation of the Commission is that Member States comply with them. On occasion, such recommendations are a prelude to more obligatory legislation such as Directives and Regulations. It is quite possible that within the next few years we may expect to see stronger European legislation in the open science area.

Implemented by extending the Open Research Data pilot into all thematic areas of Horizon 2020 funded work programmes in 2017, the Commission's Recommendation applies also to DiSSCo and to its institutional members when they enter into EC-funded Grant Agreements for project-based work under Horizon 2020, and most likely the following research and innovation framework programme, Horizon Europe. Implementation and enforcement are generally contractually via model Grant Agreement

[32] An adequacy decision, see: https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en.
[33] Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, http://data.europa.eu/eli/reco/2018/790/oj.

clauses. The Guidelines to the Rules of Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020[34] explain the application of the principles.

This framework is strongly binding for all publicly funded collection-holding institutions, whether it is explicitly transformed into an institutional policy or not. It makes clear that natural sciences data collected in these institutions must be findable, accessible, interoperable and reusable (FAIR) by default. Thus, requirements and recommendations of EC Recommendations on access to and preservation of scientific information, including research data are implemented through DiSSCo's open access guidelines (see 4.4 below) and measures for complying with the FAIR Guiding Principles (see 2.3.2 and 0).

3.9 Mass digitization

3.9.1 Characteristics

For scientific collections we define mass digitization to be an activity where entire collections, or their distinct major parts, are digitized from one end to the other, without selecting individual specimens.

Mass digitization is characterised by improved technological and procedural frameworks based on automation (both hardware and software, e.g. use of conveyors, barcodes, machine learning approaches etc.) and enrichment (link-building), although currently and for the foreseeable future the role of human operators/digitizers remains paramount in many stages of digitization. Mass digitization means implementing workflows at industrial scale, i.e., processing millions of objects annually at relatively low cost. The large scale of European natural science collections provides the opportunity and need for such industrial approaches, but also makes it imperative to maintain downward pressure on costs if this work is to be affordable at the scale required (3.9.3).

Mass digitization often contrasts with demand-led/driven digitization (also known as digitization-on-demand) (3.10 below).

3.9.2 Factors influencing digitization choices

Several major factors affect the choice of mass digitization approach, including the size of the collection to be digitized; availability of staff and their time, availability of funding; availability of space to locate equipment; and whether transportation of the collection is viewed as safe and viable. These factors broadly can be used to decide between: i) in-house digitization with own equipment and by own staff; ii) in-house digitization carried out by a contractor using their own equipment installed in the premises; and iii) digitization out-sourced to a contractor, which requires transportation of the collection to an off-site facility run by the contractor[35]. Separately from these major factors, the timing of digitization and the type of collection must also be considered.

The type of collection also is a factor. The technology of mass digitization is already well-established for herbarium sheets, which are more-or-less two-dimensional and other objects that can be placed flat (focus depth 10 cm) on (A3 sized) trays. Herbarium sheets are routinely being transported for digitization at remote facilities. Similar techniques are slowly becoming available for pinned insects as well. These are seldom shipped to remote facilities for digitization though, because of their perceived fragility and because the size of the required digitization equipment is smaller overall and often better fits into available space in museum buildings. These two collection types cover approximately 80% of all specimens. For the rest such as liquid samples, microscope slides, most vertebrate material, and non-biological specimens, mass digitization techniques still need to be better conceived; although some progress is evident e.g., [Mendez 2018, Allan 2019].

[34] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.
[35] These different approaches are analysed (with some cost comparisons) in the ICEDIG project deliverables D3.2 – D3.6.

Recommendation 17: DiSSCo strategy for mass digitization must focus on clearing the historical backlog of undigitized specimens in the next 20 – 30 years, whilst recognising that newly collected accessions, and small and private collections also each require their own organisation (workflow) of digitization to prevent new backlogs from forming. The specific collection type also dictates the appropriate technical approach and although herbarium sheet and pinned insect digitization is well-developed or advancing, greater emphasis must be placed on other collection types, including non-biological ones.

3.9.3 Affordability and achievability

There are currently a wide range of per-specimen costs for digitization, which represent workflow differences, variation between institutions, as well as variation in the aims of projects and the level/extent of digital data generated. ICEDIG has proposed templates for collecting full economic costs of digitization and has begun to record these costs (5.3.3 below). Further evidence will need to be collected on an ongoing basis. Cost effectiveness or value depends on the balance between costs and benefits/impact and may often be positive even where per specimen costs are quite high, provided the reasons for these costs are proportionate to the desired outcomes. However, even if cost effective, higher costs will limit the affordability and achievability of mass digitization for the 90% of collections remaining, so there must be a focus on achieving lower average item costs overall. It is inevitable that, even as average costs reduce, some types of collections (e.g., the largest items, those where 2D imaging does not meet scientific needs, and many of those preserved in spirit) are likely to have higher per item costs. For many types of collections, however, mass digitization workflows exist and are being refined to reduce cost. This includes collection types that account for much of the volume to be digitized, including herbarium sheets and pinned insects. Imaging millions of these specimens should be viable at average target cost of around €0.20 per specimen, and transcription of their essential (MIDS-2 level) data should cost less than €0.30 each when properly supported by technological and automated approaches, and/or done in less expensive parts of the world.

Recommendation 18: DiSSCo should plan to achieve an average digitization cost of €0.5 or less per specimen, across major collections types to which mass workflows can be applied.

Digitization-on-demand offers one key form of prioritisation, leaving choices about digitization to be largely influenced by user needs and requests. Mass digitization, however, also requires prioritization.

3.9.4 Organising mass digitization

Considering that backlogs of undigitized historical material, newly collected accessions, and small and private collections each require a different approach, and that the specific type of a collection to be digitized also dictates an appropriate technical approach, we can now try to formulate some questions to underpin decisions by DiSSCo on how best to move forward with arranging mass digitization. These questions, with some further discussion of each below are:

1. What is the balance of inhouse work to be done versus the efficiencies/economies from using outsourced facilities?
2. What is the need of specialisation (and centres of excellence) as per collection type and digitization technology?
3. How to move from one-time digitizing of the backlog to ongoing digitization and effective data mobilisation that is internalised as normal business in the institution?
4. How to motivate and support small and private collections to digitize their material?
5. How many out-sourced digitization facilities are needed in Europe and where and for how long they should operate?

6. Will DiSSCo centrally own and operate any mass digitization facilities, or shall they (only) belong to its members and/or private industry?
7. Can (some) collections permanently be warehoused next to the digitization facilities?

1 Inhouse vs outsourced: Centralised, specialist digitization teams have become the norm within many European collection-holding institutions [ICEDIG MS44]. These enable expertise to be developed within and across workflows, as well as providing training and a team environment to digitizers. However, mass digitization at low cost in-house requires a continual flow of material suitable for digitization, and often imaging set-ups that can adapt to a variety of materials, in order to justify the investment in kit; data pipelines and storage; space; and staff. For very large collections where the collection is ready and funding available 'up front', out-sourcing to a facility with multiple parallel lines might be a sensible solution to achieve the maximum speed and economies of scale, and to avoid major short-term hiring or purchase of machines that take considerable space and could later become redundant such as conveyors. A key factor for both outsourcing and inhouse approaches is whether the collections are sufficiently prepared – often imaging workflows can be very quick but can create backlogs if, for example, data mobilisation requires further work to create or clean data such as taxonomic lists. For smaller collections, outsourcing may offer efficiencies or resources that cannot be provided in house; but equally with less challenges of scale and backlog, an inhouse approach learning from the best available workflows may also be able to digitize collections at suitable cost / time. Only institutions can decide what is right in their context for their collections, however DiSSCo should provide support for these decisions, setting out the likely relevant factors, pros and cons. Cost information for inhouse and outsourced approaches is also likely to help in understanding the relative business case.

Recommendation 19: DiSSCo should develop a decision support tool to assist institutions to decide on the optimal strategy for digitization of their collections in-house, outsourced, or mixed approach.

2 Specialisation: Developing new and more efficient digitization solutions requires experience and knowledge of new technologies. While the related research and experimentation can be done at collection-holding institutions, putting the results into operation at large scale requires knowledge of industrial engineering and methods not always found in such institutions (or taking time and resource to develop). It might be necessary for DiSSCo to establish some thematically specialised DiSSCo Centres of Excellence (DCE) in key areas (3.9.5). This could be organised through competitive bidding targeted to interdisciplinary consortia. Some of these centres could perhaps be combined with out-sourced digitization factories.

Recommendation 20: DiSSCo should launch early calls for consortia to establish specialised Centres for Excellence on mass-digitization, in readiness for entering the operational phase of DiSSCo.

3 Managing 'flow' versus addressing backlogs; and sharing costs: Institutions have different scales of mass digitization backlogs, and of acquisition rates. For the largest collections, clearing backlogs through mass digitization is an activity with no end yet in sight, and acquisition rates may also be at a scale requiring ongoing mass digitization approaches. For others, however, clearing their backlog may only last for a defined period of years, after which mass digitization equipment is likely to be depreciated, and eventually become redundant or only used at lower scale. Rates of acquisition may not be high enough to support outsourced approaches. The [ICEDIG MS44] study concluded that there is an increasing interest in developing in-house digitization capacities, especially to internalise ongoing digitization as the normal business i.e., becoming digital by default with a permanent digitization infrastructure. What is needed by many is an efficient small-medium size imaging solution that fits in, for example one office room and can be operated by one person only, ideally across several types of material. A camera over a lightbox may fit these criteria, for example.

Another possibility, after backlog has been cleared at one institution, is to move the equipment to another with a backlog awaiting digitization. This allows costs to be shared and might work best if the equipment is not owned by an individual institution but either shared with other institutions or leased from a specialist company or DiSSCo Centre of Excellence who knows how to move such systems.

Recommendation 21: DiSSCo should promote re-use and/or cost-sharing of digitization equipment across institutions and projects where possible, particularly for smaller collections.

4 Small and private collections: There are many small and private collections across Europe that could, with appropriate support be encouraged and assisted to digitize. However, the approach is different from that for large collections. The incentives and support for this are considered in 3.11 below.

5 Decentralisation: Currently there are three facilities for outsourced mass digitization in Europe, run by private sector businesses: Picturae[36] at Montpellier (FR) and Heerhugoward (NL), and Bioshare[37] at Joensuu (FI). There are presently no facilities jointly owned/shared by collection-holding institutions themselves. Picturae's facilities can annually digitize one million herbarium sheets each and Bioshare's half a million. Both are developing insect digitization capabilities. Dinarda[38] are a recently established non-profit association with the purpose of operating and promoting the digitization of natural science collections, specifically insects as standard 3D models published in an archive on the Internet. Since there have so far been no satisfactory, routine solutions for the digital modelling of insects, Dinarda has developed a new scanner (DISC3D) as an open project, with the aim that it be used at numerous locations in the future.

When DiSSCo goes into full operation, we can expect a digitization rate of about 40 million specimens annually. While much of that will be done in-house, more capacity from the established out-sourced facilities will be demanded, and new facilities might be established in other parts of Europe. Another option, as has been done by the Smithsonian is to actively cultivate a collaboration of imaging equipment suppliers working with photographic or manuscript imaging companies to build expertise in natural science and other specimen types – i.e., by taking a more active role, a marketplace for outsourced solutions might be encouraged sooner rather than waiting to see if these occur organically. There are, of course several companies in Europe that might be interested in such an approach.

6 DiSSCo's role: DiSSCo will provide central infrastructure for data, and guidance and standards for digitization, but will DiSSCo own and operate mass digitization systems and services? This has not been decided yet. Many other RI's do own and control their central facilities but in the DiSSCo case there are likely to be constraints on the movement of physical specimens, especially across national boundaries. In-house systems would clearly belong to their respective institutions, but it would be the DiSSCo Centres of Excellence and their digitization facilities (if any) whose ownership and role need discussion. Four service clusters (digitization, programme, infrastructure, data management) can be identified as necessary for DiSSCo to offer, as outlined in Table 6.

Table 6: Four service clusters of digitization-related services identified for DiSSCo

Digitization	Programme	Infrastructure	Data
• High-throughput imaging services	• Training • Case studies	• Data preservation and storage solutions	• Quality Assurance • Audience/user insights

[36] https://picturae.com/en/.
[37] http://www.bioshare.com/.
[38] https://www.dinarda.org/disc3d.

• Boutique digitization • Specimen logistics management • Digitization cost models • Transcription and translation services • Workflow design • Pre-accession digitization • Pre-digitization curation	• Funding advice, support and coordination • Building human networks • Communications and advocacy • New workflows and techniques • Expert consultancy (e.g., readiness evaluation, behavioral change, standard operating procedures)	• Data policies and standards • Data preservation and storage brokerage • Collections management systems • Project and programme skills and tools • Holding and lending specialist equipment	• Tracking benefits and impact • Data access and discovery platforms (public and/or research) • Data enhancement services • Dealing with aggregators (e.g., GBIF nodes)

These service clusters fit at different organisational levels within DiSSCo – institutional, national, regional, or pan-European. The programme and data service clusters are best suited to being organised at the pan-European level, while digitization services would best fit at institutional level and in regional collaborations.

Recommendation 22: DiSSCo should design a portfolio of services and support fitting to several organisational levels that supports the ambition to organise and consolidate a distributed system of scientific collections across Europe.

Recommendation 23: DiSSCo should determine the required number, locations and specialisations of digitization (or related) facilities across Europe, including Centres of Excellence where appropriate.

7 Warehousing: Scientific collections are typically housed in museums located in city centres at expensive property. Although largely historical, this choice remains relevant today, as museum exhibitions must be located where large numbers of visitors can be easily attracted. On the other hand, combining exhibitions and collections under the same roof is not always strictly necessary. Collections take space. Industrial digitization equipment and logistics also need space and access. One alternative is to move collections of the 'B' category (e.g., of common species, a lot of specimens present in the collection, or specimens taking up a lot of space) to a dedicated warehouse after digitization. While digitization can sometimes increase demand for collections, it can also satisfy much of this demand digitally if digitization-on-demand can be offered to the appropriate levels of detail. A 'digital by default' access model would take this a step further by taking a needs-based approach to access. In such a model digital access is provided as the norm where possible and physical access is provided only for research etc. that cannot be accomplished digitally. These are all considerations when looking at where scientific collections can be effectively, securely and cost-effectively housed. 'Warehouse' type spaces for collections may also allow for improvements in storage unit standardisation, and in automated and robotic approaches. This has been demonstrated in libraries but not yet in natural science collections at scale. It may also be possible to combine demand-driven digitization with mass digitization at such facilities. It is recommended that DiSSCo explores the possibility to work with one or more relevant institutions to build and operate such a facility for an 'out-of-town' fully automated, industrial scale specimen storage and digitization facility.

Recommendation 24: DiSSCo should consider building an experimental facility (DiSSCo Centre of Excellence) for an 'out-of-town' fully automated, industrial scale specimen storage and digitization facility.

3.9.5 Centres of Excellence for harmonising approaches in DiSSCo

Answers to the above questions are not 'yes' or 'no' but require optimisation and achieving the correct balance. The questions and their answers are inter-dependent with one another, with different institutions perhaps coming to different conclusions as they balance off the factors.

DiSSCo must offer structured leadership and in digitization and data management approaches, helping institutions reach the right decisions for themselves whilst operating within a consistent framework of standards and best practices, and understanding the full range of choices available to them. This means aiming to develop and disseminate the highest standards of best practice, delivering training for staff, and helping institutions achieve the best digitization results they can in terms of quality (best, according to agreed specifications), time (highest throughput, fast), and cost (lowest, minimal per specimen) for each kind of digitization activity.

Recommendation 25: For each kind of digitization and collection type, DiSSCo should offer structured leadership in digitization approaches, proposing best practice approaches to its institutional members, and helping them to achieve the best digitization results in terms of quality (best, according to agreed specifications), time (highest throughput, fast), and cost (lowest, minimal per specimen) for each specific kind of digitization activity.

Such leadership can be achieved through a 'centres of excellence' model (considered in detail in [ICEDIG MS45]) in which small teams of dedicated individuals collect and craft new knowledge, develop and embed competencies, and train and disseminate to others. Each centre will possess specialised resources (machinery and trained staff) dedicated to each narrow area of digitization speciality giving them capability not only to develop new techniques but also to taken on outsourced digitization activities.

Topics for DCEs can be found by looking at the areas where rapid progress is needed and is viable, given the current state of technology readiness. Topics requiring significant expertise pooling and resources would also be suitable, as would a focus on addressing workflows not yet developed to mass scale. DCEs could vary from offering training to active digitization support, or in any combination in between. The ICEDIG project was already organised along these lines, but as a design study ICEDIG did not build anything operational. However, gaps of knowledge and lack of solutions were found in several areas such as robotics and automation of 3D imaging (WP3); AI/ML in image/data analysis and transcription (WP4); involvement and integration of private collections (WP5); and very-large-scale data management (WP6). The next phase of DiSSCo must investigate these areas further with the aim of developing operational mass-digitization solutions. This is the role of specialised DCE consortia.

3.10 Digitization-on-demand

3.10.1 Characteristics

Intuitively, it's not enough to just work to mass digitize an entire backlog of specimens, at a relatively shallow level of data mobilization (3.9). In contrast to mass digitization, demand-led/driven digitization (also known as digitization-on-demand) is based on specific requests for selected specimens, and/or on requests for a deeper level of data (e.g., additional images or full georeferencing).

Nevertheless, digitization-on-demand should where possible make use of mass workflows. It can also help to develop and refine new mass workflows; and if adopted as a 'digital by default' access route to collections can at some point reach mass scale (i.e., as the level of demand/requests rises) and become efficient, affordable and cost-effective.

Matching digitization to contemporary research needs by prioritising and selecting what to digitize is one very important mechanism for directing mass digitization and maximising the effectiveness of resource allocation. Making digitization available 'on demand' in response to specific requests is also essential to meet immediate needs of the community, and to manage access to collections in a world where

collections (and user) resources are often very stretched. Similarly, digitization-on-demand should also be flexible enough to include dealing with newly collected material, which is often planned in association with current research projects in progress.

Demand-led and research-ready digitization thus has several facets to it, including:

a) A framework of prioritisation criteria to examine the feasibility of digitization and the order of groups to be tackled during mass digitization;

b) Incentives for digitization of private collections – to encourage inclusion of specimens held in private collections into ongoing digitization efforts; and,

c) Offering digitization-on-demand for selected specimens or groups of specimens, to deal with specific research needs including those requiring more data fields that typical mass approaches – which can also include dealing with newly collected material;

d) How to efficiently pull out selected specimens from a large collection and establish suitable criteria for digitization-on-demand – it may not always be cost-effective to digitize e.g., very few specimens for a single research project. Some wider benefit may need to be demonstrable (as with SYNTHESYS+ Virtual Access, which aims to address demand from the community of researchers aligned to major challenges);

e) Collection Digitization Dashboard (CDD) service, offering an online accessible location where information about the extent of digitization across European institutions and collections is brought together and kept up-to-date, and where Collections Descriptions can be used to identify holdings where specimen level digitization is needed.

These are each considered in the following subsections.

3.10.2 A framework of prioritisation criteria

For digitization by DiSSCo to be successful it is essential to have a framework of transparent criteria that enables a demand-driven prioritization of the digitization of natural science collections.

Collection relevance, economic relevance, scientific relevance and social relevance are all important categories of criteria, with those of scientific relevance being the most frequently mentioned as used and important for prioritizing digitization of natural science collections[39]. This is especially true for criteria linked to stimulating fundamental research, research focused on understanding natural sciences processes and trends, and for enhancing access to primary data that support a wide-range of study types (taxonomy, ecology, evolution, extinction and/or climate change, food security, animal-mediated illness[40], for example) where there is wide agreement on this as a principal *raison d'etre*. Relevance of collections, especially in terms of having digital data about important specimens (historic, fragile, type, etc.) and digital availability promoting usage of the collection is also seen as an important. With both scientific relevance and collection relevance closely intertwined, a framework of prioritisation criteria should be based on both i.e., it's important to determine both the relevance for fundamental and other research and/or whether specimens are data-important from the collection relevant perspective.

Decision trees to assess feasibility, scoring methods to assess relative importance and expert panel review for well-informed decisions on relevance are all appropriate component tools for establishing and operating such a framework, to which a clear, harmonized common digital (scientific) research (see 3.1 above) is also an essential input.

[39] ICEDIG project deliverable D2.1, doi: 10.5281/zenodo.2579156 identifies more than 100 criteria collated across these four categories and investigated for their importance through a community survey during 2018.
[40] COVID-19, Ebola, Plague, hantavirus diseases, etc.

Note: The planned SYNTHESYS+ Virtual Access (VA) work will build implementation experience in this area. The VA review panel will have to prioritise digitization-on-demand applications through a set of criteria to be determined by that project.

Recommendation 26: Based principally on scientific relevance but also considering collection, economic and societal relevance and feasibility / cost-effectiveness, DiSSCo must establish a framework of prioritisation criteria and a set of tools and procedures for making and objectively justifying consistent digitization prioritisation decisions.

3.10.3 Offering digitization-on-demand for selected specimens

A Digitization-on-Demand (DoD) service to deal with specific requests from scientists (both individual and groups) and others to digitize selected specimens will be an important high-value service that DiSSCo can offer. Such a service also offers a means of easily and quickly dealing with newly collected material. DoD services can sit alongside other (mass) digitization initiatives.

The SYNTHESYS+ Access programme is presently (during 2020) trialling a 'digitization-on-demand' (DoD) model that can form the eventual basis for DiSSCo DoD services and workflows[41].

Deciding what data to make available from DoD, or indeed designing how to accommodate a broad range of different kinds of data requests will be an important element of the design. Digital data from DoD services can potentially include basic/regular data (cf. MIDS levels 1, 2), digital photographic images, 3D scans and other complex image types, digital molecular or chemical data, etc. Valuable lessons will come from the previously mentioned SYSTHESYS+ Access programme work on which kinds of digital data are truly valuable.

Substituting DoD for loans of specimens is an important use case here. When a loan request arrives, DoD could be offered first. If that is not enough, only then would specimens be sent. Some institutions have already stopped loans of certain kinds of material (e.g., type specimens) owing to risk. A DoD practice would enable such material to be used digitally, as well as serve as a digital backup in case of loss or damage.

Recommendation 27: DiSSCo should design and promote Digitization-on-Demand services and workflows appropriate to different collection and specimen categories that can be adopted by collection-holding institutions to become part of their normal business of digitization, including for accession of newly collected materials.

3.10.4 Pulling out selected specimens

Most major collections are organised by a biological taxonomy. This makes it easy to pull out selected specimens for DoD based on belonging to a specific taxonomic group. However, pulling out specimens on any other criterion, such as geography, time, collector, etc., is slow, difficult or even practically impossible.

Pulling selected specimens out of a collection for on-demand digitization creates a tracking challenge, both in the collection itself (cabinet, drawer, etc.) and the collection management system to know what has been digitized and what has not. This must be addressed as part of the introduction of DoD processes.

Establishing a good inventory across entire collections at MIDS-1 level would dramatically improve this situation (as well as providing a comprehensive cohort of Digital Specimen data for indexing by the ECOI). Such an inventory can be made in several ways, such as mass imaging with no manual transcription, or with minimal transcription exploiting OCR (Optical Character Recognition) and AI (Artificial Intelligence)

[41] https://www.synthesys.info/access/virtual-access.html.

assisted image analysis (see 4.6.3), imaging full drawers, rapid cataloguing of new accessions, and digitization of historical field notebooks. Such an inventory facilitates DoD. In fact, mass imaging and DoD are not opposites but facilitate each other.

Recommendation 28: DiSSCo institutions should consider quickly creating MIDS-1 level inventories of their entire holdings to facilitate access to specimens and planning of more detailed digitization activities, and to create a comprehensive cohort of Digital Specimen data.

3.10.5 Collection Digitization Dashboard

There is no single location where information on the extent of digitization of European natural science collections is gathered; where it can be easily consulted and used by decision-makers and scientists. Such information can be helpful when planning digitization (either mass or DoD) of certain taxonomic or geographic parts of collections. The most suitable way to gather and present this information is as a visual dashboard supported by an underlying service that collects, transforms and collates the data.

Different kinds of visualisation are possible, with analytical dashboards having greatest complexity and needing large amounts of collected data to drive them. Initially, DiSSCo's dashboard focus is on collection-level information. The Collection Digitization Dashboard (CDD) will be an online visual dashboard presentation that makes European natural science collections visible and discoverable. Highlighting institutional holdings and contributions, their strengths and weaknesses, it primarily presents high-level collection data (ref. 2.3.1) for general communication, for future digitization planning and as a data discovery tool. A CDD can also assist a collection to discover its own uniqueness.

Two prototype Collection Digitisation Dashboards were explored in the ICEDIG project, also in conjunction with a Roundtable discussion. The first is available via the DiSSCo website: https://www.dissco.eu/network/ (near the bottom of the page) and the second through The Netherlands country page on the DiSSCo website: https://www.dissco.eu/network/.[42] Their design has been based on analysis of needs of expected user groups (collection-holding institutions, researchers and collectors, education, policy and decision-makers including research funders, non-governmental nature and natural heritage environment organisations, etc.). Two principal collection classification schemes characterise natural science collections in a standardized way at a metadata level. These are 'taxonomic' classification and 'storage' classification that exist in parallel and are based on a scientific view or a collection-managers' view, respectively. For further description of geodiversity collections, a third parallel 'stratigraphic' classification is used. In addition, 'geographic' and 'digitization' classifications are used to further characterize the spatial coverage and levels of digitization of the collections. The most important informational elements included in the CDD are institution, country of institute, 'taxonomy', geography and digitization.

The Biodiversity Information Standards (TDWG) organisation is presently working on a new standard for Collection Descriptions[43] that will facilitate automated metrics using standardised collection descriptions and/or data derived from specimen datasets (e.g., counts of specimens), and a global registry of physical collections (either digitized or non-digitized). The DiSSCo CDD must be compatible with and implement such standard(s) when this becomes available.

[42] At the time of completing the present document (March 2020), the Dutch country page on the DiSSCo website was not yet available, pending the agreement of all the national institutions in the Netherlands. In the meantime, a temporary link to the mentioned CDD is this one: https://app.powerbi.com/view?r=eyJrIjoiNDQ1YmQzMzMtNmY5YS00MDQzLWI5M2YtNmRhOTM2MTg2NTU0IiwidCI6IjhjZDI0OTg0LTBhYTMtNGZjNS1iMDliLTRkNmVjZmFhNThmYiIsImMiOjl9.

[43] https://www.tdwg.org/community/cd/.

Recommendation 29: DiSSCo's Collection Digitization Dashboard (CDD) service must be compatible with and implement TDWG Collection Description standard(s) when this becomes available.

Further work is needed now to fine-tune and improve the prototype dashboard implementations based on relevant feedback from the main user groups (which must be obtained), and to move from prototype to robust, sustained service. This includes automating the collection, transformation/collation and presentation of the data as much as is possible for efficiency and reliability reasons. However, for an interim period it is likely that a manual data entry mechanism will be needed whereby collection-holding institutions can login, enter and adjust their own data on an institutional page, which is then rolled-up to the aggregate dashboard pages.

Recommendation 30: The service underlying the Collection Digitization Dashboard should automate as much as possible the collection, transformation/collation and presentation of collection-level information from collection-holding institutions. For an interim period, manual data entry may be necessary to ensure early public availability of collection-level information (i.e., while work to complete automation of data collection is in progress).

Note: The SYNTHESYS+ project has a task dedicated to "integrate[ing] and expand[ing] institutional collection assessments", where this work can be further pursued. SYNTHESYS+ can help by contributing to a design that is scalable, elegant, easy to link to collection management systems (CMS), and employs the use of people identifiers (ORCIDs) (and their roles) to enhance the usability and automation possible. These developments can also furthered by the MOBILISE Cost Action.

CDD styling ('look and feel') should adhere to DiSSCo common design specifications (see 3.12.1.5).

3.11 Incentives for digitization of private collections

Results from a survey of over 1,000 private collection owners carried out across Europe in 2018[44] suggest the number of specimens owned by these 1000 private collectors lies between 9 and 33 million. The overall number is certainly much higher, as not all private collectors were reached or responded and there were big differences on how the various countries were covered. Even with limited data about size, composition and usage, private collections are known to be important. They hold a significant potential to add to the growing amount of digitized specimen data[45]. However, if not approached carefully there are dangers in just adding to the growing backlog of specimens to be digitized and in holding back those progressive collection-owners that wish to digitize rapidly.

65% of the respondents surveyed said that they already manage their collection data fully or partly electronically. Over 90% are interested in sharing their data in some way, preferably through a public website, and by listing metadata of their collection in a public register. Most private collection owners (55%) indicated they need tools, such as a dataset template or a web-based digitization platform, together with guidelines (36%) and physical equipment (27%).

Thus, small/private collections will need to receive more attention in the future to identify how these can be most easily incorporated within DiSSCo, both to benefit the community at large and the collections themselves. Potentially, European small/private collections could be digitized with very complete records

[44] Presented in ICEDIG deliverable report D2.2.

[45] ICEDIG project deliverable D2.2, doi: 10.5281/zenodo.2582995, based on survey results and other data provides information about the characteristics of private collections in Europe.

within a matter of a small number of years (e.g., 5-7) if approached correctly and supported, say with CMS-as-a-service and an appropriately configured DCE (3.9.4 (2), 3.9.5).

Efforts to help private collection owners to digitize their collections should focus on providing information on how to get started with digitization and encouraging deposition/registration via the European Collection Objects Index (ECOI) service.

Recommendation 31: DiSSCo should provide guidelines for how private actors can digitize their collections and share data via the ECOI service and should ensure that the European Collection Objects Index (ECOI) service offers catalogues of private collections.

Among private collections owners, digitization and data sharing activities are considered important overall. These activities are already undertaken in some cases. For those owners thinking to digitize their collections, various kinds of new support from DiSSCo (e.g., tools, platform, training) would be of great help, and might act as an extra incentive to begin. Such support should be targeted initially towards those private collections of special significance and relevance. Lack of time is also often mentioned as a reason not to have started with digitization, suggesting that use of additional volunteers can offer an important additional means to encourage private collection owners.

Accounting for differences between collection types, their owners and the owners' motivations, communication strategies towards private collections owners should focus on the benefits and mechanisms of digitization and sharing of data online. As well as informing and educating, communications should offer access to appropriate digitization guidance, training and the available supporting tools. In this context it is of utmost importance to be clear about the meaning of the term 'digitization', and what is involved and expected. This could be explained, for example as achieving successively more comprehensive levels of digitization over time by aiming for specific levels of minimum information published about and by a private collection (cf. MICS, MIDS for open access, section 4.4).

Recommendation 32: DiSSCo should develop a package of support measures (communication of benefits, education/training in digitization, digitization tools and facilities, access to data sharing platform, use of volunteers, etc.) targeted towards private collection owners in line with digitization prioritisation decisions, to increase digitization of these kinds of collections.

There are many variabilities and uncertainties about private collections, their owners and the extent of existing digitization so maintaining flexibility in communications and support will be essential. Because there is currently no complete, up-to-date list of private collections, it can be difficult to choose the most appropriate channels for communicating. It is also important to bear in mind that besides private collection owners there are other stakeholders (associations, museums, international organisations) that can play an important enabling role in bringing private collections online. Any plan to coordinate this effort should include policy development that discusses and develops strategy for moving private collections into institutions at some point when suitable. To capture only the data without planning for long-term care of the specimens as well is short-sighted. And, it risks giving the impression that the specimens aren't needed once the data is captured.

As part of its remit to publish minimum information about available collections, DiSSCo should maintain an online inventory (e.g., a website) of available private collections and their characteristics, with an associated protocol for keeping this up to date. Note, that this could be implemented as a standalone webpage/site and/or as a specific filter/facet of entries in the European Collection Objects Index service (see 2.3.1 and 4.2).

Recommendation 33: As part of its remit to publish minimum information about available collections, DiSSCo should maintain an online inventory (e.g., a website) of available private collections and their characteristics, with an associated protocol for keeping this up to date.

3.12 Software engineering, deployment and operations

3.12.1 Software sustainability and maintenance[46]

3.12.1.1 Adopting off-the-shelf components

The decision to depend on specific software is as important as any decision to depend on a specific scientific instrument and should be taken equally carefully. The ICEDIG decision to favour CORDRA object server software, the Handle system and the Digital Object Interface Protocol (DOIP) as key components for implementing the adopted architecture (4.1.1 below) is made with full knowledge of the track record and pedigree of the organisations (cnri.org, dona.net) behind these initiatives. This is backed also by evidence of success provided by several well-established leading-edge use cases that include journal article and dataset identification by the International DOI Foundation (doi.org) and the film and television industry supply chain (eidr.org).

Adopting off-the-shelf components enables quick bootstrapping towards prototype and eventually, workable systems, as is being shown through work implementing the Digital Specimen Demonstrator in the ICEDIG project and the ELViS system in the SYNTHESYS+ project. Valuable experience is being gained. Nevertheless, much new software, which is surely essential for many of the innovative functionalities and services that the DiSSCo vision foresees, is still needed. This is software that will persist, becoming key to the culture and practices of the working community it serves. As such, it must be engineered with care as it will be mission critical.

3.12.1.2 Investing for software development

Adequate investment, and time for development, testing, deployment and maintenance, must not be underestimated. Nor must it be thought that such software will just magically happen without active initiatives to create, train and enable the software engineering team that will be needed (3.12.2 below). Proper design and robust implementation within the overall steerage and constraints implied and imposed by the adopted architecture (4.1.1 below) will be essential.

We could provisionally estimate that 8-10 trained engineers full-time over 4 years (i.e., 32-40FTE) plus one senior technical manager/principal architect are needed to bootstrap DiSSCo with the software critical for its first two years of operation. The minimum cost of this is €4m; more (~€6.4m) if experienced, skilled engineers are to be used. Competing with the private sector for competent software engineers is expensive. Costs going forward beyond first two years of operation are expected to be similar i.e., no reduction in effort is expected due to ongoing maintenance and support needs, as well as the need to introduce new capabilities. Regardless of accuracy and confidence in such an estimate, it provides a clear enough view of the reality to allow DiSSCo leadership to make reasonable decisions and financial provisions about how to control the construction to meet its targets [McConnell 2006]. One and a half million euros per year would not be unreasonable for ICT development and operations.

Recommendation 34: Adequate investment (c. €1.5m per annum) and time (4 years) for new software development, testing, deployment and maintenance must be made if the innovative functionalities and services foreseen by the DiSSCo vision are to be realised.

[46] For readers less familiar with some of the issues covered in this section, the EC-funded ENVRIplus project deliverable D5.1 "A consistent characterisation of existing and planned RIs" is a helpful resource; here: http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf.

3.12.1.3 Strategy for development, maintenance and sustainability

Much has been written elsewhere on the strategies available for software development and deployment, ranging from buying or co-developing software with a commercial vendor, through engaging with relevant open-source projects to carrying out bespoke developments. For DiSSCo, a co-development approach around open-source Digital Object Architecture components as presently being nurtured by, for example the C2CAMP initiative[47] is appropriate, supplemented with bespoke developments specific to the domain.

Several mission-critical 'engineering framework' components remain to be identified as the basis for DiSSCo development of shared infrastructure. This subset must also include common engineering tooling for all developers. Such components, chosen from well-supported open source software must become the standard basis for all developments, thus avoiding that individual engineers choose their own 'favourites'; which can introduce maintenance and compatibility difficulties later. Critical decisions about these software systems must be made carefully, by appropriately skilled engineers to avoid that substantial time/money is not wasted in the future. This is a key responsibility of the DiSSCo Technical Team, supported by the DiSSCo Technical Advisory Board (respectively, 3.12.2.2and **Error! Reference source not found.** below).

3.12.1.4 Timing of development

Overlapping development, deployment and operation will be needed, based on a minimum viable service offering definition. Note, that such development has already begun with the first-generation ELViS system being developed by the SYNTHESYS+ project. The present cost estimate excludes that work. Whereas the DiSSCo programme intends to begin a modest, soft-start to operations in 2024, then development of core components should begin no later than early in 2021.

Recommendation 35: Design and development of core software components needed by DiSSCo should begin no later than early 2021 to allow modest, soft-start to DiSSCo operations in 2024.

3.12.1.5 Common design standards

All software engineering should adhere to common design specifications ('look and feel') for DiSSCo services and components, especially where those have interactive user interfaces. These specifications must be developed.

Recommendation 36: DiSSCo should develop common design specifications, especially for 'look and feel' of interactive user interfaces, that software components and services should adhere to.

3.12.2 Organisation of engineering development and operations (DevOps)

3.12.2.1 Resilience in a community endeavour

How to build a sense of team – of community endeavour and resilience – in a multinational, distributed engineering team? This is the main question facing the organisation of engineering development and operations (DevOps) in DiSSCo. Behind the question lie historical difficulties the biodiversity informatics community has had with weak cooperation across boundaries on software development and infrastructure operations. There have been multiple reasons for this that have included needs to generate personal research outputs as a means of career progression, lack of institutional buy-in to community endeavours and strategies, shortages of cash and resources, and difficulty of sharing funding between institutions. Not least has been lack of appreciation of the importance of using architecture, standards and robust engineering tooling to guide design and implementation coupled with a culture/attitude of 'I want to do it my way'. This leads fundamentally to interoperability difficulties. DiSSCo is a once in a generation opportunity at the European level to change this for the better.

[47] https://github.com/c2camp/core/wiki.

The entire DiSSCo DevOps team must be led to understand and buy into the common goals to be achieved. They must work from the beginning in close collaboration with non-IT staff that includes collection managers, curators, digitization technicians and scientists/researchers. There are many ways this can be achieved including, for example the use of shepherds/buddies for liaison, focus groups and secondments, and co-design methodologies such as user involvement through prototyping [Nieva 2016].

Development teams must understand the steerage and constraints implied and imposed by the adopted architecture of the DiSSCo ICT infrastructure (4.1.1 below), and the tools and selected components to be used. Individuality, competence and autonomy of team engineers are to be granted, but membership, collegiality and generosity must be demanded and strongly reinforced to encourage that effort is focussed as DiSSCo management desires and directs. Creating a strong, motivating 'wiifm' (what's in it for me?) for team members is essential.

Example wiifm (what's in it for me?) for a potential engineering team member: Well-architected, engineered and operated software makes it easier to support users and to introduce enhancements and new functionality as their needs evolve. While using standard tools and development/support practices that include respecting architecture decisions and guidelines to protect and enhance DiSSCo's essential characteristics, team members have the freedom to express their individuality and competence as professionals to support and interact with users and develop new software, contributing autonomously and responsibly to collective endeavours to meet present and future needs.

The leadership approach to be taken must be based on developing the desired culture through appropriate staff appointments, training and tooling (including harmonisation of tooling) that facilitates teamwork. This will need investment of effort by DiSSCo to reduce costs over the longer term.

Recommendation 37: DiSSCo must establish, train and equip a motivated and cohesive engineering development and operations (DevOps) team where team members have the freedom to express their individuality and competence as professionals to support and interact with users and develop new software, contributing autonomously and responsibly to collective endeavours to meet present and future needs.

Software engineering, deployment and operations (DevOps) are overseen by the DiSSCo Technical Team (day-to-day) and Technical Advisory Board (strategically).

3.12.2.2 DiSSCo Technical Team

A DiSSCo Technical Team is responsible for developing the direction of DiSSCo technologies and plans, and for the day-to-day management and direction of software engineering, deployment and operations across the portfolio of projects in the DiSSCo Programme.

Key responsibilities of the DiSSCo Technical Team include:

a) Making informed decisions about the choice of technical sub-system components, supporting frameworks and tooling in the context of the steerage and constraints provided by the DiSSCo ICT Architecture concept, DSArch (4.1.1 below);
b) Selecting and promoting the standards (data, protocol, etc.) that should be adopted and implemented across DiSSCo Facilities and DiSSCo Hub;
c) Ensuring that technical choices and decisions fit sympathetically to complete the picture puzzle of needed development and operations;
d) Ensuring adherence to common design standards ('look and feel') for DiSSCo services and components, especially where those have interactive user interfaces.

Complementing the Technical Team, a DiSSCo Technical Advisory Board (TAB) (6.2.4.2 below) provides strategic and expert consultation services in all areas related to the technical sphere of operation of DiSSCo.

4 Architecture, tools and technologies

This section develops the maturity of the technical concept for DiSSCo digitization and data management infrastructure, and describes the technological innovations needed for constructing the digitization infrastructure of DiSSCo. It synthesises the various aspects, such as machinery, ICT architecture, imaging techniques, etc., describing possible design alternatives for each process and technical element.

4.1 Technical concept for data management

4.1.1 DiSSCo Digital Specimen Architecture (DSArch)

The information and communication technology (ICT) infrastructure supporting DiSSCo data management principles is based on exploitation of three modern approaches to ICT architecture design that combine to create the DiSSCo ICT architectural concept we refer to as 'DiSSCo Digital Specimen Architecture' (DSArch). These approaches are:

1. Digital Object Architecture (DOA) [Kahn 2006, Wittenburg 2019a], whereby the data entities to be managed (principally, specimen and collection data and the various kinds of transaction associated with those) are represented as digital objects, each having a persistent identifier, metadata description and a type definition.
2. The FAIR Guiding Principles to make data 'findable, accessible, interoperable and reusable' (FAIR) [Wilkinson 2016, Mons 2017, Wittenburg 2019b, Lannom 2020] and its derivative FAIR Digital Object Framework (FDOF) that combines the DOA approach with principles of explicit semantic relationships.
3. Evolutionary architecture for guided, incremental change over multiple dimensions with several characteristics protected throughout the DiSSCo lifetime [Ford 2017].

These approaches have been chosen for several reasons:

A. The infrastructure to be deployed by/for DiSSCo is intended to operate over much longer timescales (at least 25-30 years, and possibly longer) than has previously been envisaged for any kind of previous natural sciences infrastructure;

B. Over such long timescales, technologies evolve. Each of the chosen approaches separates the 'what' from the 'how' of the DiSSCo objectives that must be achieved for mass digitization and data management, thus allowing changes in technology to be adopted more easily;

C. The infrastructure affects and will be contributed to by multiple collection-holding institutions in DiSSCo. These institutions already have their own internal ICT infrastructures and procedures, and already expose their data to the outside world in a variety of ways. The chosen approaches allow deployment of DiSSCo infrastructure initially as a centralised (hub) deployment to which existing institutions' infrastructure can interface. Potentially, later migration to a fully decentralised infrastructure can take place without substantial change of architecture.

4.1.2 Principal components of DSArch

4.1.2.1 Digital Object Architecture (DOA)

Several technical approaches were considered as the basis for DiSSCo ICT infrastructure, including the approaches of Semantic Web (Linked Open Data, RDF, Triples) and Object Reuse and Exchange (OAI-ORE, aggregations of Web resources described by resource maps). Historical and current patterns of infrastructure development show these as evolutionary steps in the technology of Web infrastructure. Much more interesting is an emergence of new data architectures. Such approaches include decentralised

applications (d-apps) enabled by blockchain technologies, data intensive federations and marketplaces, and Digital Object Architecture (DOA). The last of these is considered from the research data perspective as a new kind of data fabric – the Internet of FAIR Data and Services. Differing from all other alternatives, it is a fundamental extension of the basic Internet architecture[48], responding to the 'Big Data' explosion in scientific research that has been in progress for the past two decades. With its own communication protocol (Digital Object Interface Protocol, DOIP), Digital Object Architecture (DOA) sits alongside Web approaches [Kahn 2006, Weigel 2017]. It is gathering strong interest from multiple ESFRI research infrastructures across Europe as the means of implementing the European Open Science Cloud (EOSC). DOA is the principal component of DSArch and is DiSSCo's choice reflecting this trend and its basic need to be able to efficiently manage research data pertaining to natural sciences specimens as 'specimens on the Internet'.

4.1.2.2 The FAIR Guiding Principles and FAIR Digital Objects

Two decades of scientific research based on 'Big Data', coupled with political movements towards open access to publicly funded research has led to recognition of the need to make scientific data increasingly findable, accessible, interoperable and re-usable (FAIR). These four attributes form the basis of the now widely adopted FAIR Guiding Principles [Wilkinson 2016]. DiSSCo intends to take an active approach to data management planning and stewardship, with focus on achieving maximum accessibility and reusability of data according to these core principles, longevity of data and data preservation, community curation, linking to third-party information and reproducible science. The FAIR Guiding Principles [Wilkinson 2016, Mons 2017] and their intrinsic support by DOA [Lannom 2020] are manifested as FAIR Digital Objects (FDO) [De Smedt 2020] through a Joint Statement on a FAIR Digital Object Framework, which DiSSCo Coordination and Support Office (DiSSCo CSO) and DiSSCo Technical Team both endorsed (Appendix A). Thus, this represents the second principal component of DSArch.

4.1.2.3 Evolutionary architecture with protected characteristics

Several characteristics of DiSSCo data management are essential to protect throughout and ultimately beyond the lifetime of the DiSSCo data infrastructure for engendering community trust in the value, veracity and reliability of the data to be managed. These are described in detail in the provisional Data Management Plan for the DiSSCo infrastructure [DMP 2019].

Recommendation 38: Nine characteristics (centrality of the digital specimen, accuracy and authenticity of the digital specimen, FAIRness, protection of data, preserving readability and retrievability, traceability (provenance) of specimens, annotation history, determinability (status and trends) of digitization and securability) must be protected throughout the lifetime of the DiSSCo research infrastructure.

Nevertheless, considering the expected lifetime of DiSSCo ICT infrastructure, it is inevitable that the infrastructure, its design and implementation will evolve over its lifetime – both to meet new needs from users and organisations but also as underlying technologies change. The 'evolutionary architecture' approach [Ford 2017] recognises and addresses such evolution by assigning protected status to dimensions considered essential (the characteristics mentioned above) to the integrity of the infrastructure over the very long-term. This is the third principal component of DSArch.

Protecting the essential characteristics mean that proposals for design decisions and changes (technical, procedural and organisational) must be assessed for their effect on those aspects. 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.

[48] See the 'Berlin presentation' here: https://www.rd-alliance.org/rda-11th-plenary-joint-meeting-ig-data-fabric-wg-research-data-collections under agenda item '2.3 Digital Object Principles' by Wittenburg/Strawn.

Recommendation 39: 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.

This implies the need for a formal and responsive change management procedure that includes analysis and assessment of proposed design changes against each of the protected characteristics with sign-off by a delegated authority, such as a DiSSCo Technical Team. This should include, as explained by [Ford 2017] the use of fitness functions for each protected characteristic to provide objective integrity assessments of proposed design changes.

4.2 Implementation strategy

4.2.1 Action steps and phasing

The current landscape of ICT infrastructures of the DiSSCo collection-holding institutions is a fragmented patchwork of heterogenous proprietary systems reflecting specific needs, policies and procedures of individual institutions. This includes extensive use of proprietary specimen identification schemes. Drawing these heterogeneous approaches together towards more harmonised infrastructure to support Europe-wide services cannot be a revolutionary 'out with the old, in with the new' style of change. A sympathetic but strategic approach based on continuous adaptation is needed. In the short-term, this should yield immediate big benefits for a few small changes, whilst minimising disruption to existing systems and procedures. Over the longer term it should encourage convergences towards similar ways of working across multiple institutions.

The starting point for DiSSCo construction recognises that many collection-holding institutions today are already engaged with:

- Programmes of collection and specimen digitization that operate on a variety of levels from basic and manual data entry tasks recording the existence of collections and specimens to highly specialised, partially automated, high-throughput digitization lines that lead to extensive image and data capture for the contents of entire collections;
- Museum catalogues based on computerised databases that make use of institution codes, collection codes and catalogue numbers for indexing digital information about specimens in collections;
- Established and emerging institutional data portals, making collection and specimen data available for external use e.g., via Web browsers and programmatic interfaces (API);
- Publishing institutional data to GBIF, for example as Darwin Core Archives via the GBIF Integrated Publishing Toolkit;

In the future, the DiSSCo virtual collection of specimens will be indexed by the European Collection Objects Index (ECOI) service. The indexed data about collections and specimens can be served in multiple formats (human readable, JSON, XML, RDF, etc.) to meet the various needs of both native and non-native processing applications that will include new Europe-wide services such as the Collection Digitization Dashboard (CDD), European Loans and Visits System (ELViS), European Curation and Annotation System (ECAS), as well as many others. On creation, individual digital specimens will each be given a globally unique Natural Sciences Identifier (NSId) that acts as a permanent and unambiguous reference and an anchoring point on the Internet to all the information known about a specific specimen over decades. It will become increasingly possible to create and identify digital specimens at the point of digitization in digitization projects and programmes, at the time of accession into collections, and even earlier at the time and place of gathering in the field. Curation by the community experts, long-term preservation of the digital specimen data, and positive feedback to points of origin will become the norm, as will publishing to aggregators such as GBIF by DiSSCo on behalf of its member institutions.

The action steps of an implementation strategy and the phasing of that strategy must support both innovative collection-holding institutions that want to advance more quickly and those acting more conservatively in moving towards DiSSCo goals. For the near future, a key assumption is that museum catalogues remain as core business of the collection-holding institutions with retention of control over the authoritative information about specimens in their collections. However, within ten years, this is expected to change towards a model of curation by acknowledged and appropriately authorised community-experts outside of any specific institution.

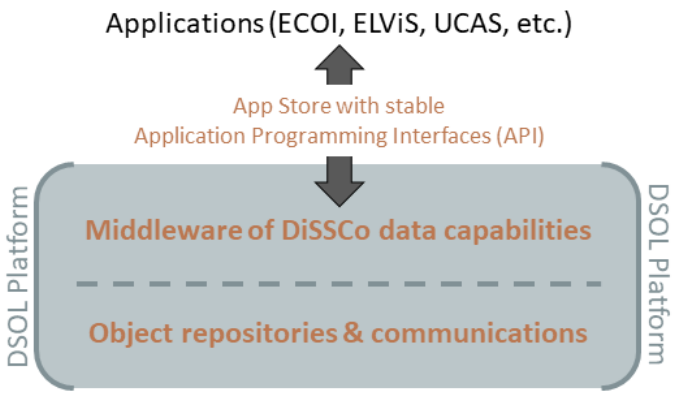
Recommendation 40: Within ten years the institution-centric collection curation model should evolve to support complementary digital curation by appropriately authorised community-experts.

DSArch should thus be implemented aggressively to allow soft-start operations to commence at the earliest opportunity (3.12.1.4) via a multi-stage construction approach around two key lines of activity: i) hub infrastructure indexing Digital Specimen and other object types, and offering added value services such as ECOI, ELViS and ECAS; and ii) data coupling to populate the initial systems/services with relevant data.

Recommendation 41: DiSSCo Prepare should follow an aggressive ICT implementation strategy and construction plan based on two key lines of activity that include i) DiSSCo Hub implementation indexing Digital Specimen and other object types, and offering added value services such as ECOI, ELViS and ECAS; and ii) data coupling of DiSSCo Facilities to populate the systems and services with relevant data.

4.2.2 Hub infrastructure

At launch, the DiSSCo Hub infrastructure is a platform[49] comprising a middleware of data capabilities and object repositories (as illustrated in Figure 2) that together constitute a Digital Specimen Objects Layer (DSOL). Below this (and not shown in the figure) is a layer of virtualisation, the data coupling (4.2.3) that connects the collection-holding institutions. And above, application programming interfaces (APIs) allow services like ECOI, ELViS, ECAS, etc. to interact with the Digital Specimens and other object types stored in the object repositories.



The DiSSCo platform sits between layers 1 and 3:
1. Applications users can access and launch
2. Digital Specimen Objects layer – the platform
3. Virtualisation layer, connecting collection-holding institutions (not shown)

Figure 2: The DiSSCo Hub data platform

Figure 3 begins to illustrate how the notion of he DiSSCo Hub data platform manifests as a set of application programming interfaces (API) set against the phases of the DiSSCo research data lifecycle, ultimately manifesting as the illustrative system design below.

[49] In systems engineering terms, a platform is a bed or core of stable hardware and software components upon which a variety of functions and services can be built to serve users. Evolvability to adjust to changing needs and demands is controlled to ensure protection of vital characteristics (see 2.3.1 and 4.1.2.3).

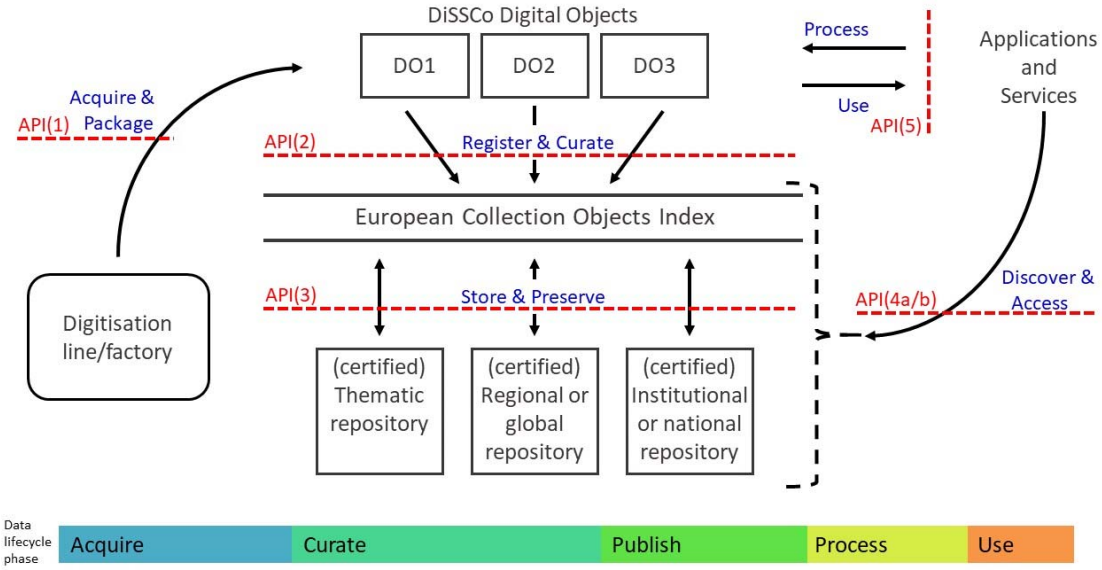


Figure 3: From Hub platform to APIs, data lifecycle and system design

DiSSCo Hub infrastructure will be implemented with a system design based on multiple instances of the CORDRA[50] object repository software as the core elements of the DiSSCo data platform, as shown by the illustrative design in Figure 4. Separate CORDRA instances index different DiSSCo digital object types, such as Digital Specimens, Digital Collections, Provenance Events and Annotations.

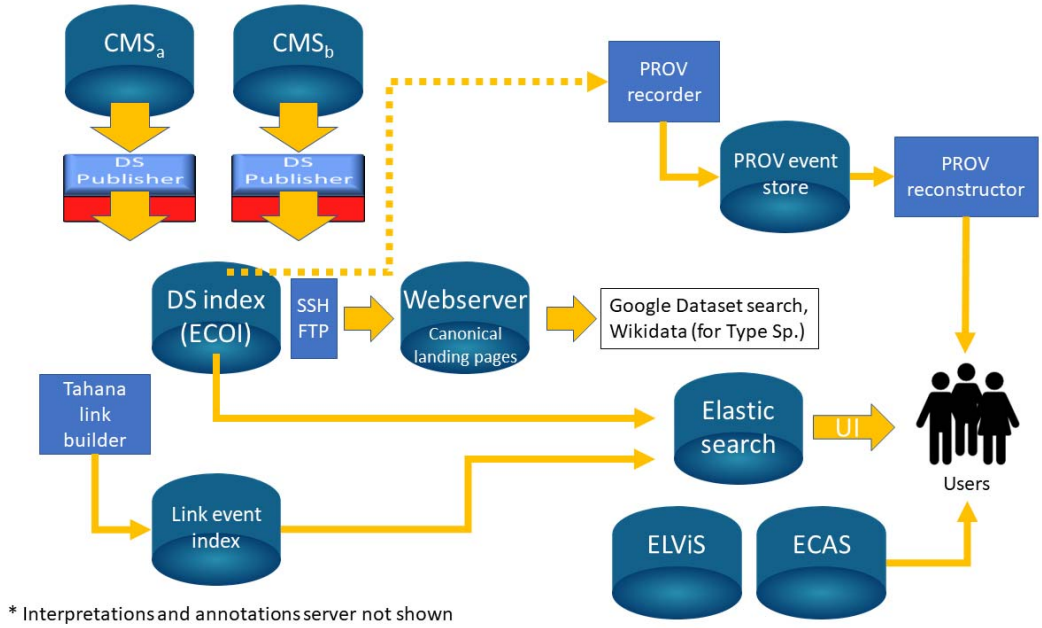


Figure 4: Illustrative system design based on multiple CORDRA object repository instances

[50] https://www.cordra.org/cordra.html.

Enrichment of Digital Specimens either/both at time of creation or/and later is made possible by the algorithms of the 'Tahana link builder', a software component that finds and exposes the third-party data associated with natural science specimens to build the DiSSCo knowledge graph. Services such as ECOI Elasticsearch make it possible to query and find data of interest across multiple object types/repositories. Each Digital Specimen will have its own human-readable Web 'landing' page, as well as being both machine-readable and machine-actionable[51]. Selected important Digital Specimens, such as type specimens for example, can be promoted to external resources such as Google Dataset Search and Wikidata.

Built-in DSOL behaviours (operations) act on Digital Specimen objects directly, for example to update/modify them, to enrich them with new data, to tag and classify them, to access their images and rights and to query across them. The built-in behaviours are to be standardised as extended operations in the specification for open Digital Specimens (see 4.3).

A DSOL API allows service developers to build demand-led applications that can make use of and extend the built-in behaviours to interact directly with Digital Specimens. One example of this is the eventual integration that can be made with the SYNTHESYS+ Specimen Data Refinery (SDR)[52]. This tool combines artificial intelligence (machine learning, computer vision) and human-in-the-loop methods to extract, enhance and annotate data from digital images and records of natural science specimens at scale. Being able to directly update the data of Digital Specimens with the data extractions from images immediately increases data quantity and value.

Recommendation 42: DiSSCo Prepare should specify the Application Programming Interface(s), API needed to allow third-party software applications to be built on top of the DiSSCo Hub (core) infrastructure.

4.2.3 Data coupling

Data coupling begins with harvesting existing Darwin Core Archives or ABCD-A files published by collection-holding institutions (for example, using respectively the GBIF Integrated Publishing Toolkit, IPT or the BioCase Provider software), and transforming their content to create Digital Specimens from the records contained within those Archives[53]. During the process, newly created Digital Specimens can be immediately enriched with basic third-party data, such as Species2000/IT IS Catalogue of Life, Plazi TreatmentBank, varied literature sources, EMBL European Nucleotide Archive, etc.

Recommendation 43: DiSSCo should commence further development and hardening of the specimen data harvesting and transformation process as the means of creating Digital Specimens and populating DiSSCo data infrastructure.

Initially, this is a one-way publication or data sharing process based on proactive harvesting by DiSSCo Hub. It will need to accommodate expected changes to the way data is published by institutions, for example phasing out support of Darwin Core Archives in favour of Frictionless Data packages. Later, data coupling moves towards a more sophisticated bidirectional interfacing for circular near real-time updates of records in the CMS of collection-holding institutions, based on (for example) digital curation and annotation activities that take place around the Digital Specimen itself.

[51] Being machine-actionable (i.e., knowing how to process the data in specific circumstances) is more than being machine-readable (i.e., understanding the data, its meaning and context).

[52] See, for example https://osf.io/bsyac/ and doi: 10.3897/biss.3.37647 including its accompanying Powerpoint presentation, https://biss.pensoft.net/article/37647/download/media/372189/.

[53] Exploratory work in the ICEDIG project has developed prototype software capable of reading Darwin Core Archives and transforming their content records into Digital Specimens. This software can be found at https://github.com/DiSSCo.

Bidirectional transactional updating will be piloted first between DiSSCo's own CMS-as-a-Service and ECOI and later be packaged and rolled out as an interface component for adoption by DiSSCo Facilities.

4.2.4 Beyond the initial phases

Later in the evolution of the data infrastructure involves change from a centralised DiSSCo Hub deployment to a more decentralised deployment whereby collection-holding institutions become more involved and responsible players in operating the core elements of a distributed DiSSCo ICT infrastructure. Although many years away, this is mentioned here as the overall trajectory to keep in mind for the DiSSCo lifetime.

4.3 Innovations needed for data management infrastructure

4.3.1 Bringing technical innovations to required readiness level

Multiple innovations are needed for the implementation and operation of the DiSSCo data infrastructure based on the proposed approach and direction (2.3). Many of these have been worked out at high level during the progress of the ICEDIG project. They are listed in Table 1 – Table 4 (pages 28 – 30). A few of specific importance require substantial further work during the DiSSCo Prepare Preparatory Phase project to bring them to the necessary level of technical, organisational and financial readiness. They are highlighted in the following sub-sections 4.3.2 – 4.3.5.

Recommendation 44: DiSSCo should ensure and provision the further work needed (during the DiSSCo Prepare Preparatory Phase project) to bring necessary innovations needed for data management infrastructure to the required level of technical, organisational and financial readiness. These innovations include: i) NSId PID scheme, ii) openDS standard, iii) MIDS/MICS minimum information standards, iv) FAIR Digital Object Framework

In addition to the key innovations mentioned above, further technical work is needed on many infrastructure building blocks that include (for example):

- **Metadataschema(s):** to be used and how these are EOSC compatible;
- **Authentication, Authorization and Accounting (AAA) infrastructure:** AARC-based, FIM4R compatible;
- **Attribution model**: Research Data Alliance (RDA) recommendations for the representation of attribution metadata[17]; extending W3C PROV;
- **Data packaging:** of multiple specimens, for example;
- **Index hierarchy / DiSSCo Data Model:** 3 x object type hierarchies (resource type, storage, annotation and provenance – latter to be based on notion of event objects) and type definitions for each of the various object types. Extensions to PID record (kernel information);
- **DiSSCo specific ontology:** rooted in BCO.
- **System design:** Finalisation of system design based around APIs and CORDRA technology – an outline exists (4.2.2) but review and approval is needed by DiSSCo TAB, as well as GA endorsement;
- **DiSSCo Type Definitions and Extended Operations:** Actions needed to specify the types, the extended DOIP operations and DiSSCo APIs.
- **DS enrichment strategies:** Further work on, and Tahana link builder software.
- etc.

Maximising flexibility of DiSSCo to support unimagined uses of the data in the future implies an approach based on a foundational platform, APIs and an app-store for DiSSCo scientists/users. Visually, this can be represented as explained/illustrated in section 4.2.2 above.

4.3.2 NSId PID scheme

A decision on adoption of an appropriate persistent identifier scheme, meeting the requirements set out in Appendix B is needed as soon as possible. Lack of decision constitutes a substantial barrier to beginning pilot operations of the DiSSCo infrastructure because of the implications of permanently and persistently assigning identifiers to digital objects that might be difficult to change later.

A proposal for a Handle-based scheme such as the Natural Science Identifier (NSId) scheme proposed by ICEDIG (5.3.2.7) must be made to the DONA Foundation for a persistent identifier scheme to be used by DiSSCo. This is imminent and urgent, as there are other initiatives in other places (IGSN, EUPS, Extended Specimen Network, etc.) with political implications.

4.3.3 A standard for open Digital Specimens (openDS)

A first draft specification of a standard for open Digital Specimens (openDS) is needed as soon as possible. A MOBILISE COST Action workshop took place in Warsaw, Poland 11th/12th February 2020. The outcomes of that workshop must be acted upon. Specifically, what is needed to support open Collections is an essential first step.

4.3.4 MIDS/MICS minimum information standards

Standards covering the Minimum Information about a Digital Specimen (MIDS) and Minimum Information about a Digital Collection (MICS) are needed as soon as possible. A first draft MIDS specification document (v0.8, http://bit.ly/MIDSv08) exists but is incomplete. It needs further improvement, validation and agreement. Also see the related discussion in ICEDIG project deliverable D6.5. Little work exists so far for MICS, which also needs to align to both CDD and TDWG CD work. Also needs implementation and evaluation work. The current expectation is that version 1 of the TDWG CD standard might be available by September 2020, with a version of the GBIF World Collections Catalogue available by then as well. DiSSCo is also planning its own Collection Description Identifier Registry (CIDR) as part of its ECOI service.

4.3.5 FAIR Digital Object Framework (FDOF)

Finalisation and stabilisation of the FAIR Digital Object Framework (FDOF) must be achieved during 2020, including addressing and solving the inherent security of digital objects, i.e., by ensuring relevant identification and authorisation mechanisms are embedded in FDOF. An important open issue is to ensure common understanding of the difference between being machine-actionable and being machine-readable, and to ensuring that FDOs (and by extension, Digital Specimens) are machine-actionable[51]. Machine-actionable first is a core principle.

4.4 Open access guidelines

As noted in 3.8.3, the legal framework around open access to research data cannot be ignored by DiSSCo and its member institutions. A detailed analysis by the ICEDIG project of the existing data policies from several collection-holding institutions[54] concludes that natural sciences data must be open, findable, accessible, interoperable and reusable by default. This leads to several principles that should each become incorporated into the open access policies of collection-holding institutions. These are explained in the following sub-sections.

4.4.1 Minimum information standards

Minimum information standards (such as the proposed standards for Minimum Information about a Digital Specimen (MIDS) and Minimum Information about a Collection (MICS)) can assist with mobilising incomplete data. Even partially available data can be useful for some purposes and it is recommended that data and media be made openly accessible after minimal delay. Digital Specimen objects must be

[54] ICEDIG project deliverable D6.5, doi: 10.5281/zenodo.3465285.

findable and accessible already at the lowest (MIDS-0) level. Collection information must be findable and accessible, even at the level of overview information (MICS-1 level).

Recommendation 45: Open access policies of DiSSCo and its collection-holding institutions should include that Digital Specimen objects must be findable and accessible, even at the lowest level of available information (MIDS-0 level). Collection level information should be findable and accessible, even at the level of an overview (MICS-1 level).

4.4.2 Use of public data repositories

For long-term storage and archival of image data, which involves storage of many files with large size, institutions (and DiSSCo as a whole) have choices of institutional, national, EUDAT, Zenodo or other public repositories and choice can vary from one institution to another (see 3.7.1, and also 4.7). It is a matter for each DiSSCo institution to determine the best choice according to its own needs. Nevertheless, such choices must keep in mind the need to achieve and maintain FAIRness i.e., to comply with the FAIR Guiding Principles as set out by DiSSCo policy and the data management plan for the DiSSCo infrastructure [DMP 2019].

For non-image data (meaning principally the authoritative data about specimens and collections, as well as pointers to supplementary data)[55], DiSSCo requires it to be maintained and indexed in one logical place, named as the European Collection Objects Index (ECOI), where that data can be stored and processed consistently. Physically, ECOI will be a single, central index (as planned initially). In the future it can become a federated decentralised index distributed among the DiSSCo partner institutions if that is desirable for operational reasons.

Recommendation 46: Open access policies of DiSSCo and its collection-holding institutions should include that: i) image data and its immediate metadata should be deposited in a trusted public repository of the institution's own choice, and ii) that other (non-image) data should be deposited in the European Collection Objects Index (ECOI).

In separating storage of image and non-image data, care must be taken with duplication of data/metadata in multiple places to avoid future synchronisation difficulties. Best practice is that media asset management systems used for long-term storage of image data should contain minimal metadata pertaining only to the characteristics of the image itself and should avoid metadata stating anything about the content of the image and what it may represent. When using public data repositories such as EUDAT, Zenodo, Wikidata and others, where often there is a requirement to submit metadata about the content of the item deposited (images, in this case) careful thought should be given to the reasons for choosing such repositories, the nature and status of the items to be deposited there and the likelihood of future modification of the metadata.

Recommendation 47: DiSSCo should develop guidelines for its member institutions for determining whether and when it is appropriate to deposit specimen data, such as images of specimens in long-term public repositories such as EUDAT, Zenodo, Wikidata and others, having regard both for the purpose of such depositions and for the stability of the metadata describing the content of the deposition (i.e., what, where, when, who).

[55] See section 5.2.3 of the provisional data management plan for the DiSSCo infrastructure, [DMP 2019] for definitions and scope of authoritative data and supplementary data.

4.4.3 Unrestrictive licensing, as open as possible

As far as possible, projects must enable third parties to access, mine, exploit, reproduce and disseminate this data by using a copyright waiver such as CC0 or an open access licence such as CC-BY.

Recommendation 48: Open access policies of DiSSCo and its collection-holding institutions should include that as far as possible third-parties must be able to access, mine, exploit, reproduce and disseminate data by using a copyright waiver such as CC0 or an open access licence such as CC-BY.

However, there are essential legal and ethical reasons for restricting access to data. Thus, DiSSCo policy on open data access should be "as open as possible, as closed as (legally) necessary".

Recommendation 49: Open access policies of DiSSCo and its collection-holding institutions should include that access be "as open as possible, as closed as (legally) necessary".

Exceptions to the open as possible policy must be stated clearly and must be justified strictly according to objective criteria defined by national security, legislation or other regulatory compliance, sensitivity of collection information, and third-party rights (such as personal privacy). Restrictions that do not have a justification based on objective criteria in legislation are legally invalid and not permitted (see also 3.8.2).

Recommendation 50: Exceptions to the 'as open as possible' data access policy of DiSSCo and its collection-holding institutions must be justified based on objective criteria, stated clearly and strictly limited to reasons of national security, legal or regulatory compliance, sensitivity of collection information, and third-party rights.

Robust technical mechanisms must enforce an "as open as possible, as closed as necessary" policy. However, such mechanisms must not place onerous and convoluted obligations on data publishers nor must they decrease usability for users. Digital Specimen objects must be secure by design. The precise mechanism to be used remains for further study[56] but could, for example be based on ciphertext-policy attribute-based encryption (CP-ABE) [Bethancourt 2007] whereby suitably authorised users have a key that unlocks sensitive data encrypted by the owner/publisher. Alternatively, but less desirable is an approach based on stretched-perimeter access control, whereby 'policy enforcement points' are instantiated to the places where access policy is to be enforced [Burnap 2012]. Note, however that neither approach addresses the inherent trust issue, which is that once access has been given, there is little that can be done to prevent further (malicious) dissemination of the controlled sensitive data. At that moment, a disciplinary or legal recourse becomes the only viable response.

4.5 Service portfolio management

DiSSCo should undertake its service management responsibilities in compliance with a lightweight IT Service Management (ITSM) framework focussed on the holistic organisational/business perspective of service delivery. This will help to ensure that all the individual business aspects of delivering services portfolio in a professional, robust and reliable manner – including the upskilling of staff and increasing the digital capabilities of DiSSCo Facilities – are properly taken into account through defining, producing, and providing the required services, and responding to changes in the environment in which they are offered. Such a framework provides a mechanism to align ICT needs with governance and business model, provide a process-oriented approach with clearly identified process owners and managers, emphasise avenues

[56] Inherent security of Digital Specimens has been identified as a fundamental topic of the FAIR Digital Object Framework that will be addressed during 2020. See section **Error! Reference source not found.**.

for continual improvement, training and awareness. In short, provide a tool to manage a full-service life cycle from concept to deployment to daily operation.

There are two candidate frameworks appropriate to consider[57]: VeriSM and FiTSM.

Recommendation 51: DiSSCo should adopt a lightweight ICT service management framework for the holistic delivery of its service portfolio.

4.6 Digitization design alternatives

4.6.1 Mass imaging (2D)

4.6.1.1 Microscope and other slides (D3.2)

Microscope slides form part of natural science collections in herbarium, museums and many other collection-holding institutes. They are unusual compared to the other preservation types or collections covered in this section as they are rarely curated as separate collections but stored as supplementary collections alongside a range of "classical" collection categories including entomological (both as whole slide mounts and preparations of parts like genitalia), botany, zoology, palaeontology and mineralogy.

The specific preservation methods, labelling practices, dimensions and storage are very variable. It is probably due to these properties that there have been limited mass-imaging methodologies published and considered for slides in general. While there have been several pilot projects that have used specially modified histology slide scanners adapted for natural science specimens, they cannot accommodate damaged slides or slides with non-standard thickness or length - issues that can be frequent in natural science collections.

Within ICEDIG two mass-imaging workflows were developed to digitize microscope slides as discrete collections with adaptable methodologies for inventory (MIDS level 0-1) [Allan 2019] and more detailed specimen-level digitization [Allan 2018].

Recommendation 52: For mass digitization of microscope slides, the recommended approach in the first instance is to digitize on the lowest MIDS level (0,1) capturing an image of the whole slide including its labels. These images can be used later for extended data entry (MIDS 2,3), conservation assessment and subsequent research-grade imaging.

Like other collection types, the labels on microscope slides are well suited to automated image processing and information extraction techniques like OCR to capture label data. Due to the spatial constraints of the labels they are potentially more amenable to these techniques than other collection types (see 4.6.3).

4.6.1.2 Vertebrate and other dry three-dimensional collections (D3.3)

Mass-imaging of vertebrate material (skins, bones, nests, eggs, etc.) and other dry three-dimensional collections (molluscs, fossils, etc.) has had limited take-up to date. Currently only 2.9% of phylum Chordata preserved specimens in GBIF have an associated image (furthermore, a significant subset of these images is images of the catalogue, not the object itself). These collections are smaller in number and are not as easy to handle and position as herbarium sheets and microscopic slides. No existing automated imaging solutions have been found for these types of specimens, although conveyor-driven imaging lines commonly used for herbarium sheets could (in principle) be rigged to image anything that fits on a tray narrower than the conveyor belt (typically 60 cm) and covered by the camera focus depth of field (maximum 25-40 cm)[58]. Due to the variability of object types, shapes, volumes and research interest,

[57] https://verism.global/ and https://www.fitsm.eu/.
[58] ICEDIG project deliverable D3.3, doi: https://doi.org/10.5281/zenodo.3364385.

mass-imaging for scientific purposes is not considered to be realistic in many cases (with exceptions perhaps for smaller, flatter objects). Imaging of specimen labels for data entry and databasing purposes is deemed workable. We recommend an approach to image every label combined with quick minimal data entry (i.e., MIDS level 0 - 1) from the images. Considering the expertise needed to interpret label data, risks associated with multiple handling movements, the limited range of most data capture and the extended potential of label images, this is an effective approach. The label images can then be used, during a follow-on phase, for more detailed data entry.

Recommendation 53: For mass digitization of vertebrate and other dry three-dimensional collection objects, the recommended approach in the first instance is to digitize on the lowest MIDS level (0-1) combined with label imaging. These images can be used later for extended data entry (MIDS 2-3).

The extended potential of the images of labels lies in the use of optical character recognition (OCR) and handwritten text recognition (HTR) (and as training for those purposes). This approach also allows the use of thesauri for quicker data entry with less transcription errors, while still being able to check the data against the original label without the time-consuming task of retrieving the object.

Because this approach uses imaging for labels and subsequent data entry, inclusion of the specimen in the image is optional. Reasons to include the specimen comes from the added potential for curation, research and narrowing down loan requests, as well as future developments such as AI for species recognition.

It is acknowledged that there can be cases where imaging is so complex or where data entry can be done so fast that this approach is not efficient. The GBIF task force on accelerating discovery advised a tiered strategy with rapid and least expensive steps during the first phase, with a second phase for more detailed data capture and imaging [Krishtalka 2016]. We propose, instead, that imaging is the fast, initial step that requires minimal expertise. [Nelson 2012] also concluded that data entry from the images was the most efficient way of (large scale) digitization. Setting up imaging protocols and stations for any type of collection should involve a professional (or experienced) photographer to establish the proper lighting and camera settings for the collection. The workflow should be designed in such a way that there are minimal adjustments needed by the operators. For three-dimensional collections, depth of field is especially important which can be maximised through adjustment of various parameters.

There can be an efficiency gain from combining with other tasks: all handling actions have some risk for errors or damage to the specimen/label, so minimising the handling of specimens is crucial. Combining tasks such as barcoding, cleaning, repacking and tissue extraction with digitization efforts can lead to an efficiency gain in certain situations. On the other hand, more complex tasks as part of a larger process can produce more errors than when these tasks are done separately by specially trained workers.

Most imaging of these kinds of collections is done with a camera mounted on a copystand or tripod. For imaging for overview and inventory purposes some existing mass digitization systems can be adapted, such as SatScan type scanners and herbarium conveyor belts, taking into account the great depth of field requirement for these collections. Picturae, for example uses a configuration for imaging that photographs from two sides simultaneously. This could be adapted for imaging vertebrate and other dry three-dimensional specimens and their labels from two opposite sides.

Recommendation 54: Setting up imaging protocols and stations for digitizing any type of collection should involve a professional (or experienced) photographer to establish the proper lighting and camera settings for the collection.

4.6.1.3 Liquid preserved specimens (D3.4)

Few institutions hold wet collections on a 'mass' scale (as defined elsewhere in the present document) if compared with other collections types such as herbarium sheets or insects. Mass digitization of samples stored in liquids has rarely been done, as their handling and transportation requires special care, but also because of the limited view a two-dimensional image gives of these specimens. Handling of liquid specimens comes with several risks and impediments, including fragile containers, and dangerous chemicals. Imaging of liquid collections can essentially be divided into two categories: whole jar/container imaging, and specimen imaging by removal from the container. Limitations occur due to various factors, including (for closed container imaging) distortion by the container, discoloured liquid and specimen, obstructing labels, multiple specimens in a container (sometimes in separate vials) and so on. Liquid samples are present in all life science collections. Generally, almost all the different taxonomic collections within an institute have their own liquid collections, storage protocols and digitization status.

For mass (or rapid and affordable) digitization purposes, it is deemed unrealistic to open containers for imaging the specimen, unless containers are homogenous and easy to open. Label data capture can be achieved through whole container imaging, for a part of the collection. Whole container imaging can give an impression of the specimen. Further, the condition of the container can be recorded. Like with dry three-dimensional collections, the recommended focus is on capturing the label data in the first instance, and any related field books or ledgers. Data entry can be done directly from the object to the appropriate MIDS level. When enough data is legible on the labels without opening the container, the same approach as for dry three-dimensional specimens can be followed: to digitize on the lowest MIDS level (0-1) combined with label imaging. These images can be used later for extended data entry (MIDS 2-3) and allow for the potential uses described for dry three-dimensional objects.

Recommendation 55: For mass digitization of vertebrate and other dry three-dimensional collection objects, the recommended approach in the first instance is to digitize on the lowest MIDS level (0-1) combined with label imaging. These images can be used later for extended data entry (MIDS 2-3).

A test was done of imaging solutions to deal with curved containers. When labels are so large that they are not legible from a single photo because they curve around the jar, then multiple photos are required, and data entry is impeded by having to switch between multiple photos. Stitching of images to create a virtual "inverse panorama" or "rollout" would allow data entry from a single image. In the future, NHMUK plan to test the output with their ALICE software for label extraction [Price 2018]. Picturae has conducted tests that show promising results of virtual rollouts. Using a controlled turntable and capturing either multiple images (36) or a video (2,500 frames) and cropping each image or frame to the central pixel column(s) and subsequent stitching of these, the label data was legible in a single stitched image. Placing the object on a controllable turntable would be simple but is likely to cause difficulties with processing due to differential movement of label and specimen relative to container due to inertia. To avoid the issue of inertia, moving the camera instead of the object would be advisable for further development. These tests should be further developed to allow the use of these methods for mass-imaging of wet collections containers.

A new approach to digitize wet specimens outside of their container has been developed using a flatbed scanner and 3D printed containers [Mendez 2018]. Where camera-based imaging of specimens in liquid suffer from surface reflections and lens distortion, these are non-issues for flatbed scanners because each point is scanned as the scanner is moved across the object. This method is only recommended for small specimens due to limited depth of field. Specifically, the containers were designed with dimensions of 75mm x 50mm x 1mm. This workflow poses a DNA contamination risk as the containers for imaging are used in a workflow for many specimens a day.

In conclusion, some hardware solutions have been proposed to speed up imaging, but efficiency must be mainly sought in workflow design and size of the project. In summary, a lean process is suggested as described for vertebrate digitization.

4.6.1.4 Pinned insects (D3.5)

The main challenges in mass digitizing insects stem from the facts that pinned insects, with their equally pinned labels, are i) basically three-dimensional objects and ii) their numbers in collections are huge; up to one billion objects in Europe. The fastest available digitization systems of today [Tegelberg 2014, Tegelberg 2017, Hereld 2017, Price 2018] can make images of up to 1,000 specimens per working day utilising just one operator. However, a ten-fold increase in the speed of insect digitization from the current state of the art must be sought in order to make enough progress during the next 30 years[59].

An example of one innovation towards this goal is the ENTODIG-3D prototype built by the ICEDIG project, which employs Tensorflow AI/ML image analysis, multiple webcams on rails to make focus-stacked images of insects and labels from various angles and builds 3D models of individual insects and visible parts of their labels [Ylinampa 2019]. Three-dimensional (3D) modelling is increasingly being used to digitize cultural heritage. However, no available solutions have been designed with mass production in mind. To modify them for mass production might be possible but will require redesign.

The big question is: Do we really want to make pictures of 1 billion insect specimens? Would it be better to just transcribe the labels? This is really what most curators prefer, because in many insect groups a picture of the specimen is not enough to determine the species. Examination of its genitalia is required and that cannot be done from a whole specimen image. The question is not yet answered but it can be pointed out that transcription of the labels, if done without imaging of them, must necessarily be done in-house at the institution that owns the collection. This defies the possibility to do the transcription in other countries, where labour costs could be lower and local knowledge of language, geography, and taxonomy often is available. Transcription without imaging also means that users of this data are unable to view and verify the verbatim labels.

Ultimately, it is a question of cost. To reduce costs the number of human operators must be reduced as much as possible. A fully automated line where humans just bring the insect drawers to the digitization line would be ideal. The ENTODIG-3D solution mounted on a conveyor shows how this can work.

Another approach to reduce the costs could be to use inexpensive but fully automated imaging stations. If the capital cost of each is well below €10,000, dozens could be installed in parallel in each museum. Human operators would only bring one drawer to imaging once or twice a day. For a collection containing 50,000 drawers with up to one million unit-trays, ten imaging stations operating in parallel 24 hours per day for twenty years would do the job.

In these fully automated scenarios one problem remains: How to attach unique identifiers such as a QR-code to each physical specimen? It does not take much time to attach them, but still would mean that each insect is handled by a human for a few seconds. One possibility that could be explored is to make the image of the insect itself the unique identifier! An image taken from a standardised position could be hashed, for example into SHA256 file hash checksum, which becomes the practically unique identifier for the specimen. If we look closer, like human faces, no two insect specimens are identical. Great progress has already been made in automatically identifying insect species from each other automatically, e.g., [Valan 2019], so what about individuals? We recommend testing this approach in pilot projects.

[59] ICEDIG project deliverable D3.5, doi: 10.5281/zenodo.3520667 describes state-of-the-art, potential technologies, and three experiments the ICEDIG project carried out to find new innovations in mass digitization of pinned insects.

Recommendation 56: Further explore combinations of new technologies such as robotics, 3D modelling, machine learning, etc., in novel ways to achieve imaging (including of the labels) of 5,000 insect specimens in 24 hours by one workstation and operator.

4.6.1.5 Herbarium specimens (D3.6)

The imaging of two-dimensional objects, such as herbarium specimens, is perhaps the most advanced and well understood area of collection digitization. Nevertheless, the preparation of herbaria for digitization is complex and requires considerable organization and planning. It requires a workforce that may need training to fulfil the needs of the digitization workflow and there are infrastructural needs for digitization that need consideration. For example, space is required for the imaging equipment. A staging area is needed for specimens and for decontamination – a walk-in freezer is the preferred method. Herbaria are housed in a wide variety of buildings, each with their own idiosyncrasies. The width of the doors, the number of floors and the height of shelves all need consideration.

A digitization project often starts with the estimation of the size of the collection and the different types of objects within it. Counts are needed of unmounted specimens, or those needing restoration, also non-standard specimens and those with envelopes that may need opening. Decisions must be made at this point about how digitization is going to be done, whether it is onsite or offsite, how much mounting is going to be done and the timetable for completion. All this information must be fed into the tendering process and attention to detail at this time will ensure the process will run smoother later on[60].

Some processes are peculiar to the preparation of a herbarium for digitization. Pest management must be specifically considered. As all the collection is going to be moved during digitization it makes sense to integrate pest management processes into the workflow. This also ensures that pest problems are not made worse by cross-contaminating collections and moving pests between rooms.

Each institution has a unique combination of infrastructure, user requirements, staffing and funding that mean that no two herbarium digitization projects are alike. However, underlying these projects is the need for a quality digital product, minimum damage to the collection and long-term needs for the collection's preservation. Case studies from European herbaria[61] give some idea of the variety of options and available solutions to digitization. These studies will give other institutions that are embarking on this process some idea of the options open to them and the decisions they must take.

The complexity of herbarium digitization can appear trivial until one starts to examine the complex decisions and dependencies in the process. Often it is easy to concentrate on the imaging equipment and the digital infrastructure. However, personnel and physical herbarium constraints are just as important. ICEDIG project work in this area[61] has, for the first time, consolidated expertise in herbarium preparation for digitization. It is clear from the results that early and detailed preparation and management are needed throughout the process to ensure the goals of an imaging project are achieved and within budget. Each herbarium that has gone through this process is able to point to innovations they made in the process and by learning from these pioneering herbaria, future projects are likely to be easier, cheaper and quicker than they would have otherwise been.

Recommendation 57: DiSSCo should prepare and promote community-built guidelines and checklists for future digitization projects, to assist the detailed preparation, costing and management of such projects.

[60] ICEDIG project deliverable 3.6, doi: 10.5281/zenodo.3524263 examines the process of herbarium imaging and the preparatory steps needed. Some of the issues are common to any digitization project, but each step is important.
[61] Described in ICEDIG project deliverable 3.6, doi: 10.5281/zenodo.3524263.

Fully automated conveyor-driven imaging can produce nearly a terabyte of data in a working day. This data must be quality controlled as part of or soon after the acquisition process, and then be further processed to detect barcodes and other pertinent details from the images. It takes a complex and powerful imaging and data processing system to do these tasks. Such systems currently are proprietary to the industrial service providers of mass digitization solutions. Data then must be transmitted to the institution that owns the collection, and imported in the institution's CMS, possibly through a separate transcription system. DiSSCo Facilities and their service providers must automate this pipeline somehow. Eventually this requires standardisation so that the various steps can be seamlessly integrated, and so service providers can be switched as competitive tendering determines. Ideally, an automated mass imaging system would output MIDS compliant Digital Specimen objects to begin with. It might make sense to create an open-source package to process data from imaging pipelines to harmonise, simplify and bring down the cost of large-scale deployment of conveyor-driven mass digitization lines.

Recommendation 58: DiSSCo should investigate standardization of the interfaces of the components of mass digitization/imaging lines and their downstream data stores and processing elements (including quality control) and encourage open-source tools development in the area.

4.6.2 Manual transcription

4.6.2.1 Available options

Most specimens are documented on paper before their details are entered into a database. In many cases, specimen data only exist as handwritten data on labels, in catalogues or in field notebook. This presents an interoperability challenge between digital and paper records. Accurate and efficient transcription is clearly an essential step towards making these data findable, accessible and useable. Different options exist for the transcription of these data, such as in-house transcription, transcription outsourced to a commercial company, and volunteer crowdsourcing from the general public. Each of these approaches comes with advantages and constraints.

Essentially, the accuracy of all human transcription methods proves to be similar and the choice between them comes down to the options open to the collection-holding institution[62]. For example, outsourcing transcription can be fast, but comes with fixed costs and cannot easily be changed during the process. Crowdsourcing takes longer and is likely to give rise to more issues around data quality and rework. However, it has considerable value in public engagement with the collection.

4.6.2.2 Improving manual transcription

In furtherance of improving manual transcription, DiSSCo could provide additional cross-checking tools and aids. It is clear, for example that data quality can be improved with the use of lookup lists. Speed can be increased two to threefold by comparing the details of specimens being digitized with other previously digitized specimens from the same collecting events [Mononen 2014]. Providing handwriting samples of collectors and determiners is another useful aid, as is use of social media to access expertise of people outside the immediate transcription process.

An interactive nomenclatorial reconciliation service would allow transcribers to verify that names are correctly spelt. However, being able to distinguish, after transcription has been completed between actual (spelling) errors in the data as it was originally recorded, and errors introduced during transcription is essential. Thus, in such discovered cases of original misspelling, the verbatim data must be annotated to highlight the fact. It is essential to retain the data verbatim, as well as to provide it in corrected/improved forms. The quality control and assurance plan of a transcription project or process must be clear on how to proceed in this and similar cases, and community-agreed best practices should be applied.

[62] ICEDIG project deliverable D4.2, doi: 10.5281/zenodo.3364509.

Giving attention to data standardization and total quality in manual transcription pays dividends, especially through the greater use of automation. Focussing on identifying the points in transcription projects, processes and workflows where errors can arise and designing for their prevention, combined with strict, enforced use[63] of data standards and automating selected quality control tests and assertions from those defined by the TDWG Data Quality Interest Group[64] goes a long way to improving results. This should all be documented by a quality control and assurance plan for transcribed specimen data that aims towards ensuring that completeness and quality of transcription aligns to one of the levels of Minimum Information about a Digital Specimen (MIDS levels) that are presently being developed (4.4.1). Standardized data records coming out of a digitization transcription process should indicate their completeness in terms of one of the MIDS levels reached.

Recommendation 59: DiSSCo should work with collection-holding institutions to improve the standardization of transcribed data in line with the emerging Minimum Information standards for Digital Specimens and Digital Collections (MIDS/MICS) and should seek to introduce community-agreed quality control and assurance plan and procedures for transcription.

In the longer-term, the European Collection Objects Index will prove to be an invaluable resource for the correction of errors and the overall improvement of data quality by the application of centralised analysis, dynamic visualization of current specimen data and curation by the community at large.

4.6.2.3 Field notebooks

A special case of manual transcription is digitization of field notebooks. Each major museum collection-holding institution has hundreds of them from the times before the 1950's. With ink pen it was not possible to make small labels for insects, so only collector name and a catalogue number (collector´s field number) was attached to the specimen and all other details were written in the notebook. Only data equivalent to MIDS-1 level data is available from the specimens while the corresponding field notebook has not been digitized. Field notebook digitization projects have been carried out at least by the Smithsonian Institution[65] and the University of Helsinki[66]. Digitizing them is easy and inexpensive, since they require only 2D imaging, and transcription can be done by remote (volunteer) workers.

Recommendation 60: Digitization of field notebooks should be a priority and should precede digitization of any related specimens.

The richness of information in field notebooks can be amazing. It might be a practice worth reintroducing by using electronic means, although it may not be applicable for all collection types. A physical specimen would only need a QR code label as its specimen label. All other data would be in an electronic field notebook, entered for example via a smartphone or tablet computer and almost immediately accessible through the DiSSCo infrastructure.

Recommendation 61: DiSSCo should consider developing and offering a digital field notebook service.

4.6.2.4 Georeferencing

To maximise the use of specimens in research it is helpful to be able to accurately determine the point of collection from the textual information on the label. This process is known as georeferencing and is largely carried out manually today, being time-consuming and costly. There are currently several tools available

[63] For example, by using tools for data entry such as RightField, https://rightfield.org.uk/.
[64] https://www.tdwg.org/community/bdq/, and https://github.com/tdwg/bdq.
[65] https://siarchives.si.edu/about/field-book-project.
[66] http://digit.luomus.fi/

to assist with and to some extent automate georeferencing[67] but the general conclusion is that to make most use of these tools users need to have good data management and basic programming skills.

New techniques from natural language processing and geospatial information analysis can contribute to improved semi-automated and automated methods to speed up georeferencing. These would have impact not only on time and cost, but also on quality and utility of data and the georeferenced specimens that data represents

Recommendation 62: Adopting new techniques from natural language processing and geospatial information analysis, DiSSCo should investigate the improvement of georeferencing techniques for identifying more precise locations from natural language descriptions of locations that appear on specimen labels.

4.6.3 Automated text digitization

Optical Character Recognition (OCR) can be used to digitize text appearing on images of specimens, for example where labels appear in images[68]. However, recognising text quickly and accurately from these images can be a challenge for OCR because the non-textual component of the image pixels often accounts for most of the image content. This can overwhelm OCR algorithms and lead to false positive (faulty) recognition. Performance can be improved by segmenting specimen images into their component parts prior to applying OCR, ensuring that only images that are essentially the text-bearing labels are submitted for OCR processing as opposed to whole specimen images. Image segmentation can be done automatically using a deep learning approach, for example [Nieva *in prep.*]. The open-source Tesseract OCR software version 4.0.0[69] offers promising text recognition accuracy with segmented images.

Recommendation 63: AI-assisted image segmentation should be further developed by DiSSCo for routine use as a step in the digitization and transcription process.

Not all text on specimen labels is nicely printed. Handwritten text varies considerably and does not conform to standard shapes and sizes of individual characters. This poses an additional challenge to OCR but the application of techniques such as those utilized by Google Cloud's Vision AI product[70] has allowed for significant recent advance in this area. With support for 50+ languages and pricing beginning at €1.45 per label recognition, such commercial products can take DiSSCo a long way towards negating the need for humans to routinely transcribe handwritten text.

Recommendation 64: DiSSCo should consider an agreement (operational or strategic) with, for example Google for AI-assisted text recognition of specimen labels, including handwritten labels at industrial scale.

Determining the countries and collectors of specimens has been one aim of previous automated text digitization research activities. An area of Natural Language Processing (NLP) known as Named Entity Recognition (NER) has matured enough to semi-automate this task, recognizing location names and person names within the text extracted from segmented images via Tesseract version 4.0.0. NER can be used in conjunction with other online services, such as those of the Biodiversity Heritage Library to correlate the named entities to places and people mentioned in the biodiversity literature. It can be used

[67] Reviewed in section 6 of ICEDIG project deliverable D4.2, doi: 10.5281/zenodo.3364509.
[68] ICEDIG project deliverable D4.1, doi: https://doi.org/10.5281/zenodo.3364502.
[69] https://github.com/tesseract-ocr/tesseract.
[70] https://cloud.google.com/vision.

with new, emerging georeferencing techniques that can infer locality from sparse and vague descriptions on labels (0).

OCR, automatic language identification and NER, correlation and georeferencing are just a few of the main components needed for improving the efficiency and quality of transcription and the utility of transcription data products. Further work should be undertaken to stabilize appropriate techniques and bring these together into deployable robust, semi-automated and automated pipelines (workflows)[68].

Recommendation 65: As part of the DiSSCo programme portfolio, DiSSCo should sponsor a research and innovation project investigating novel approaches to label segmentation, transcription (i.e., text OCR and digitization) and interpretation (i.e., named entity recognition, georeferencing, people referencing, etc.) and synthesis and deployment of robust production pipelines (workflows) to improve efficiency, quality and utility of transcription data.

4.6.4 3D capture methods

Many systems exist for performing three-dimensional (3D) imaging of cultural and natural heritage artefacts. From these, several rapid 3D methods are considered as potentially viable for integration into mass 3D digitization of natural science objects[71].

Viable rapid 3D methods include multi-plane photography, which is a technique that creates two-dimensional picture sets or stacks, that can subsequently be manipulated by software to give pseudo-3D views of real-life 3D objects. In contrast, other techniques directly create 3D computable models. Photogrammetry uses methods of image measurement and interpretation to derive the shape and location of an object from one or more photographs of that object. Structured Light Scanning uses the methods of stationary fringe projection based on a fixed grid of fringes generated by a projector and observed by one camera. Laser Scanning describes the three-dimensional measurement of the surface of an object by analysis of the reflected light from a laser beam which is scanned over the object surface. For further extensive details on these techniques and their use to date by collection-holding institutions, readers should consult the relevant report[71].

However, moving 3D techniques forward into mainstream mass digitization projects is not presently considered feasible within current resource constraints, and may not even be needed at large scale. It is most likely that specimens to be the subject of 3D imaging will be hand-picked in relation to specific research and exhibition project needs. Nevertheless, there are opportunities within the context of rapid 2D mass digitization pipelines to insert procedures that can identify suitable specimens for 3D treatment and to extract these for further special handling in a separate 3D imaging pipeline. This 3D pipeline itself should be designed to be achieve a balance of speed and efficiency whilst being as cost-effective as possible[72]. From study of existing uses of 3D techniques by several collection-holding institutions[71], it is evident there are opportunities at multiple points in the generic 3D workflow for improvements with potential to increase the overall throughput. Some of these possibilities, given in Table 7 are dependent on the acquisition method whereas others, given in Table 8 are possible later in the workflow where there is no dependence on the acquisition method.

[71] ICEDIG project deliverable D3.7, doi: 10.5281/zenodo.3469531.

[72] See, for example the work of oVert (an ADBC Thematic Collection Network) which is aiming to produce 3D images of more than 20,000 vertebrates: https://www.idigbio.org/wiki/index.php/OVert:_Open_Exploration_of_Vertebrate_Diversity_in_3D.

Table 7: Method dependent digitization tasks with areas/opportunities for improving throughput

Workflow Step	Workflow Task	Improvement Areas	Improvement Opportunities
Acquisition	Specimen handling	Techniques and devices for specimen handling before and after image acquisition.	Automated specimen handling and mounting can create faster end-to-end digitization workflows.
	Specimen mounting	Techniques and devices used for specimen handling during image acquisition.	
	Imaging station setup	Techniques and devices for specimen scanning (rigs, platforms, lighting).	Digitization stations can be implemented with commercial-off-the-shelf (COTS) equipment (e.g., photogrammetry) acquired from outside vendors or can combine different techniques (e.g., using images for colour texture acquisition for laser and CT scans). Optimise number of passes, determine best angles and appropriate lighting. Map the type of primary models to be created and the associated types of derivatives that can be obtained.
Processing	Scanned data processing	Integrate and verify raw data sets. Prepare data sets for building 3D models e.g., removing outliers and duplicates.	
	Build 3D model	Models with higher quality increase the number and types of derivatives that can be created.	
	Build model derivatives	Derivatives clearly defined beforehand can streamline the curation and publishing tasks.	

Table 8: Method independent digitization tasks with areas/opportunities for improving throughput

Workflow Step	Workflow Task	Improvement Areas	Improvement Opportunities
Curation	Model identification	Identification of 3D models.	Curation tasks add information on digital specimens. This can include adding data about the physical specimen, the digitization process, and the 3D reconstruction process.
	Information extraction	Extraction of metadata from digitized specimen (from tags, barcodes, etc.)	
	Specimen annotation	Annotation of digital specimens, including model metadata and specimen data (from catalogues).	
Publishing	3D model publishing	Providing access to and long-time preservation of 3D models.	The publishing processes for digitized materials must follow FAIR principles.
	3D model display	Rendering and accessing 3D models.	Supporting preview of the models and their associated metadata is among the goals of digitization.

The ICEDIG work on rapid 3D capture methods has been distilled to propose a design for an ideal 3D digitization workflow[73] that can be integrated into collection digitization projects. The proposal serves as a high-level model for adding 3D digitization and can be given as guidance to DiSSCo collection-holding institutions that can help to:

- Identify appropriate 3D imaging technologies that can be applicable for each specimen type;
- Determine the attributes of physical specimens that make them candidates for 3D imaging;
- Describe the ways in which the current rates of 3D digitization have been achieved and possible areas of improvement; and

[73] See section 5 of ICEDIG project deliverable D3.7, doi: 10.5281/zenodo.3469531.

- Act as a call for action on establishing 3D imaging metadata guidelines between manufacturers, industry, natural science collection institutions and cultural history collection institutions.[74]

Nevertheless, future directions for 3D imaging in DiSSCo are still dependent on the basic expectations among DiSSCo member institutions and the wider community about the requirements around and prospects for the use of 3D imaged specimens.

Recommendation 66: DiSSCo should identify and further study the scientific needs that demand the use of 3D imaged specimens, as well as the scientific opportunities opened up by the availability of appropriately imaged 3D specimens to inform future planning for introducing 3D imaging on a more widespread basis.

4.6.5 Quality

4.6.5.1 The definition of quality

Effective quality management is essential for producing digital data and images that are fit for purpose. DiSSCo should aim to treat quality management in a simple and straightforward manner, avoiding meaningless terms such as high-/low-quality or good/poor quality. Such terms are relative, vague and without a reference point, impossible to define. We want to treat quality objectively and in absolute terms by defining it as 'conformance to requirements' and not as goodness.

Against such a definition, DiSSCo should be concerned with applying quality management approaches in two fundamental areas, namely i) digitization quality management as preventative work, and ii) data quality improvement as curational work[75].

4.6.5.2 Digitization quality management as preventative work

First and foremost, DiSSCo must ensure that its digitization processes and other services deliver digital data (including images of specimens) fit for scientific purpose and meeting users' needs. DiSSCo must do this without introducing new errors during such processes. Thus, an imaging process, for example should always deliver in-focus images and a transcription process must deliver accurate transcriptions; and both must relate to the requirements of users. This means properly understanding what these requirements are in relation to the purpose users need data for and aiming to satisfy those requirements. This requires proactive management, especially in relation to prioritisation of digitization (3.1) and designing processes to prevent the occurrence of errors in digital outputs. The aim is to produce the right digital specimen or other data product without defects at the right time, for the right cost whilst correctly meeting the users' requirements. Producing outputs with defects (or producing the wrong outputs) leads to waste and rework, both of which have an associated cost.

Prevention (or quality assurance) activities are normally performed by the digitization team (digitization technicians, in-house digitization teams, and the service providers). Paying attention to prevention, continuously improving and automating the digitization process reduces and eventually minimises the need for re-digitization or for post-digitization corrective actions. This improves overall quality of the output and can lead to significant benefits in both collection management and use.

[74] The Community Standards for 3D Data Forum (CS3DP) has done some work in this area. See https://groups.google.com/forum/?utm_medium=email&utm_source=footer#!forum/community-standards-for-3d-data-preservation-cs3dp. This group was created to aid in the organization of nationally shared resources for the preservation and management of 3D digital research outputs and the development of shared community driven standards. Also, some work done by the oVert Thematic Collection Network (footnote 72), especially on archival strategies and access.

[75] Traditionally, these can be thought of as quality assurance and quality control respectively. Quality assurance is process oriented and focuses on defect prevention, while quality control is product oriented and focuses on defect identification.

Analysis of the present capacities of collection-holding institutions (5 below) reveals that quality policies, procedures and standards are commonly lacking, minimal or ad hoc. It is an aspect of digitization that often appears under resourced and lacking in specific expertise.

Recommendation 67: DiSSCo should establish a common quality policy and standard for digitization, including adopting a prevention approach to digitization process and workflow design and appropriate training of personnel.

Philip B. Crosby's four absolutes of quality are a useful framework for thinking about prevention in digitization. The first absolute is that the definition of quality is conformance to requirements. This means a digitization process is a 'quality process' when it is producing digitized outputs that are fit for the purpose the user requires. The second absolute is that the system of quality is prevention, which means digitization process designers and operators focus on preventing errors and mistakes from happening in the first place ('do it right the first time') rather than accepting that it's ok for errors to creep in and to fix them after the fact. The third absolute says that the performance standard is zero defects, meaning not that mistakes never happen but that there is no acceptable number of errors that should be allowed to occur, and that effort should be put into reducing and removing opportunities for error. To achieve zero defects means monitoring and measuring non-conformances in the process and its outputs, and the fourth or final absolute is concerned with that; saying that the measurement of quality is the price of non-conformance i.e., that metrics of such measures should be expressed as a monetary cost i.e., €. Only by taking such an approach is it possible to drive mass digitization costs downwards. Crosby's approach to quality is just one of several approaches to Total Quality Management (TQM) developed over the past fifty years that include: W. Edwards Deming's Deming Cycle, Joseph Juran's 'the Vital Few and the Useful Many' philosophy, Six Sigma, Continuous Improvement and Lean Manufacturing, among others. Although perhaps not so well known these days, Crosby's approach has been shown to be practically effective and straightforward to implement, with the possibility to obtain impressive results.

4.6.5.3 Data quality improvement as curational work

In addition to ensuring that preventative quality management in digitization is effective (above), DiSSCo must provide mechanisms for assessing data to discover and correct errors that exist prior to digitization and/or to enhance elements after digitization takes place. The following are all examples of aspects of data quality that might be improved during curation work:

- Misspellings on labels will pass verbatim through digitization and transcription processes to be identified and corrected separately later;
- Geographic coordinates can be added later by a georeferencing process;
- Completion of empty/unknown fields with data gleaned from another source, such as collector name where this can be identified from similar specimens collected at the same time and place, for example;
- Use of controlled terms/vocabularies and authority files/sources, for example for disambiguated collector names and geo/place names; especially semantic enhancement by linking directly to such sources rather than copying values;
- Alignment of nomenclature;
- Versioning of changes over time, addition of annotations and preservation of verbatim data;
- Improvement and disambiguation of identifiers; addition of globally unique identifiers.

Improvement activities are normally performed by data curation and publishing teams (curators, collection managers, data publishers and data users, data aggregators). From first beginnings of establishing principles of data quality of GBIF data [Chapman 2005], there has been significant progress

in recent years, largely through the work of TDWG volunteers in biodiversity information standards; specifically through the TDWG Data Quality Interest Group[64] that has led to both a conceptual 'fitness for use' framework for quality assessment and improvement of natural sciences data [Veiga 2017] and a data quality solutions library of quality tests and assertions that can be applied to data [Chapman 2020]. DiSSCo should consider adopting such a conceptual framework for quality assessment and improvement of natural sciences data and to selecting and implementing appropriate data quality tests and assertions to remediate data errors.

Recommendation 68: DiSSCo should adopt a conceptual framework for quality assessment and improvement of natural sciences data and should select and implement appropriate data quality tests and assertions.

There is scope for substantial further innovation in tools and processes for improving data quality, for example:

- With new techniques to visualize data and allowing fixing of the mistakes seen – a kind of "using the data to improve the data" approach; and,
- With new dissemination mechanisms for communicating and adopting new approaches and efficiencies in data quality improvement that is as automated as possible, reaching to the entire DiSSCo network (and beyond).

4.6.5.4 Image quality management

Digital specimens composed of images and other data can be considered, for many purposes to be viable research surrogates for the physical specimens from which they have been derived. To be suitable for research i.e., to be of research quality images must meet specific requirements/criteria in respect of a whole range of different characteristics that can include size, resolution, colour space, colour accuracy, focus/sharpness, cropping, metadata and more. Varying requirements can apply for different purposes as well as for different collection types but in general these serve to define the criteria to manage the expectations of digitization processes and guide the acquisition of equipment. Table 9 provides an overview of the image elements that often need to be captured for different types of specimens[76].

Table 9: Overview of image elements

Legend: **R** = Required **C** = Conditions apply **NR** = Not required **O** = Optional

Imaging Workflow	Specimen	Background	Colour Chart	Scale Bar	Labels	Barcode	Institution Name	Other Elements	Conditions
Microscopy Slides	**C**	**C**	**NR**	**NR**	**C**	**R**	**NR**	**O**	● Specimen can be hard to capture without special equipment due to size. ● Background is generally white to facilitate viewing of slide elements. ● Labels can be placed on both sides of the slide, requiring additional images per slide. Special type labels are important for classifying specimens.

[76] For more detail refer to each of the specific workflow sections in the ICEDIG project deliverable D3.1, doi: 10.5281/zenodo.3469521.

Imaging Workflow	Specimen	Background	Colour Chart	Scale Bar	Labels	Barcode	Institution Name	Other Elements	Conditions
Skins and Vertebrate Material	**R**	**C**	**C**	**C**	**R**	**R**	**O**	**O**	● Background must maximise the identification of the specimen, avoiding glossy or reflective materials that can hinder border detection. ● Some specimens may not require colour chart as colour is not a main feature (e.g., bone samples). ● The placement of the scale needs to consider the depth and angle at which images are acquired.
Liquid preserved specimens	**C**	**C**	**O**	**O**	**C**	**C**	**O**	**C**	● The containers can contain more than one specimen and require additional handling. Sometimes the specimens are removed and imaged outside the container, but this take longer time. ● Background must be neutral, especially for see-through containers. ● Barcodes can refer to one container with multiple specimens. ● Labels can be hard to image due to the shape and placement in container. ● Paper records which describe the specimens in a container will need to be digitized as well.
Pinned Insects	**R**	**C**	**R**	**R**	**R**	**R**	**O**	**O**	● Background must maximise the identification of the specimen, avoiding glossy or reflective materials that can hinder border detection.
Herbarium Sheets	**R**	**NR**	**R**	**R**	**C**	**R**	**R**	**O**	● Labels can be hard to image due to overlapping with other labels and with the specimen parts. Additionally, some labels can be placed at the back of the sheet, requiring additional imaging.
3D Specimen Models	**R**	**C**	**C**	**NR**	**C**	**C**	**O**	**O**	● Background must maximise the identification of the specimen, avoiding glossy or reflective materials that can hinder border detection. ● Colour charts can help if model includes colour or texture. ● Labels may be captured separately, but some cases such as pinned insects, 3D scanning may help in rapid imaging while minimising specimen handling. ● Barcodes can be part of the label set

Ultimately, there is always a trade-off between image quality, time to produce and cost. However, aiming to produce images against more stringent requirements whilst meeting constraints of time and cost makes images potentially more useful for a wider range of purposes over the long run. This is because once such images have been created, derivatives of them meeting reduced (less stringent) requirements can be easily produced.

In several domains of heritage digitization, quality criteria (such as size, image resolution, colour, etc.) have been developed for digitization of printed materials (pictures, documents, books). To an extent these can be re-applied for collection materials (such as microscope slides, and herbarium sheets) that are close to two-dimensional (2D) representations, as well as to digitization of supplementary materials such as labels, physical registries/catalogues, and field/lab books, etc[77]. Similar generic requirements can be applied across different collection types, e.g., relating to image bit depth, colour space, colour accuracy,

[77] ICEDIG project deliverable D3.1, doi: 10.5281/zenodo.3469521. Recommendations are organised by collection type i.e., for i) microscopy slides; ii) skins and vertebrate material; iii) liquid preserved specimens; iv) pinned insects; v) herbarium sheets; and vi) 3d digitization. Recommendations cover: quality criteria to be met, the digitization workflows, the quality assurance activities to be performed during digitization, the quality control activities that are performed after digitization, and the software and hardware tools that can be used to support these quality management activities.

metadata. Other requirements can include file naming, maximum file sizes, dealing with duplicate imaging, and how to avoid badly cropped images.

However, requirements/recommendations for imaging of three-dimensional (3D) objects are much more variable and less well harmonised. The current literature on quality management of 3D models is more focused on the model structure (point cloud, mesh, polygons) than on the colours and illumination of the object. Nevertheless, 2D quality management methods can be applied to the 2D source images to ensure that the derived model textures are accurate. More importantly, in the case of photogrammetry, 2D image quality methods are needed to ensure the accuracy of the photogrammetric 3D model.

Currently, image quality is predominantly verified by visually inspecting a sample of the images in a digitization batch. Besides taking place 'after the event', this is neither sufficient nor practical when thousands of images are produced daily.

As explained above (4.6.5.2) prevention must be practised. Imaging requirements and meeting them through accurate operation of digitization processes and pipelines avoids unnecessary work and cost later. The aim should be to prevent errors in the first place and to be able to detect when these are happening. Automated, computer-assisted checks using computer vision techniques – for example to detect inappropriate cropping, and to quantify image sharpness, contrast, evenness of lighting, colour balance, proper detection and reading of barcodes, etc. – combined with statistical process control can avoid critical process elements drifting outside established tolerances. Instrument calibration (camera exposure, scanning speed), illumination, conveyor speed, etc. can all be monitored and controlled these days.

Recommendation 69: DiSSCo should harmonise and establish quality requirements for image characteristics for common image classes expected from digitization and should aim to prevent quality defects arising in digitization processes through the use of automation, computer-vision and statistical process control techniques.

A DiSSCo DCE (3.5) might be proposed to specialise in this area.

4.6.6 Use of automation and robotics

Several case studies of automation from e-commerce and the library sector were explored as well as two examples of a robotic arm in the heritage sector to provide an understanding of the current state of robotics and automated hardware as relevant to natural science collections for digitization, handling and storage[78]. Robots and other automated systems are very good at repetitive tasks, but natural science collections are mainly heterogeneous, organically shaped and fragile, which is not easily compatible with automation of manual handling.

Warehousing automation can improve retrieval times from storage, space efficiency and climate control. However, implementation of automated warehousing solutions requires significant adaptations of existing storage space or designs for new builds substantially different from existing collection storage solutions. Automation is likely to be greater interest when new storage spaces are being built where, for example shelving units compatible with robot operations must be used.

Presently, a fully automated solution from storage to imaging and back to storage is not realistic for the complex context of natural science collections. However, by developing independent components (including storage and retrieval, transport, object picking, and imaging) that can be connected in the future, progress can already be made towards an end-to-end solution. For the imaging component, due

[78] ICEDIG project deliverable D3.8, doi: 10.5281/zenodo.3719101.

to the great variety of natural science collections it is inevitable that multiple automated imaging systems are needed.

Recommendation 70: To work towards a future automated end-to-end digitization solution, development should focus on independent components (including storage and retrieval, transport, object picking, and imaging) which can be connected in the future.

The natural science sector will need to work with commercial partners to develop and integrate such components. This is most likely to involve working with SMEs[79] specialising in the sector as they have the required knowledge and experience but also because the small potential market size makes the sector relatively challenging and financially uninteresting for larger players. On the other hand, larger players can bring a wider range of expertise and resources to bear. Natural science collections have very different requirements than other sectors where automation and robotics is applied, so companies need to be provided with clear requirements and information, as previous experiences show. Competitions and tenders will need to allow for bidders to see collections in action and to ask questions to gather information and experience that they require for development.

If there is a real desire to use automation solutions in natural science collections for warehousing and/or digitization, then some steps are needed on the part of the natural sciences community. We propose that a series of pilot projects is established, which SMEs can participate in.

Recommendation 71: If there is a real desire to use automation solutions in natural science collections for warehousing and/or digitization, a series of pilot projects should be established, which companies can participate in and in which collection holders collaborate.

DiSSCo, and its Centres of Excellence, could play a further role in developing the expertise to better communicate with companies. Also, DiSSCo can lead a concentrated effort for research and development in this field, to ensure the various pilot projects are aligned.

Recommendation 72: DiSSCo should play a role in the development of expertise in automation, in communication with companies, and aligning efforts in automation of natural science collections.

4.7 Long-term data preservation alternatives

4.7.1 Investigation of EUDAT, Zenodo, and national cloud alternatives

Digitization of Natural science Collections (NHC) evolved from transcription of specimen catalogues in databases to web portals providing access to data, digital images, and 3D models of specimens. These increase global accessibility to specimens and help preserve the physical specimens by reducing their handling. The size of the NHC requires developing high-throughput digitization workflows, as well as research into novel acquisition systems, image standardisation, curation, preservation, and publishing. Nowadays, digitization workflows (and fast digitization stations) can digitize up to 6,000 specimens per day. Operating those digitization stations in parallel, can increase the digitization capacity. The high-resolution images obtained from these specimens, and their volume require substantial bandwidth, and disk space and tapes for storage of original digitized materials, as well as availability of computational processing resources for generating derivatives, information extraction, and publishing. While large institutions have dedicated digitization teams that manage the whole workflow from acquisition to publishing, other institutions cannot dedicate resources to support all digitization activities. This motivated the research into alternatives for long-term storage of digital collections. National and

[79] Small and Medium Enterprises.

European e-infrastructures are viable alternatives supporting different parts of the digitization workflows. To this end, three different e-infrastructures providing long-term storage were analysed through three pilot studies: EUDAT-CINES, Zenodo, and National Infrastructures.

The EUDAT-CINES pilot centred on transferring large digitized herbarium collections from the National Museum of Natural History France (MNHN) to the storage infrastructure provided by the Centre Informatique National de l'Enseignement Supérieur (CINES), a European trusted digital repository, part of the EOSC infrastructures. The upload, processing, and access services are supported by a combination of services provided by the European Collaborative Data Infrastructure (EUDAT CDI) and CINES.

The Zenodo pilot included the upload of herbarium collections from Meise Botanic Garden (MBG) and other European herbaria into the Zenodo repository (hosted by CERN). The upload, processing and access services are supported by Zenodo services, accessed by MBG.

The National Infrastructures pilot facilitated the upload of digital assets derived from specimens of herbarium and entomology collections held at the Finnish Museum of Natural History (LUOMUS) into the Finnish Biodiversity Information Facility (FinBIF). This pilot concentrates on simplifying the integration of digitization facilities to Finnish national e-infrastructures, using services developed by LUOMUS to access the data hosted at the Finnish IT Centre for Science, CSC.

The data models employed in the pilots allow defining data schemas according to the types of collection and specimen images stored. For EUDAT-CINES, data were composed of the specimen data and its business metadata (those the institution making the deposit, in this case MNHN, considers relevant for the data objects being stored), enhanced by archiving metadata, added during the archiving process (institution, licensing, identifiers, project, archiving date, etc). EUDAT uses ePIC identifiers (ePIC 2019) to identify each deposit. The Zenodo pilot was designed to allow defining specimen data and metadata supporting indexing and access to resources. Zenodo uses DataCite Digital Object Identifiers (DOI) and the underlying data types as the main identifiers for the resources, augmented with fields based on standard TDWG vocabularies. FinBIF compiles Finnish biodiversity information to one single service for open access sharing. In FinBIF, HTTP URI based identifiers are used for all data, which link the specimen data with other information, such as images.

These pilots are described in design reports[80] which include features, capacities, functions and costs for each model, in three specific contexts which can be relevant for the implementation of the Distributed Systems of Scientific Collections (DiSSCo) research infrastructure, informing the options for long-term storage and archiving digitized specimen data. The explored options allow preservation of assets and support easy access. In a wider context, the results provide a template for service evaluation in the European Open Science Cloud (EOSC) which can guide similar efforts.

Each of the three services offer cloud service to access data on-line, and a long-term data preservation service where data is in a hierarchical storage accessible through grid technology (IRODS). These long-term services were not tested in the ICEDIG project.

The cloud services were tested and significant differences in data upload and download performance were found. However, the results were not carried out using standard datasets and therefore are not fully comparable. During the tests random errors were encountered in some cases. Despite that use in principle has been shown to be possible, further work is needed to evaluate these (and other) storage options.

[80] ICEDIG deliverables D6.2, doi: 10.5281/zenodo.3364533; D6.3, doi: 10.5281/zenodo.3346782; and D6.4, doi: 10.5281/zenodo.3469490.

Recommendation 73: DiSSCo should develop harmonised policies, procedures and best practices covering the different kinds of storage solution available for a wide range of anticipated data storage needs by collection-holding institutions.

4.7.2 Reproducible research through research objects

Collection-holding institutions also have a requirement to be able to support long-term preservation of data sets used for research; for example, aggregations of selected specimens combined with other data sources, the intermediate and final results of analyses and details of the methods used i.e., 'research objects. The RO-Crate initiative[81] is one such approach that should be studied further for compatibility with DiSSCo Digital Specimen Architecture.

5 Culture, skills and capacity building

Technical advances outstrip the ability of social structures to adapt to changes. Personnel involved in collections to be digitized must be part of the process of change and must become enthusiastic and engaged; especially since collection agents are a large set of end-users of both the eventual DiSSCo research infrastructure and the data produced and stored within.

The present section firstly investigates social aspects (work, skills) of digitizing biodiversity by understanding the current practices, responsibilities and roles of the collections holders and considering the cultural differences between countries and organisations. Secondly, it considers current capacities of collections holders to perform digitization and makes suggestions for capacity building, including needed skills profiles in digitization, data management and analytics, and taxonomy. Thirdly, it explores how opening the natural science collections affects scientific knowledge exchange, collaborations, and interdisciplinary research. It will advise on applicable working methods and approaches to science, given expanding access to vast amounts of digital data.

Analysis of the results of an online survey sent by CETAF as part of ICEDIG project work to targeted natural sciences collection-holding institutions at the end of 2018 forms the basis of this present section 5 on culture, skills and capacity building. The survey results, representing a large pool (143) of participating European collection-holding institutions provide a snapshot of the current level of competencies and the need for capacity building to support collection digitization.

5.1 Current practices, responsibilities and roles

5.1.1 Types of collections being digitized

The collections mostly being digitized today are, unsurprisingly, the botanical collections (herbaria), followed closely by lichens and algae, mycological collections (fungi), entomological collections (insects) and vertebrate collections. Amongst the least digitized natural science collections are anthropological and ethnographic collections, archaeological collections, and documentation archives of various kinds. The type of collection or their storage method strongly influences the digitization effort (e.g., Herbarium specimens are easier to digitize than specimens stored in jars with liquid preservatives).

Documentation (books, field notes, documents and illustrations) is sometimes being digitized as part of specific departmental or collection-oriented projects as well but the libraries of collection-holding institutions and general documentary assets are also being digitized outside of those departments directly involved in natural sciences. The latter are often part of an overall institutional digitization initiative and can end up being treated as a separate digital collection.

[81] https://researchobject.github.io/ro-crate/.

5.1.2 Current digitization efforts

The majority (83%) of European collection-holding institutions are active in digitizing their collections, spanning from large scale (systematic digitization of whole collections or sub-collections, 44%) to small scale digitization (digitization of a limited series of objects/specimens, on-demand digitization, 39%). Only 8% of those interviewed are currently not performing any digitization at all. Additionally, carrying out digitization of collections can mean very different things for each collection-holding institution. This goes from turning physical objects into complete digital objects with added metadata and georeferencing included, down to just creating a database record for a specimen with minimal data and without images.

Digitization is still mainly performed in-house, with about 40% performed solely by internal staff without any external assistance or support. In only 1% of the cases is the digitization effort completely outsourced to an external partner, and in 6% it was partly outsourced. Over half (54%) of digitization efforts also involve the participation of volunteers and other unpaid collaborators or temporarily paid helpers (such as job students), who can contribute in various tasks of the digitization project (e.g., transcribing label data and making images). On average, full time digitization staff members make up 36% of the total workforce, 30% are part time staff, 18% are in-house volunteers (not paid), 9% are temporary paid positions, and 6% are outsourced external collaborators.

5.1.3 Current technical capacity

5.1.3.1 Tendency towards on-site digitization

All the collection-holding institutions represented in the ICEDIG project consortium have significant proportions (>50%) of their collections that still require digitization. A lack of common terminology and the variance in defining the extent of digitization makes institutional comparisons challenging. In terms of established digitization workflows, all institutions can digitize herbarium sheets at small scale. This reflects the simplicity of digitizing sheets that are largely two-dimensional, digitization methodologies well established and that herbaria are a common collection type across institutes.

As noted above, most of the institutions are actively engaged in digitization at large scales, and there is a strong current tendency towards on-site digitization, even when employing the services of an external contractor. This suggests that all institutions have some suitable space on-site for mass digitization, but further conversations would be needed to assess how much those activities could be scaled up in those locations. It was also notable that outsourcing, both currently and historically, centred almost entirely on the digitization of herbarium sheets. Macrophotography setups are ubiquitous but other more expensive digitization equipment such as laser scanners, computed tomography equipment and scanning electron microscopes are less common. There is potential within DiSSCo to negotiate discounts with suppliers for bulk ordering of digitization equipment.

Opportunities should be made available to encourage wider sharing of digitization expertise by encouraging networks and disseminations channels that can be joined by digitization professionals.

Recommendation 74: DiSSCo should encourage and provide opportunities in various forms (newsletters, fora, blogs, networks, conferences, etc.) for sharing expertise and knowledge among its digitization professionals.

5.1.3.2 Specialist digitization teams

Specialist centralised digitization teams are now the norm within several institutions. They are mostly trained in either photography or scanning. These teams are sometimes trained to use other visible light imaging equipment and may have technical skills to use OCR software or development software or scripts to support digitization. As a community we will need more staff trained in these skills, along with the capability to evaluate and utilise new software and approaches to digitization.

Those institutions not digitizing yet will require training and capacity building in almost all aspects of digitization and data mobilization, especially for the skills needed to operate the specialist hardware and software, but also in fundamental ICT support-related skills such as databasing of digital data and management of the digital archives and collections. Providing data to a suitable standard (for example to aggregators such as GBIF) can be as challenging as imaging and requires training and support as well. There is clearly a need for either externally provided training or better knowledge exchange with institutions that are more experienced because of their own digitization activities.

Recommendation 75: DiSSCo must plan and implement a comprehensive training programme covering all aspects of modern digitization and data mobilization.

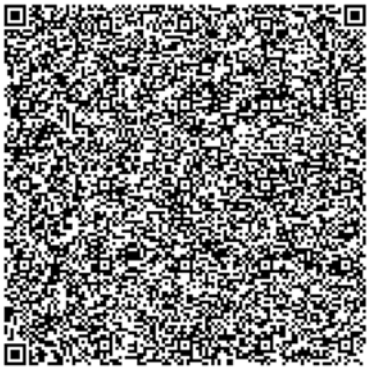
5.1.3.3 Internal documentation and tracking

In terms of internal documentation most of the institutions have specimen handling and digitization process guidelines that collectively cover imaging, georeferencing, data entry and transcription. While some of this documentation may be specific to an institution, there is undoubtedly value in DiSSCo collating this documentation and sharing it as best practice guidelines. This would also help to clarify terminology and definitions of digitization terminology.

Recommendation 76: DiSSCo must offer a platform for sharing documentation and best practice guidelines of workflows from the digitization projects of its Facilities, so that new projects can start faster and learn from each other, with the appropriate citation.

Barcoding specimens has become standard practice across the surveyed institutions for maintaining a traceable link between digital and physical specimens, at least for centralised, large scale digitization projects. Despite the large-scale deployment of barcodes, knowledge of the differences and the pros and cons of each of the various barcode standards[82] is largely lacking among collection managers. One dimensional (1D) barcodes are not ideal for image analysis and impossible for small objects such as insects. Many specimens are now receiving several barcodes, for example from different institutions as a collection is moved from one institution to another. Different practices among institutions have resulted in a variety of different 1D and 2D (two dimensional) codes being used, with no standardization as to what to encode[83], leading to recurring problems across digitization projects. With increasing barcode use, especially of 2D matrix codes[84], standardisation is now essential; both in choice of 2D matrix code itself and on what data to encode. When barcodes contain more information than just a running number (such as institution code), multiple codes can be automatically read and interpreted, either directly from the physical specimen labels themselves and/or from label images to reveal information about the specimen and its history. This is illustrated by the example in Figure 5 on the following page.

[82] https://en.wikipedia.org/wiki/Barcode.
[83] Although at least one proposal has been made, by Diazgranados and Funk (2013), doi: 10.3897/phytokeys.25.5175.
[84] https://en.wikipedia.org/wiki/Barcode#Matrix_(2D)_barcodes.



nsid: 20.5000.1025/486a7e883f14f88bba37
physical specimen BMNH:2006.12.6.40-41

Encoded as JSON (listed below), the QR code on the left provides information about a specimen in the Natural History Museum, London with the catalogue number BMNH:2006.12.6.40-41. It includes a hyperlink (stableIdentifier) to the Web page for the specimen in the museum's data portal. The information encoded is about one quarter of the maximum amount (4296 characters) that can be contained in a QR code.


```
{
  "type": "DigitalSpecimen",
  "attributes": {
    "content": {
      "id": "20.5000.1025/486a7e883f14f88bba37",
      "creationdatetime": "",
      "creator": "",
      "midslevel": 2,
      "scientificName": "Holorchis castex Bray & Justine",
      "country": "New Caledonia",
      "locality": "Rocher a la voile",
      "decimalLat/Long": [-22.3, 166.42],
      "recordedBy": "J L. Justine",
      "collectionDate": "2006-06-01“,
      "catalogNumber": "2006.12.6.40-41",
      "otherCatalogNumbers": "NHMUK:ecatalogue:7072219",
      "institutionCode": "NHMUK",
      "collectionCode": "ZOO, Parasitic worms",
      "stableIdentifier":  "https://data.nhm.ac.uk/object/e90b81bc-1642-47ca-b587-6aa8885cd6a0/1558569600000",
      "physicalSpecimenId": "013258549",
      "Annotations": "Type status = paratype. Holotype =  MNHN JNC 1848 –D1",
      "gbifId": "https://www.gbif.org/occurrence/1826086349",
      "catOfLifeReference":
"http://www.catalogueoflife.org/col/details/species/id/828cbc4eecaa2402b09cd9223754171e",
      "literatureReference": "https://doi.org/10.5281/zenodo.175744",
      "treatmentbank": "http://tb.plazi.org/GgServer/html/03BA87825543FFBFD895FD27FAE1DDAD",
      "enaBiosample": "None available",
      "enaSequence": "https://www.ebi.ac.uk/ena/data/view/FJ788436",
      "LiteratureReferenceRelated": "https://doi.org/10.2478/s11686-009-0045-z"
    }
  },
  “elements": []
}
```


Figure 5: An example matrix (QR) code encoding information about a specimen

Recommendation 77: Stop using one-dimensional barcodes and move over to using a standard two-dimensional matrix code, with standardised data content, that can be read automatically from digital label and other images.

The provenance of the specimens from the field and through the various collections to their present disposition can often be tracked by analysing in sequence the various labels and annotations made on the specimens. The value of such information can be debated, but at least it forms a part of cultural history

and discussion about it could be encouraged. Tracking the location of specimens by electronic methods appears to be somewhat underdeveloped generally, although there are several good systems employed within various institutions to act as references for others. The natural science collections community has yet to routinely add provenance data to digitized specimens. Making such documentation more easily findable, accessible and part of specimen metadata would make specimen tracking much easier.

5.1.4 Cultural differences

More so than cultural factors, the institution size, collection types to be digitized and the availability of digitization staff all play important roles as potential limiters of digitization. The cultural background or the country in which the collection-holding institution is located appears to be neither determinant nor decisive for the decision-making towards digitizing. However, there is one facet where cultural aspects can become a factor to consider. This has only been observed when referring to volunteer work and citizen science support. In northern and western Europe, there is a long history in volunteering, and this has become a sign of progress and involvement among the civil society[85].

Speaking generally, the more mature the collection-holding institution is and the more economically and socially developed the location country is, the larger the external commitment towards digitization appears to be. Large institutions in developed countries are preceded by their fame and reputation. They can count on a bigger and more stable external workforce that become attracted, among others, by the size of the collections, the easy access for collaborative work, and the possibility to gain credit for work done. This is also promoted by the collection-holding institution's themselves through their communication and dissemination campaigns that can be stronger and more impactful locally. This increases motivation and engagement of volunteers.

Recommendation 78: DiSSCo should assist its collection-holding institution members to develop and strengthen their external profile (marketing) with funding agencies, professional and citizen scientist groups and local communities appropriate to their location and sphere of collections related operations (i.e., research, education and exhibition).

5.1.5 The limitations of current capacities to perform digitization

One third (36%) of digitization staff receive on-the-job training. Within this already small group, half of them only receive training initially at the beginning of their digitization activities without follow up or additional training to stay up to date. Only a minority (10%) receive training multiple times per year.

When training is provided, it covers skills in basic digitization work, such as handling of collections, use of digitization hardware (cameras, computers…), utilisation of digitization software and the use of a collection management system. About half of the trainees also receive training in more advanced aspects such as data management and archiving. Less than 20% of respondents receive training that deals with related critical activities such as ICT support, working with volunteers, and new developments in digitization. Finally, only 5% learn about implementation of automation and robotics. Specific aspects of training currently missing in many cases are knowledge and skills in technological aspects, working with shared or common data references and identifiers, and skills in manipulating and working with physical collection objects.

These results suggest that many staff employed to perform digitization tasks are not adequately trained to do so. This observation is supported by many respondents' answers provided verbally, reflecting their own self-assessments. Lack of training to effectively operate the hardware and/or software used to digitize collections, basic general digitization skills and associated elementary collection actions such as handling specimens or a rudimentary knowledge of taxonomy and interpretation of taxonomic data are

[85] Contrast this with the USA, for example where unionisation prevents the use of volunteers by some institutions.

all missing from a large proportion of staff involved in digitization. There is an overall and urgent necessity for both general and specialised training for staff covering all aspects of digitization at all complexity levels and an urgent need for professionalization of such work.

A different approach was taken by Digitarium[86] in Finland in 2011-2013, when specific courses on digitization of natural science collections were designed and implemented. Two courses, each lasting about 8 months and each enrolling 10 students were held. The trainees were academic unemployed people and the activity was financed through the European Social Fund. At the present time (January 2020) three of the trainees are now employed in digitization related work in fixed positions and two are temporary staff in ongoing digitization projects.

Recommendation 79: DiSSCo should organise a training curriculum for its member institutions covering: i) technological aspects, such as features and operation of equipment and software; ii) standards, i.e., museum and archival practices including data standards, in particular unique and persistent identifiers; iii) efficient digitization workflows in various situations, including quality management; and iv) for museum leadership.

5.1.6 Limitations in resources and funding

Often mentioned, insufficient funding is the biggest limitation for digitization, principally restricting the ability to hire, train and keep experienced staff. Funding when available is often mostly project based and is rarely enough to hire staff that are either highly qualified (and who could then act as trainers to others) or that are fully dedicated to digitization for prolonged periods of time. Additionally, there are also insufficient dedicated funds for those costs of digitization that are not related to staff, such as the acquisition of the necessary technical equipment and the use of the most up-to-date and sophisticated software (which is often only available commercially). About 30% of the survey respondents mentioned these latter shortcomings.

Currently therefore, digitization activities are most often relying on external funding, which suggests that digitization is a priority for allocating core budget in collection-holding institutions only in a limited number of cases.

However, despite these apparent limitations there are opportunities, especially using European Structural Funds[87] as exampled by the Finnish Digitarium initiative from 2010 - 2017.

A national centre of digitization expertise in Finland, Digitarium was launched in 2010 and operated until 2017, funded by a series of grants from the European Structural Funds. Totalling €2.1 million, this funding covered 70-80% of costs, with the remainder coming from the host city and the two participating universities. The funding was used to build the technological base for mass digitization as well as the human capacities. Additional funding of €2.0 million was obtained from EU FP7 research projects, national research infrastructure projects, and commercial mass digitization services. This model of funding in eligible parts of Europe can be attractive for DiSSCo as a means of establishing the needed digitization factories and Centres of Excellence. It is worth noting that the ESFRI LifeWatch Research Infrastructure is being largely built on a similar basis.

Recommendation 80: DiSSCo institutions should look for opportunities to use EU structural and investment funds to build up the digitization capacities in eligible countries and regions. DiSSCo should centrally support this activity with application packages and support for proposal writing

[86] http://www.digitarium.fi/en/content/our-expertise-and-knowledge-base.html.
[87] http://www.digitarium.fi/en/content/digitization-centre.html.

targeted specifically for these funding sources which are not research oriented but aim for economic and social development.

5.1.7 Digitization becoming business as usual

Mass digitization and Digitization on Demand have been described (3.9, 3.10) as processes (activities) both taking place today and foreseen to take place at larger scale in and across the collection-holding institutions that make up DiSSCo. Beyond one-off funded projects (albeit lengthy ones in some institutions) when does digitization become 'business as usual' and what is needed to make that happen? What is meant by digitization business as usual? What is 'digital by default'? Indeed, how will digitization adapt and look like in the future as new innovations are introduced, as funding models changes, and as organisations collaborate more and co-organise in pursuit of digitizing and digitally exploiting Europe's 1.5billion specimens?

Many museums are already organising or beginning to organise their digitization beyond specific projects by establishing specialised functions and teams. LUOMUS (Finland), for instance has two teams (botany, entomology) for mass digitization. Naturalis (The Netherlands) has coined the term "permanent digitization infrastructure" as everything required within an institution dealing with digitization on a day to day basis e.g., the digitization of specimens sent out on loan, a type specimen that needs imaging, volunteers that digitize a private collection, etc. That doesn't include mass digitization and digitization-on-demand. In other collection-holding institutions, hi-resolution digitization (e.g., of specimen parts) and selective digitisation (e.g., of type specimens) becomes more focussed in specialised units with appropriate expertise and equipment.

Each collection-holding institution has a different balance between what it considers to be 'business as usual' and what is done by central teams, functions or projects as a separate workflow. Working collectively towards the DiSSCo aim of digitally unifying all European natural science assets under common curation and access policies and practices that make the data easily FAIR is bound to bring further and sweeping organisational changes and these should be studied to identify best practices.

Recommendation 81: DiSSCo should investigate and promote best practices for operating models within collection-holding institutions whereby digitization becomes business as usual and digital by default.

5.2 Effect of opening collections on knowledge exchange, collaboration and research

To determine the effects of opening natural heritage collections to wider access by means of digitization, we here consider the effect on collaborations and research (5.2.1), on the mobility of collections in terms of loans and visits (5.2.2), and on education, citizen science and private collections (5.2.3).

5.2.1 Effect on collaboration and research

When considering the real need for digitization of natural science collections, one important argument is that the collections would be opened for a far broader usage, not only for the community of natural sciences researchers but for society at large.

In cases where digitization has already had a noticeable effect, that is considered overwhelmingly to be positive or very positive. In general, digitization broadens the impact and strongly facilitates many aspects of collection-based work, both internally and externally, with positive effects attributed to one of the following categories:

- **New collaborations and networks:** Digitization activities have a positive effect on connecting museum departments within an institution and research groups acting across institutional

boundaries, as well as external partners that previously were less engaged. Digitization helps departments within the same collection-holding institution to work more closely together and to develop more streamlined workflows. The same effect is visible with external partners from different collection-holding institutions, since digitization directly stimulates knowledge creation and exchange among scientists with similar focus.

- **Improved access to collections and collection data:** Digitization makes collections more visible to external researchers, allowing easier access to collection data and thus stimulating their use. Researchers and taxonomists who work with collections can save much time and effort - at least for initial assessments - by consulting the digital data, rather than having to physically visit the collections or request a loan and extract the information manually.
- **Sharing of knowledge, standards, working methods and practices:** Digitization results in more efficient networking, a faster and more dynamic exchange of experiences between collection-holding institutions and within departments, and stimulates the creation of research groups and of larger consortium projects (such as ICEDIG) that aim at improving the existing practices and setting the standards for larger and more harmonised digitization efforts in the future.
- **Effects of digitization activities on obtaining funding opportunities for new projects:** As digitization initiatives expand across institutions and countries and as the awareness of the positive effects of having digitized collections increases, access to funding opportunities becomes easier and broader. Public institutions realise more and more the need to improve access, physical and digital, to their collections in their budget proposals. National initiatives provide increasing support with the inclusion of digitization in the national research Roadmaps. The European Commission has shown their support to this endeavour when it approved the DiSSCo RI for inclusion on the ESFRI Roadmap Update in September 2018 and thus characterising digitization as pivotal towards the achievement of the goals of European open science.
- **Increase in visibility of collections and stimulation of scientific/taxonomic investigations and output from these collections:** Many collection-holding institutions that digitize their collections are noticing an increase in research outputs (scientific publications) produced with the inclusion of (digitized) specimens from bio- and geo-diversity collections. Researchers and taxonomist can more easily detect the presence of interesting specimens in collections and can include them in their projects in a more efficient and relevant manner.

Recommendation 82: DiSSCo should deploy metrics (key performance indicators) to monitor impact and progress in collaboration and research facilitated by digitization and should publish the results annually.

One possible side effect of this success story that should not be ignored however is the changed workload (perhaps increased) on the shoulders of the personnel conducting digitization and those responsible for arranging and shepherding access to physical collections, which may increase as a consequence of increased external visibility and accessibility.

5.2.2 Effect on mobility of collections

Opening access to the collections also influences the mobility of collections in terms of the numbers of arranged loans and visits. Half of the interviewed respondents noticed there is a difference in terms of the number of loans and visits when comparing the current situation of (totally or even partly) digitized collections to times before digitization began. The effects are the following (expressed in rounded-up numbers, with the other 50% noticing no effects):

- Decrease of both loans and visitors: 7%,
- Fewer loans but same number of visitors: 9%,
- Fewer loans but more visitors: 5%,
- No change in number of loans but fewer visitors: 3%,

- No change in number of loans but more visitors: 5%,
- More loans but fewer visitors: 8%,
- Increase in the number of both loans and visitors: 10%.

The very broad distribution of responses in terms of effect correlates with the completeness of the digitization as well as to the type of data provided (e.g., with or without images) in each institution. Collections that are fully digitized may experience fewer loans and visits, since researchers can check in advance if the collection is worth visiting and what individual specimens of interest are stored in the collections, and can often use the digital data e.g., for identification. The same is true for collections with an image even if they are not fully digitized. Still, the need to send material on loan seems to decrease. For collections that are available online, but without an image, both loans and visits seem to increase, indicating the need to verify the reliability of the available information, e.g., to check if the provided identification is correct, and the impact of saying what is in collections without providing further details (I.e. awareness increases but the digital data are not research-ready by themselves).

Additionally, when the entire collection is digitized, the requests are also more specific. When only the type specimens are digitized, there are more requests for specimens (more specimens per loan), but this does not impact the overall amount of loans or visits. Finally, digitization-on-demand is regarded as a good alternative to loans of physical specimens when no further examination of the specimen is required (e.g., dissections, taking detailed measurements, etc.).

Digitization, perhaps particularly when associated with other changes such as collections moves, can be an opportunity to revisit institutional processes more broadly. Given the uncertain impact of digitization on physical access and loans, DiSSCo may wish to pilot 'digital by default' access approaches which focus on digitization-on-demand first, followed by physical access or loans only where the digital data cannot meet the research need. DiSSCo may also wish to put in place metrics that show how digitization is spreading across all collections in a museum, as a new standard of practice.

Recommendation 83: DiSSCo should deploy metrics (key performance indicators) that show how digitization is spreading across collections and to monitor the changes in mobility and usage of collections. The impact of digitization should be assessed on a regular basis.

5.2.3 Effect on education, citizen science and private collections

Collections are an invaluable resource for education, both for formal and informal education. Formal education is usually directly supporting school curricula and taking place on the school premises while informal education can take place virtually everywhere and is not bound to school curriculum.

As illustrated in Figure 6, the usefulness of collections on education can be seen in three aspects – physical specimens, occurrence data and images of specimens.

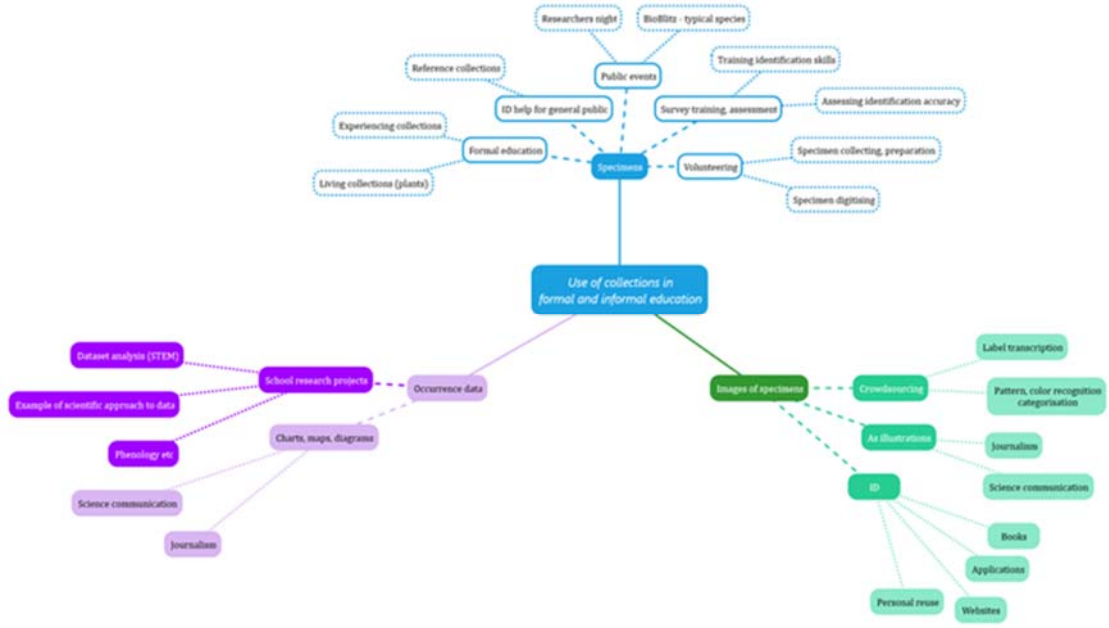


Figure 6: Some uses of natural science collections in formal and informal education

Digital content of natural science collections (images, maps, etc.) is already actively used in informal education[88]. Moreover, the perceived potential of the educational use of digital information is higher, compared to public exhibitions and access to physical collections. Publishing digital content with educationally accessible licenses, easy-to-use web interfaces and providing open access via API would further facilitate the use of natural science collections in education.

Recommendation 84: As well as providing access to Digital Specimens for research purposes, DiSSCo should consider the additional and different aspects that can pertain to providing access for educational purposes.

Although private natural science collections play an important role in building material evidence for taxonomy worldwide, they go largely unnoticed as a form of citizen science, whereas observation-based initiatives such as iNaturalist gain high visibility and impact. Representation of collection-based projects in citizen science portals is rather modest[89].

Recommendation 85: DiSSCo should actively promote the role of private natural science collections as a form of citizen science.

These portals, of which there are many, often utilise metadata of citizen science projects for their search engine backbone. However, these search engines are missing appropriate data fields that would allow discovery of projects based on private collections. This is most likely because the development of relevant metadata standards by, for example the European Citizen Science Association (ECSA)[90] and/or the Citizen

[88] ICEDIG project deliverable D5.3, doi: 10.5281/zenodo.3364541.
[89] ICEDIG project deliverable D5.2, doi: 10.5281/zenodo.3364519.
[90] https://ecsa.citizen-science.net/.

Science Association[91] are not being well supported by collection-holding institutions through active membership in relevant working groups. The situation needs more attention. DiSSCo collection-holding institutions, working on behalf of private collections must review all the relevant data and metadata standards and work to ensure that missing data fields and/or vocabulary terms that help to link private collection information is are added. Specifically, the development of the PPSR-Core metadata standard under the auspices of the Citizen Science Association[92] should be kept on the radar.

Recommendation 86: In working to promote private natural science collections as a form of citizen science, DiSSCo should take the lead to ensure that the metadata definitions needed to make private collections more publicly visible becoming incorporated into appropriate citizen science metadata standards, such as PPSR-Core.

5.3 Improving working methods and approaches

5.3.1 Re-organising work

In current institutional practice, staff are normally organised around fixed roles, each having specific tasks attributed to them. Examples of such roles are 'collection curator', 'collection technician', 'digitization agent', etc. However, to properly undertake digitization activities according to this template means cooperation is needed between actors holding different roles. As the means to introduce a greater flexibility in the management of skills and competencies within the digitization process, we suggest the idea of working with so called functional units rather than with traditional roles. A functional unit (in the sense of organising competencies) is a collection of competencies needed to perform specific tasks within the different steps of the digitization process. We believe this approach can allow a more fluid distribution of tasks among digitization personnel. Staff often already think in that direction, shown by the fact that 68% chose to identify themselves with more than one role during the survey. A tentative structure for digitization related functional units has been suggested (Appendix 1 in ICEDIG MS49 report) including a proposed list of tasks that can be attributed to each potential functional unit. In a small institution one person could be qualified for, and combine, tasks from one or more functional units while in a large institution several people could be capable of performing several tasks within one functional unit. How functional units are distributed amongst the staff can depend on different factors, such as the capacities of the institution (i.e., number of available staff), or the desired scale of digitization to be undertaken. In some institutions, especially the smaller ones with a limited number of staff, some people are already performing tasks that transcend their habitual role and combine tasks that would otherwise belong to people who hold different roles. This phenomenon was clearly supported by some of the survey respondents who consider this to be common practice within their own organisation. It is, for example, possible that the collection curator also performs tasks that can be attributed to a collection technician and/or a digitization operator.

Once the DiSSCo consortium is fully implemented, collections-related work could evolve from a research institution-based approach to a distributed research infrastructure-based approach. This means that while currently most digitization activities are taking place in the context of single institutions, with DiSSCo this could evolve and increasingly include more coordination between institutions and countries. Better coordination of digitization strategies and priorities would efficiently create the best possible digital collection of bio and geo-diversity specimens.

This would involve *new and improved unified methods* to tackle the challenges of collection management and digitization of collection objects and their associated data. It could free organisations from following

[91] https://www.citizenscience.org/.

[92] PPSR-Core is a set of global, transdisciplinary data and metadata standards describing contextualized details about Public Participation in Scientific Research (PPSR) projects i.e., citizen science projects. For further details, see here: https://github.com/CitSciAssoc/DMWG-PPSR-Core/wiki/About-the-PPSR-Core.

the strict definition of roles and could act as a facilitator for the mobility of work forces and researchers as well as the creation of international competence groups. At the very least, it would provide institutions a common vocabulary when discussing collaborative actions, common projects or very specific things like comparing annotation workflows and when hiring / or designing professional development.

This is a concept that falls in the realm of *Business Process Re-Engineering* which has been practiced in the commercial sector for about two decades. Introducing a new European RI should certainly introduce some new and harmonised working and business practices across the DiSSCo membership that will need further discussion and could only be developed so far in the context of ICEDIG. How the work in each department will be organised will of course always remain at the discretion of each collection-holding institution.

Recommendation 87: DiSSCo to develop a strategy for aligning and unifying the work practices across its facilities.

5.3.2 Re-organising data management

5.3.2.1 Interoperability of Collection Management Systems with DiSSCo Hub

Collection Management Systems (CMS) are database systems used by collection holders to keep track of the data they possess about their specimens. As the primary data sources for the DiSSCo Hub, they will play a critical role in the infrastructure. Specimen data will have to flow efficiently and in a standardized way from those systems to the hub, where the data can be consulted alongside data from other sources. This way, interplay with other data is also possible for enrichment and validation purposes. After such enrichment and/or validation services have been applied, data should also flow back seamlessly into the CMS to maintain the CMS as the central authority.

Surveys indicate many different CMS solutions are presently in use. Some solutions are bespoke, in-house developments, while others are commercial products or the result of national consolidation initiatives. Small collections may not use a dedicated system at all. Migration from one system to another is a time-consuming and difficult process, made worse by an imprecise picture of which systems are available/in use and how they compare to each other. A landscape analysis of potential systems is needed, with a list of minimal specifications – at least those related to the needs of the DiSSCo Hub. Local needs and preferences may be harder to generalize. This will facilitate migration from outdated systems, but also the adoption of a first CMS by smaller collections. It will also discourage the development of further bespoke systems, which is to be avoided.

Recommendation 88: DiSSCo should prepare a minimum specification of an ideal collection management system (CMS) and select/recommend preferred alternatives from the available product solutions to meet member institutions' various needs.

Some recommendations for CMS developers were drafted in ICEDIG project deliverables D4.3[93] and D4.4[94]. Most of these relate to the interactions with the DiSSCo Hub. The ease of data flowing in and out of systems varies considerably today. Data publishing pipelines to institutional or national portals and international aggregators are becoming more common, but issues with a lack of standardization persist. This hinders interoperability. Implementations of links with external services, such as persistent identifier services for people or taxonomic names, would address some of these problems. The consistent and distinct use of persistent identifiers for both physical and digital specimens is related to this.

A further obstacle to efficient data flow from the DiSSCo Hub back into local CMS is the problem of trusting these new or updated data, in conjunction with the lack of a proper version history mechanism in most

[93] ICEDIG project deliverable D4.3, doi: 10.1093/database/baz129.
[94] ICEDIG project deliverable D4.4, doi: 10.5281/zenodo.3361598.

CMS. Most CMS are also unable to systematically store annotations relating to specific records or specific data fields of records. The overhead of dealing with multiple versions or annotations to records remains a separate problem to solve, but data should not get stuck in between different nodes of the pipeline.

Some CMS aim to cover a wide range of data types, while others are tailored for very specific needs. The institutional organization also has an impact, as libraries are often managed separately from biological or geological collections. Because of this, they use different systems and even different standards, despite the potential links that exist between books and specimens, or the overlap in certain types of data such as those related to people. For technical reasons, multimedia objects of specimens are also managed in a different manner than specimen data. Multimedia objects suffer from similar problems as data, such as versioning, proper use of identifiers and interoperability with media from other sources. A push for better standardization might be needed here as well, using for example "triple-eye eff"[95] image interoperability framework.

Recommendation 89: DiSSCo should adopt the "triple-eye eff" (iiif.io) image interoperability framework as its basis for media management and interoperability.

Finally, not all data are currently published from CMS. This is often related to (perceived poor) data quality or availability, as well as sensitivity of data relating to threatened species or valuable specimens. Some data are not published as they fit poorly in the existing exchange standards – or not at all. This is frequently the case for verbatim renditions of data, which are typically considered obsolete after transcription and interpretation, but which nevertheless still have various viable use cases (Table 10). A prominent use case (listed as case 5 in Table 10) is providing training and validation data for automated text capture methods. This requires the unclassified verbatim rendition of a specimen's label information as this is also the format in which the data will be produced by those automated methods. Data produced this way can prove valuable for findability of the specimen in the absence of a fully atomized and standardized transcription. In turn it can be important for refining data capture algorithms.

[95] International Image Interoperability Framework – "triple-eye eff", https://iiif.io/.

Table 10: Use cases for verbatim data, with examples and notes on applications

	Use Case	Examples	Application notes
1	Facilitating data cleaning and indicating the degree of interpretation in the standardized fields.	Dates that are found to be unlikely or impossible can be easily checked for typos or erroneous transcription.	If a digital image of the label is available, then there is less need to check a verbatim transcription for validation.
2	Discovering information hidden in the typography of how text is presented on the label.	The syntax of person names can be a clue to the writer's identity and for linking related specimens.	This is unnecessary for most specimens but is valuable for enriching poorly documented specimens.
3	Increasing the findability of specimens.	Where a word, such as a place name, can be read but not understood, then the text can still be found.	Original text can be searched in the original language.
4	Accommodating partial or uncertain transcriptions, which would otherwise clutter standardized, interpreted fields.	The use of square brackets ([]) and ellipses to indicate uncertainty or a failure to read part of the text.	Other transcribers can build on the initial attempt, and it will be clear that the information is present on the label.
5	Providing training and validation source data for automated text capture methods.	Automated reading of 19th century handwriting and recognition of symbols used on labels.	Finding gold standard training data for algorithms is a common problem.
6	Accommodating data that are not sufficiently standardized for the interpreted field or that fail to comply with the restrictions of the interpreted field.	Dates that lack a year or data awaiting interpretation.	It is common to find verbatim fields containing data in non-standard formats, yet they are not transcribed data either.
7	Accommodating data following obsolete or bespoke standards.	Grid system location codes.	When a database is migrated from one system to another, then verbatim fields are used to store old formats.
8	Preserving the original language when interpretation has included translation.	Habitats can have some very specific meanings in different languages, and they are difficult to translate because there may not be a direct equivalent.	This also improves the findability of specimens written in a different language.

5.3.2.2 Support for automated data capture in Darwin Core data standard

The Darwin Core (DwC) data standard[96] does not presently fully support automated data capture because there is not a complete separation of verbatim fields from interpreted fields in DwC records. Verbatim fields are needed to capture literally the text from labels, as written, with all possible errors and free syntax. There are a lot of semantics tied up in the way information is structured on the label itself. When we atomize it, we lose much of that. Interpreted fields should use any current lookup tables for obtaining values. The verbatim fields could automatically be captured from labels by OCR and other image analysis, but the interpreted fields values (i.e., interpretations) must presently be manually entered by humans. There is presently no software that can do this. One necessary step towards solving this issue could be to add just one new field to the Darwin Core standard – "dwc:verbatimLabel". This field would contain all the text that can be extracted (by any means) from the label(s) as written. It's content could then be

[96] https://dwc.tdwg.org/.

analysed and structured further by either human transcribers and/or machine-learning techniques, or both in combination.

5.3.2.3 Dealing with blank fields in Darwin Core data standard

Currently, little regard is given to unknown and incomplete data in Darwin Core biodiversity records. It's not possible to tell whether a blank field is blank because the data is not available, is not known, has not been digitized, has been withheld or is simply missing because someone didn't bother to fill the field with a value. Digitized specimen label data are the result of a complex and multifaceted digitization workflow. This workflow is often focused on core information of the specimen label and much data from specimens remains to be digitized. Curators and those working directly with specific collections/specimens often know when data exist but are undigitized. For the wider user community this can be important information, though it is rarely communicated. In deliverable D4.3[93] we recommend a simple vocabulary for unknown data that can be used in conjunction with any other vocabulary used for fields of specimen data. This vocabulary proposes the terms unknown, unknown:undigitized, unknown:missing, unknown:indecipherable and known:withheld. Use of these simple terms would quickly communicate to users about the status of a missing datum and whether it was knowable.

5.3.2.4 Data about people

Data about people are one of the most overlooked components of specimen data. Such data are generally represented only as character strings with no specific format. Nevertheless, in recent years unique identifiers for living researchers and authors, such as ORCID identifiers[97] have become increasingly available that allow us to uniquely identify people and link their research contributions across collections. By identifying both specimens and people, we can link specimen data to biographies, as well as to other kinds of information, such as that from literature and from genetic sequence data. Current data standards still lack the ability to work with people data accurately, including coping with teams of people. However, work is in hand to change this situation and further support to the organizations and researchers proposing these changes is needed.

Recommendation 90: DiSSCo should encourage the use of unique persistent identifiers for people collecting and working in collections.

5.3.2.5 Geography

The geographic location where a specimen was collected is a very important piece of information. It allows the specimen to be associated with other data resources tied to location, such as data related to climate, vegetation, soil and other environmental variables. It is also key for species distribution modelling. Location data can be modelled in different ways and at different hierarchies, but specific point coordinates are commonplace nowadays. Coordinates are always tied to a coordinate reference system (for example, latitude/longitude in the WGS84 system[98]). However, different coordinate reference systems have different characteristics, levels of accuracy and forms of uncertainty. Some locality descriptions can be very broad, but certain parts might be excluded on the basis of land cover or incompatibility with the biological requirements of the species.

Georeferencing of a specimen can also be performed *post hoc* for historical specimens, based on interpretations of locality descriptions or information noted down in field notebooks. Biological observations in the 20th century were regularly made with the aid of national grid systems. These can be converted to coordinate point data with an uncertainty radius, but this distorts the initial extent of the

[97] https://orcid.org/.
[98] https://gisgeography.com/wgs84-world-geodetic-system/

uncertainty. It also causes extra work and further inaccuracy can be introduced because (as in ecological research, for example) data are often transformed into grids again for modelling purposes.

Names of geographic locations suffer from similar problems as names of people: they can change, they can be different in other languages, they may need to be disambiguated, etc. Hence, here too the use of persistent identifiers linked to formal geographic name descriptions (such as those to be found in GeoNames[99]) should be more strongly encouraged. These can be used to validate eventually associated coordinate pairs and improve findability of the specimens.

Recommendation 91: DiSSCo should encourage the use of persistently identified geographic name descriptions from recognised sources.

5.3.2.6 Data migration

Data migration is the process of selecting, preparing, extracting, and transforming data and permanently transferring it from one computer storage system to another. Additionally, the validation of migrated data for completeness and the decommissioning of legacy data storage are considered part of the entire data migration process.[100]

As noted above, migration of data from one CMS to another is a difficult and time-consuming process. Yet it is one that must be undertaken as part of the overall rationalisation of CMS solutions in use across the DiSSCo collection-holding institutions if levels of support and training to institutions in this area are to be improved as part of the DiSSCo programme.

Another kind of data migration occurs when collections of specimens are moved from one institution to another. This is best illustrated by an example of a collection of herbarium specimens that are presently being digitized. Most of these specimens carry a 1D barcode attached by the herbarium originally the collection. This identifies only a catalog number. No institutionCode nor collectionCode appears. Many of the specimens in this collection also carry the barcode and catalog number of a second institution, reflecting a transfer of ownership sometime in the past. The present digitization process includes attachment of a new QR code label to each specimen that contains a universally unique identifier (uuid) with a purl.org prefix. Thus, already 3 identifiers, with potentially similar records in 3 different database systems. This is a problem that needs attention urgently in terms of standardising digitization procedures for the future such that future difficulties are reduced.

5.3.2.7 Identifying data with CETAF Stable Identifiers and Natural Science Identifiers (NSId)

CETAF Stable Identifiers[101] provide the means to consistently reference the digital records about objects in biological collections at the institutional level. Being both human and machine-readable, they redirect users and computer systems to the images, websites, and metadata of the specimen of interest. Wider adoption of such identifiers can have the effect of making digital information about those specimens more easily accessible.

Recommendation 92: DiSSCo should encourage the further adoption of CETAF Stable Identifiers for the local and persistent identification of physical specimens.

[99] https://www.geonames.org/.
[100] Source: https://en.wikipedia.org/wiki/Data_migration.
[101] https://cetaf.org/cetaf-stable-identifiers.

The proposed Natural Science Identifier (NSId) scheme[102] is intended to allow and encourage the emergence of new sector-wide services that build upon the increasing availability and quantities of digital data about physical specimens held in and across natural sciences collections and other sources.

Like in the journal publishing and film/TV entertainment industries, this identifier scheme is independent of the underlying assets to be identified and their owning organisations, as well as being independent of the specific technology of the World Wide Web.

The new NSId scheme to identify Digital Specimens[103] sits alongside the CETAF Stable Identifier scheme as a new level of indirection that should be viewed as an opportunity for adding value in ways that cannot always be foreseen today.

Recommendation 93: DiSSCo should identify benefits from and opportunities for third-party, value-added services arising through adoption of a Handle-based persistent identifier scheme for Digital Specimens, presently proposed as Natural Science Identifiers (NSId).

DiSSCo's requirements for persistently identifying Digital Specimen and other object types are set out in Appendix B.

5.3.3 Keeping records of digitization costs

5.3.3.1 Costing as a new practice

Keeping records of costs is a new practice that should be adopted. It should be possible to account for:

- Initial investment costs of establishing digitization facilities and other infrastructure (e.g., ICT, helpdesk, training, support, etc.);
- Fixed costs of owning and operating such infrastructure; and,
- Variable costs of operations due to demand and prioritization[104].

Optimal digitization cost is achieved when the volume and availability of specimens ready for digitization matches the capacity of the digitization facility. Having enough specimens ready means the digitization capacity can be effectively utilised and the highest throughput can be achieved, thus leading to the lowest cost (notwithstanding other factors contributing to cost and the assumption that the digitization facility is dimensioned sufficiently for the task). Too few specimens ready means the capacity is underutilised, meaning higher cost per specimen.

What an institution wants to know is: When can certain kinds of digitization be achieved for specific levels of investment? When does it become practical/economic to start digitizing a collection? What does it cost to invest for digitization and to reach a certain level for a collection e.g., one of the MIDS levels or dashboard goals? How much is it going to cost to maintain? Gathering cost information begins to inform answers to such questions.

Work in the ICEDIG project[105] has shown the variation in approaches to recording and presenting costs of digitization and the need for guidance by DiSSCo on how to record cost details for future comparisons and for budgeting purposes. Reliance on old-style spreadsheet products, distributed and managed as files among participants is no longer the most flexible, efficient or sustainable approach. DiSSCo should

[102] Like DOIs for journal articles, this will be based on the Handle system, https://www.dona.net/handle-system.
[103] Note that at the time of writing (March 2020) the proposal is that NSId should be used to identify Digital Specimens whereas DOI are being considered as the basis for identifying Digital Collections.
[104] Of course, as well as initial costs, the costs of replacement, upgrade and decommissioning must also be accounted for.
[105] ICEDIG project deliverable D8.2, doi: 10.5281/zenodo.3724224.

consider modern alternatives to the Excel/Google spreadsheets approach for gathering, collating, and analysing cost information and for budgeting and management of DiSSCo costs.

Recommendation 94: DiSSCo must evaluate, adopt and support modern alternative(s) to traditional spreadsheet approaches for gathering, collating, and analysing cost information and for budgeting and management of DiSSCo costs.

Different currencies (sterling and euro) have been used during the cost gathering and analysis work in the ICEDIG project. To make comparisons between gathered costs reasonable, an artificial currency ('Purchasing Power Standard', PPS[106]) has been used. To such make analyses and reporting easy in the future, a helpful approach would be to convert from the local currency used for data entry to PPS for each data item entered, at the time of entry. Nevertheless, budgeting and management of actual costs are likely to be done in the local currency and in the currency of the legal entity to be established by DiSSCo (see 3.8.1).

Recommendation 95: In cost gathering, analysis and reporting, DiSSCo should convert, at the time of data entry from the currency of data entry to a standard currency for analysis and comparison purposes.

5.3.3.2 Costs versus charges

Costs must be treated separately from charges. A cost model is not the same as a charging or business model and cost calculations cannot be considered in isolation from a business/charging/organisational model, because of the influence of DiSSCo governance decisions and policy on requirements for digitization, data access and availability. Digitization can be required to a certain level. Some data may be more immediately available than other data, according to scientific demand and difficulty to retrieve (faster and easier versus slower and more time-consuming).

Analysis of potential business (and thus charging) models is tied closely to questions of DiSSCo organisation and governance (see 6). Nevertheless, it is likely that the DiSSCo business model should use the fundamental assumption that data must be 'free at the point of use' i.e., at no charge to the ultimate end-user. Within such a constraint, various charging models are conceivable, including for example: i) a research subscription model, whereby an institution or project wanting to provide its research staff with access to digital collections data pays a subscription for DiSSCo user membership; ii) an open-access model, whereby those demanding digitization of specimens pay for that e.g., through a funded digitization programme; iii) an extension/re-orientation of the current loans and visits model, whereby costs of organising loans/visits are re-allocated providing FAIR access to digital content; and iv) obtaining money from industry in exchange for "free" data.

Any business model must, however, take both depreciation of capital equipment and amortization of intangible assets i.e., data into account, such that these costs can be accounted for and recovered over the long-term.

Recommendation 96: The DiSSCo business model must take depreciation of capital equipment (tangible assets) and amortization of intangible assets (i.e., DiSSCo data) into account, such that these costs can be accounted for and recovered over the long-term.

[106] Purchasing Power Standard (PPS) is the technical term used by Eurostat for the common currency in which national accounts aggregates are expressed when adjusted for price level differences using Purchasing Power Parities (PPP).

5.4 Capacity building, maturity assessment, skills profiles

Most personnel concerned with digitization in collection-holding institutions have a positive attitude towards the future and for the continuation of their current digitization efforts, with the hope to implement improvements and to upscale them. Nevertheless, current digitization efforts are mainly project based and depend almost entirely on volatile and temporary funding sources. In some cases, especially in the smaller institutions, the current situation is dire and digitization efforts are carried out by too few people who are involved in too many other tasks and roles within their institution. For them, without more dedicated effort, digitization of entire collections is either impossible or progress will soon come to a halt until a new project or opportunity is found. For those working in an collection-holding institution currently undertaking digitization with enough supporting resources (facilities, equipment, personnel and, more importantly, with a strategic vision) the effort needs to continue to complete the digitization of their collections, or even going further by tackling additional challenges such as the addition of collections outside the collection-holding institution (private collections, etc.).

Several crucial aspects must be taken into consideration when addressing the needs for capacity enhancement and for providing proper skills to the people involved in digitization. Still, it remains difficult to draw uniform conclusions a very diverse pool of institutions is involved, some of which have the means to perform large scale digitization and even outsource parts of the digitization activities, while others are very small collection-holding institutions with sometimes only a handful of people in service (occasionally just one person who combines all digitization tasks), trying to build up digitization actions with the little means they have.

Four priorities for capacity building have been identified:

- External training for digitization staff is necessary;
- More funding is required to hire staff and finance the digitization effort in general;
- An increase in dedicated digitization staff is needed; and,
- Investments in modern efficient equipment, software, CMS and data portal solutions are desired.

For these, a capacity building pipeline is needed, and DiSSCo and institutional leadership must plan for that.

Recommendation 97: DiSSCo and institutional leadership must plan for capacity enhancement in i) training of digitization and allied personnel; ii) funding to hire digitization personnel on long-term and for the effort in general; iii) increase in dedicated digitization staff; and iv) investments in modern efficient equipment, software, CMS and data solutions.

Alongside these, there is a general need for standardisation and the systematised sharing of best practice and common approaches. These issues are relevant to both institutional, and individual staff, capability. To support institutions to understand and to improve their capability, DiSSCo should provide a self-assessment tool that identifies levels of 'digitization readiness' (including data mobilisation) i.e., a digitization maturity assessment model. This would enable organisations to understand their own needs and what next steps they might take, and to support their case for these steps e.g., to funders. For individuals, DiSSCo will need to continue work to identify digital skills and competences which can be matched to roles and functions as appropriate, setting out a framework and tools to help individuals to identify their own levels of capability and their development and training needs i.e., competency matrix and skills profiles[107]. Institutional leadership needs to know what skills are needed, and where to find these people. And people, need a career path with professional development opportunities that is responsive to fast-changing technical situations.

[107] ICEDIG Milestone Report MS48 contains detailed information on competencies and skills profiles.

Recommendation 98: DiSSCo should create focus on harmonised tools and frameworks to help institutions and individuals understand and develop their skills capabilities, needs and professionals (such as: digitization maturity assessment model, competency matrices and skills profiles, career paths) and should make the case to address these with each collection-holding institution.

Following the priorities detected above, participants in the survey identified needs for supporting activities, with capacity enhancement and training actions among them, that would help to increase the impact of collections digitization in science and society. These activities comprise an enhanced involvement of citizens that can contribute as citizen scientists or volunteers and a more targeted training in addition to the general training for the professional staff and volunteers involved in digitization activities to address possible shortcomings.

There is a strong advocacy to strengthen interdisciplinary research and to improve general public engagement by making the digitized data available on openly accessible and user-friendly platforms.

5.5 Training and working better together

Training must be provided on a regular basis. Current training in digitization is generally inadequate, in terms of frequency, content, availability, professional credit and recognition (e.g., only 50% of staff receive initial training at the start of the job and even less enjoy recurring training). There are still substantial gaps in capacity building actions today and training at regular intervals is almost unheard of, which diminishes dramatically the chances to have staff working at leading edge with the state-of-the-art techniques. Keeping track of rapidly evolving techniques and acknowledging global standardisation endeavours require focused and expert training beyond the individual institutional approach, from mature organisations and highly experienced trainers. Even more critical is the fact that digitization staff often doesn't have the adequate background to effectively operate the hardware and/or software that is used to digitize collections. Some of the digitization staff are also not trained enough in general digitization skills, even on a very basic level. There is an overall and urgent necessity for increasingly specialised training for dedicated staff that should cover all aspects of digitization at all complexity levels.

Training needs vary significantly from one institution to another and cannot easily be categorised since a very broad lack of digital and to a lesser extent, analogue skills has been detected, with data skills standing out as a specific need. This widespread need for capacity building not only impedes digitization efforts but also prevents innovation in this area.

Before that, collection-holding institutions need leadership who understand the care and feeding of data (for longevity, re-use) and can plan cyber/human infrastructure needs and training accordingly or they must have access to experts with that knowledge.

Recommendation 99: In alliance with an appropriate training provider, DiSSCo should develop and promote executive/senior level training in collection-holding institutions, with a specific focus on collection leadership, mobilisation and use in the digital information age.

When DiSSCo is implemented, collection related work is likely to evolve from an institutionally based approach to a distributed infrastructure-oriented approach. This means that while currently most digitization activities are taking place within single institutions, with DiSSCo this could evolve and at least include coordination between institutions and countries regarding digitization strategies and priorities to efficiently create the best possible European digital collection. This would involve new and improved unified methods to tackle the challenges of collection management and digitization of collection objects and their associated data. A working approach based on functional units could help to streamline the digitization efforts and improve efficiency by reorganising the competencies in respect to the existing

capacities per institution. It could free organisations from following the strict definition of roles and could act as a facilitator for the mobility of work forces and researchers as well as the creation of international competence groups. At the very least, it would provide institutions a common vocabulary when discussing collaborative actions, common projects or very specific things like comparing annotation workflows.

To tackle this challenge, DiSSCo institutions must look beyond their own realm. Finland approached this challenge in 2010-2017 by establishing a national digitization centre funded by the European Structural Funds (ESF). The operation had to change its *modus operandi* several times but was largely successful.

Recommendation 100: DiSSCo institutions should form consortia that consolidate activities by launching national or regional centres for large scale digitization and offer out-sourced digitization services, training and other capacity building for in-house digitization at the institutions. These could be funded through European structural and investment funds.

Recommendation 101: DiSSCo centrally should launch thematic DiSSCo Centres of Excellence that support regional and national centres with technological innovations needed to ramp up the speed of digitization to the required levels of output. Such Centres may or may not be connected to digitization consortia/factories.

Looking beyond Europe, such an approach might also be developed in conjunction and/or alignment with institutions in other regions (e.g., USA), helping all to better develop the 21st century workforce.

5.6 Awareness raising and promotion in the Preparatory Phase

Finally, there is an urgent need for parallel initiatives around i) further awareness raising and training, and ii) development and promotion of pilot applications and exemplars as a means of convincing those that need to engage and participate to DiSSCo of the value of doing so. The former would introduce DiSSCo to newcomers and assist participants to better understand the existing plans and implementation strategies and their potential benefits for users. The latter, through the availability of working examples of good solutions (as has already been shown with the ICEDIG Digital Specimen Demonstrator) greatly accelerates understanding and helps to convince future users of the potential benefits. The latter could include ways to help institutions with small collections to get started on their digitization journey. Both activities must contribute to showing how DiSSCo can support the research goals of individuals and how DiSSCo can actively support and enhance the work of specific stakeholder groups.

Recommendation 102: To communicate and demonstrate the value of participating in DiSSCo, the DiSSCo Coordination and Support Office should initiate: i) further awareness raising and training, and ii) development and promotion of pilot applications and exemplars. Both activities must contribute to showing how DiSSCo can support the research goals of individuals and how DiSSCo can actively support and enhance the work of specific stakeholder groups.

Such activities are important also to refresh and deepen DiSSCo's understanding of requirements (and how those may be changing over time) and of the practices and activities of individuals working with the data of natural sciences collections.

6 Governance and business model

6.1 Governance of the DiSSCo Programme

By entering the ESFRI Roadmap in October 2018, DiSSCo initiates its Preparatory Phase (PP); a period that will conclude with the formation of the legal entity of DiSSCo. The PP governance proposed in the present

document is considered here as part of a succession of models for each programme phase, as illustrated in Figure 7, extracted from the DiSSCo European Memorandum of Understanding (MoU).

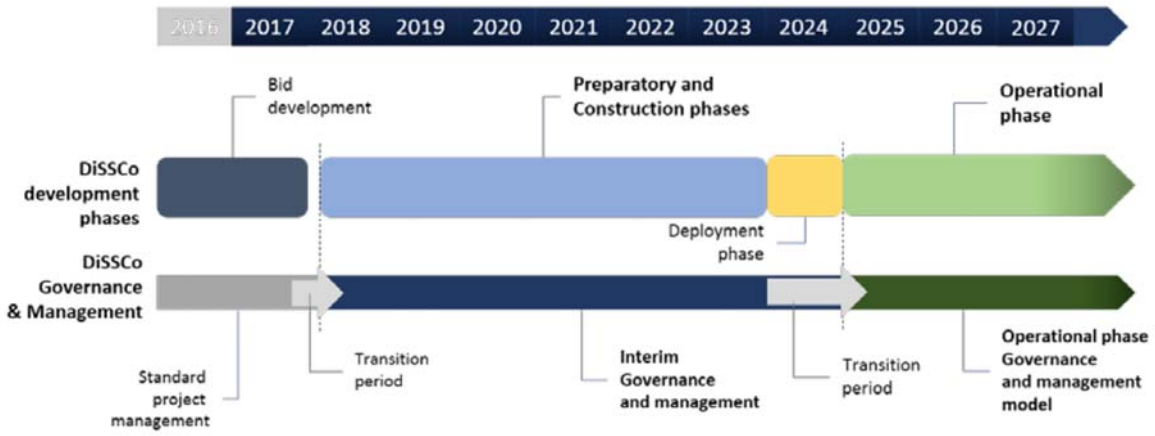


Figure 7: Governance and management models during the different programme phases

The Preparatory Phase is entirely oriented to guarantee DiSSCo reaches maturity to be constructed and fully operational afterward. Readiness in five specific dimensions (data, technical, financial, governance and scientific) demands a governance structure able to respond to new demands and requirements during the preparatory phase while guaranteeing a smooth transition towards a country-based research infrastructure.

It becomes necessary that the Steering Committee, the governance model of DiSSCo that sufficed during the first years (up to 2019) moves in 2020 to a more democratic and inclusive form of governance, a General Assembly (GA), that levels the playing field for national consortia involvement in the developments of the preparatory phase while securing the sustainability and operation of the Coordination and Support Office (CSO) activities.

It is necessary that this interim governance model guarantees the financial sustainability of the preparatory phase, anticipating the possible negative effects of the potential critical funding path (see 6.3.1 below) on the sustainability of the Coordination and Support Office activities and consequently, its direct impact in communication, engagement, the international positioning, the scientific and technical missions, service deployment & others domains.

Additionally, the interim model must ensure excellence in the way DiSSCo achieves all matters related to the scientific and technical mission and service objectives. In doing that, it is necessary to constitute several independent advisory bodies (6.2.4 below) to provide expert consultation services. These include the Scientific Advisory Board (SAB), a Technical Advisory Board (TAB), a Funders Forum (FF) and a Stakeholders Forum (SF).

During the Preparatory Phase, the research infrastructure will constitute an institutions-based General Assembly (GA) that advises and takes decisions on the implementation of the tasks defined through a portfolio of interlinked projects. At that stage, it is necessary the model guarantees the continuity of the community active in running the projects and other activities while stimulating the creation of new national consortia in those countries that have not set up one yet.

6.2 DiSSCo Preparatory Phase governance

6.2.1 Requirements for a new model

The new Preparatory Phase governance model pursues the empowerment of DiSSCo national nodes that progressively will take a key role in the engagement of DiSSCo country-level funders. The complexity of the task will be partially pared down by the Coordination and Support Office's actions, including the formulation and implementation of relevant consulting bodies, the Funders Forum and the Stakeholders Forum.

Those forums, meant to be part of the DiSSCo Prepare project, will constitute key instruments for the sustainability of the research infrastructure and its position at national, regional and international scale during implementation. The way how these two bodies engage and operate will be crucial to reach successfully the implementation phase. In that sense, there are open questions, mostly related to engagement and operation.

DiSSCo will need to establish effective ways to engage authorities at this stage when DiSSCo is still an institution-driven research infrastructure and the bodies do not have any decision-power. The research infrastructure will need to be able to prove a level of maturity, at technical and governance level, which is unusual in such an early stage. At the same time, DiSSCo will need to develop a sound engagement with existing initiatives of strategic interest at national level, in order to attract the attention towards the potentialities of the services that will be provided in the future.

For operation, it will be necessary to decide which are the necessary instruments for the operation of these two bodies that guarantee the achievements of their goals.

Tasks definition, coordination with other advisory and decision- bodies to achieve real contributions and mechanisms to communicate the content of the discussions to a realm of different national realities have still to be defined.

Furthermore, the new governance model will need to guarantee that the nodes represented receive the support needed to adopt the best infrastructure possible for the national interest. Clear targets and guidelines on requirements and best practices in advocacy and communication are still to be defined.

A third domain that will need an equal investment of resources refers to scientific and technological excellence. How to position DiSSCo in an existing fast-paced environment at scientific and technical levels will be a main challenge and requires both, scientific and technical guidelines and, a trustful stakeholder network. It will be necessary to develop an engagement strategy beyond the environmental domain, investing in participation in international fora, collaboration in transdisciplinary projects, knowledge exchange and, above all, the independent expert guidance of the Scientific and Advisory Boards (SAB & TAB).

How the recommendations of the SAB and the TAB percolate down the different levels of execution and expertise across different activities (i.e. synchronisation groups, technical team, etc.) should be analysed to guarantee the alignment necessary to guide DiSSCo future developments.

Last, but not least how the actions towards positioning DiSSCo benefits from the interactions in the Stakeholders Forum and contribute to the SAB/TAB discussions, should be also investigated to build a communication that conveys key messages in both directions.

6.2.2 General Assembly

6.2.2.1 Scope

The General Assembly (GA) will be the decision-making body for DiSSCo during its preparatory and transition phase which means the period starting with the admission of DiSSCo onto the ESFRI roadmap

and ending with the creation of a separate legal entity for DiSSCo. The financial contributions will define the levels of participation at the General Assembly. The mode of operation of the new body requires specific terms of procedure.

The GA aims to be a more inclusive form of governance and to provide active participation to all its members. Comprising representatives from all DiSSCo consortium (MoU) members, it provides the forum for multilateral discussions on DiSSCo developments during its Preparatory and Construction phases and the necessary policies and strategies to be implemented in the early stages of the Construction and Operation phases.

6.2.2.2 Membership

Different types of membership will be established, depending on the possibilities of members to contribute financially and their state of development towards a national DiSSCo consortium.

In general, DiSSCo aims for representation - on the national level - through a developed national consortium or node. Within each node, the partners will sign a national Memorandum of Understanding (MoU) to formalise their relationship and appoint a representing institution. This institution then signs a European MoU with DiSSCo in the name of the national node. Once the European DiSSCo MoU has been signed, the member can apply for membership to the GA with a written letter to the Chair and a copy to the Coordination Office.

However, DiSSCo recognises that not all countries are the same when it comes to the maturity of national networks of natural science institutions. Therefore, an exception to the desired representation mode of a national mode will be granted for the duration of the interim governance. This exception states that in such cases where a national node has not yet formally formed, natural science collection-based institutions can become a member of the DiSSCo GA.

Types of Membership:

- Members of the DiSSCo GA will be national consortia (or institutions) that have signed the DiSSCo European MoU and committed to provide substantial in-kind contributions towards DiSSCo Preparatory and Transition Phase and have signed an agreement for the provision of cash contributions to the Preparatory and Transition Phase for a minimum period of two years, which will be renewable. The Consortium of European Taxonomic Facilities is a DiSSCo member but will be exempted from the second condition. Members will have one vote in the GA.
- Associate Members of the DiSSCo GA will be national consortia (or institutions) that have signed the DiSSCo European MoU and committed to provide substantial in-kind contributions towards DiSSCo Preparatory and Transition Phase. Associate Members will not have voting right in the GA.
- Observers will be either a national consortium (or institution) that has expressed in writing its interest in becoming a member or associate member, a governmental entity that has expressed in writing a commitment to provide financial support to DiSSCo or entities with a European or international dimension, with a mission and objectives that are deemed pertinent to DiSSCo. Observers will attend the meetings of the GA but will not have a voting right.

6.2.2.3 Decision-making

As the decision-making body of DiSSCo, the GA will need to take decisions concerning:

a) the financial contributions of members and adoption of the budget for the Preparatory and Transition Phase;
b) acceptance of new members, associate members and observers;
c) the composition of the DiSSCo SAB or TAB;
d) appointment of auditors;
e) approval of the annual or interim accounts;

f) appointment of the Chair and Vice-Chair;
g) appointment (and dismissal) of the DiSSCo Coordinator, who shall carry out the day-to-day management of DiSSCo in accordance with the decisions and instructions of the General assembly;
h) approval of the annual reports (technical, scientific and organisational);
i) appointment of DiSSCo programmes and project boards;
j) establishment of non-executive advisory bodies;
k) approval of arrangements (technical, organisational and financial) proposed by the DiSSCo Coordination Team, necessary for an efficient transition into the Construction Phase.

These decisions will either be taken during GA meetings or intersessionally through a dedicated platform.

6.2.3 Coordination and Support Office

The Preparatory Phase of DiSSCo has started in 2018 and aims at the improvement of the overall Implementation Readiness Level (IRL) of the infrastructure across all the relevant dimensions of its future operation. To achieve this, DiSSCo is currently linking together a series of externally funded projects that distinctly contribute into one or more of those dimensions. These projects are part of a multi-partner multi-million work programme that includes tens of institutions and hundreds of contracted researchers, curators, software engineers and managers. As the executive body of the DiSSCo Research Infrastructure (RI) preparatory phase, the DiSSCo Coordination and Support Office (CSO), is acting as interim support office until the RI becomes operational. It holds the overall responsibility for the successful coordination of the pan-European projects linked to DiSSCo and the synchronisation with national activities.

The CSO undertakes multiple tasks, pursuant to the European DiSSCo MoU:

a) Implement agreed decisions upon KPIs and produce reports for the GA;
b) Develop and implement a strategy for the support and engagement of NTFs (incl. formation of new nodes);
c) Coordinate and support DiSSCo Advisory Boards at scientific and technical level (SAB and TAB respectively);
d) Coordinate and support the DiSSCo Funder Forum and the DiSSCo Stakeholders Forum;
e) Communicate progress of the initiative to full members, associated members and any national or European authority upon request;
f) Coordinate the development and implementation of projects relevant to the development of the DiSSCo RI;
g) Specify and resource tasks for internally managed programmes and projects developed under these programmes;
h) Draft agreements and administer contracts (in collaboration with nodes and facilities) as needed;
i) Specify and oversee procurement of software, hardware and services for the development of the DiSSCo technical infrastructure;
j) Develop in-house expertise to underpin its technical, policy, capacity building, and management responsibilities;
k) Undertake software engineering tasks towards the development of the DiSSCo RI;
l) Compile policy drafts and organise consultation rounds;
m) Monitor the progress of policy, process and operations harmonisation across all participating facilities;
n) Provide expert advice to facilities and nodes as per their request and available capacity.

To a certain extent, several aspects of the above list of responsibilities are incorporated into the work programmes of the DiSSCo-linked projects. Part of the crucial list of activities of the DiSSCo CSO, however, are not (or cannot) be included in the externally funded DiSSCo-linked projects because of limitations in the scope of those projects. Those activities will be supported, for the duration of the Preparatory phase,

by the resources contributed by the DiSSCo Consortium, and specifically the DiSSCo Governing Body full members. This new governance and financial model ensures the operation of the DiSSCo CSO during the preparatory phase with emphasis on capacity enhancement, outreach and engagement, alignment and coordination, training, business development and infrastructure piloting.

Over the last three years, the small team of three people gradually increased to a group of nine professionals (as of March 2020) working across the different operational aspects of the office. With the support of the winding-up Steering Committee and the DiSSCo nodes, the CSO has now access to an efficient toolkit, which allows it to navigate the strategic, technical and organisational complexities of the four-year Preparatory Phase.

6.2.4 Advisory Bodies

The advisory boards, formed by international experts, will play a key role in highlighting issues to consider, tabling risks to mitigate, or identifying specific new challenges to address. They will ensure that both areas – technological infrastructure and scientific coverage – are aligned and developed coherently.

In doing that, the advisory bodies may be supported by other consulting bodies and forums established during the preparatory phase.

6.2.4.1 Scientific Advisory Board (SAB)

DiSSCo is a Research infrastructure aiming at providing end-user scientific services in support of frontier data-intensive science. As such, its development and operation are heavily predicated upon its continuous ability to serve the needs of a diverse set of scientific, political and industrial users. As those needs change, DiSSCo needs to constantly develop its services through a permanent and productive dialogue with its users. To this end, the DiSSCo Scientific Advisory Board will take stock in this process by providing expert advice across the DiSSCo governance and executive structures.

a) The DiSSCo Scientific Advisory Board (SAB) mission will be to provide expert consultation services in all matters related to the scientific mission and service objectives of the DiSSCo Research Infrastructure. It has the following objectives:
b) Understand and analyse relevant scientific challenges and translate them into research infrastructure needs at European and Global scale;
c) Work closely together with science users across scientific disciplines to understand and describe current and future data and infrastructure service requirements;
d) Define metrics for assessing the overall performance of the Research Infrastructure in relation to its scientific agenda and expected impact across fields of science;
e) Provide expert advice on the needs to enhance digital capacity across DiSSCo users and partner facilities;
f) Draft annual science reports and provide input to the development of the mid- and long-term science strategy of DiSSCo;
g) Provide ad-hoc expert advice on requests from the DiSSCo governance bodies and act as external advisory boards to DiSSCo-linked projects;
h) Represent DiSSCo in external meetings, presenting the technical vision and progress of the infrastructure;
i) Evaluate internal project proposals and review key documents.

6.2.4.2 Technical Advisory Board (TAB)

DiSSCo, as a research data infrastructure, is heavily investing in its ability to link together and serve data classes related to the European natural science collections. To this end, it plans to deploy a comprehensive data model, which enables the development of world-class e-Services for a diverse set of scientific, industry and policy audiences.

DiSSCo does not develop its technical architecture in isolation. Instead, it makes use and augments existing underlying and complementary data systems and services, whilst innovating where necessary to reach the required technical readiness level.

The Data Management Plan (DMP) of DiSSCo will provide overall guidance on the future implementation and operation of the infrastructure. The implementation of this plan will require a series of key technical decisions to be made, whilst the ever-changing technological landscape mandates further amendments to the existing DMP.

The DiSSCo Technical Advisory Board (TAB) mission will be to provide expert consultation services in all areas related to the technical sphere of operation of DiSSCo. It has the following objectives:

a) Collect and analyse scientific priorities of the infrastructure and provide recommendations on how to address these in services and the technical development roadmap;
b) Provide expert advice to the DiSSCo governance and management bodies;
c) Propose metrics for measuring the performance of the technical teams across the DiSSCo projects;
d) Monitor and report on the technical preparation and implementation progress of DiSSCo and specifically in relation to the ongoing portfolio of DiSSCo-linked projects;
e) Draft annual reports on the accounts of technical developments and future technical roadmap and provide advice on the technical priorities of new projects and project proposals;
f) Provide ad-hoc expert advice on requests from the DiSSCo governance bodies and act as external advisory boards to DiSSCo-linked projects;
g) Represent DiSSCo in external meetings, presenting the technical vision and progress of the infrastructure.

6.2.4.3 Funders Forum (FF)

During the DiSSCo Prepare activities, undertakings will be pursued to develop enough level of trust by national funders in the benefits DiSSCo will bring for their national efforts and scientific agendas. Discussions at national level on their involvement in the RI will contribute to further set up a refined governance and business structure for the subsequent phases of DiSSCo.

Representatives from national authorities will participate as representatives in the future governance model of the Research Infrastructure during its construction and operational phases. In its role, the mission of FF will be to provide recommendations on both, strategic and operational planning, that guide DiSSCo on how to set up a smooth transition from a project-based model to a sustainable organisational and financial model, well-embedded into national roadmaps.

The Coordination and Support Office activities together with DiSSCo Prepare Work package 8 team will set up and support the operation of the FF, ensuring a fluent communication with the interim General Assembly.

6.2.4.4 Stakeholders Forum (SF)

Ongoing discussions with initiatives working in similar areas of interest at European and International level led to DiSSCo's involvement in a series of initiatives and projects, positioning DiSSCo as a key Research Infrastructure, a foundational layer for cross-cutting research.

During the DiSSCo Prepare project, it is planned to set up a Stakeholders Forum to ensure permanent engagement. This is intended to strengthen engagement with the stakeholders i.e., the institutions that support and develop the DiSSCo Research Infrastructure. The way stakeholding institutions relate to and work with DiSSCo differs from country to country. Most have constituted a National Task Force (NTF) with a leading spokes institution, while in other countries institutions operate individually.

The mission of the Stakeholders Forum will be to provide critical feedback from a scientific, technical, data, organisational and financial perspective during the Preparatory Phase.

The mode of operation of the Stakeholder Forum will be part of the developments within the DiSSCo Prepare project.

6.2.4.5 Industry Forum (IF)

To channel needs of the community to industry and to obtain technological and other perspectives on what is possible it can be helpful at some moment to establish an Industry Forum (IF).

6.2.5 Coordination Bodies

In addition to the DiSSCo Coordination and Support Office (6.2.3), the Strategic Alignment of Projects (SAP) group and the Synchronisation Groups (SG) have been established to improve coordination within the DiSSCo Programme.

6.2.5.1 Strategic Alignment of Projects (SAP)

Running a programme of multiple simultaneous projects linked to the DiSSCo vision (e.g., ICEDIG, SYNTHESYS+, COST MOBILISE, DiSSCo Prepare) (Figure 8) creates a complex web of tasks that must be properly coordinated to avoid overlap, duplication and conflicting results.

The Strategic Alignment of Projects (SAP) group consists of the Coordinators of the DiSSCo-linked projects and the DiSSCo Coordinator. Its mission is to guarantee the best use of the resources available and the achievement of the programme objectives.

SAP, as its title indicates, works towards the overall alignment of activities across the existing projects. SAP has recently set up five Synchronisation Groups (SG) to enhance the work led by SAP in that matter.

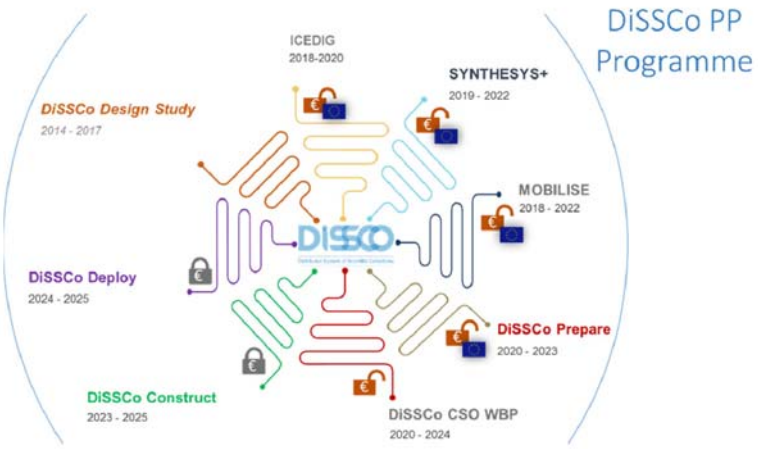


Figure 8: DiSSCo Programme of linked projects

6.2.5.2 Synchronisation Groups (SGs)

Several thematic areas of related work have been identified as cross-cutting topics relevant towards the readiness of the future DiSSCo Research Infrastructure. To enable efficient alignment and ensure adequate support between all the DiSSCo linked activities, five Synchronization Groups (SG) have been established:

1) Policy harmonisation & International Coordination;
2) Data (Standards and other common resources);
3) Tools and Services (Data models, management, publishing pipelines & services);
4) Digitization; and,
5) Training and capacity enhancement.

The SGs act under the supervision of the SAP and comprise work package and task leaders from across the DiSSCo linked projects to guarantee the alignment of outcomes. The task of the SGs is to identify gaps,

complementarities and/or overlaps among the different thematic streams of DiSSCo, ensuring that the work undertaken under different projects is sufficiently coordinated.

These SGs act as specialist groups advising across domains during the Preparatory Phase. Having accrued highly specialised knowledge during this phase and representing a highly specialised critical mass, the SGs will most likely continue in further developments during the implementation and operational phases. How that will evolve, which role they should adopt and how their relationship with the advisory bodies will be established are open questions that will need to be answered during the DiSSCo Prepare project.

6.3 Options for sustainable funding

6.3.1 The critical funding path for DiSSCo

ESFRI-endorsed European Research Infrastructures usually go through a succession of semi-standardised funding models. Those models can be summarised as follows:

A. **Design and early phase**. During this stage Research Infrastructures rely on project-based resources, which are usually self-funded (research infrastructure consortiums) and supplemented by European Commission resources. In the context of DiSSCo, ICEDIG project has significantly contributed into the design of the research infrastructure and ICEDIG resources flanked investments coming from the DiSSCo facilities.
B. **Preparatory Phase**. This constitutes the first formal development phase of each of the new research infrastructures. Infrastructures in this phase, usually secure funding from the European Commission to perform critical tasks related to the preparation of the infrastructure across key organisational, legal, financial and technical dimensions.
C. **Construction and operation phases**. Despite these phases having distinct objectives, they are financially coupled together as they usually fall under the same business model, i.e., long-term financial commitments of national governments.

Considering the different funding models, a RI needs to transition between them as it progresses from one development phase to the next. This creates a critical funding path that needs to be preserved throughout, for the RI to retain continuous funding. The critical funding path (Figure 9) can be flanked with funding coming from additional sources. It is, however, important that core operations for each phase are fully funded through resources in the critical funding path to ensure continuity of core operations of the RI. Disruptions in the critical funding path risk the interruption of core operations, especially when a gap opens between end of preparations and beginning of implementation due to unsolved political difficulties.

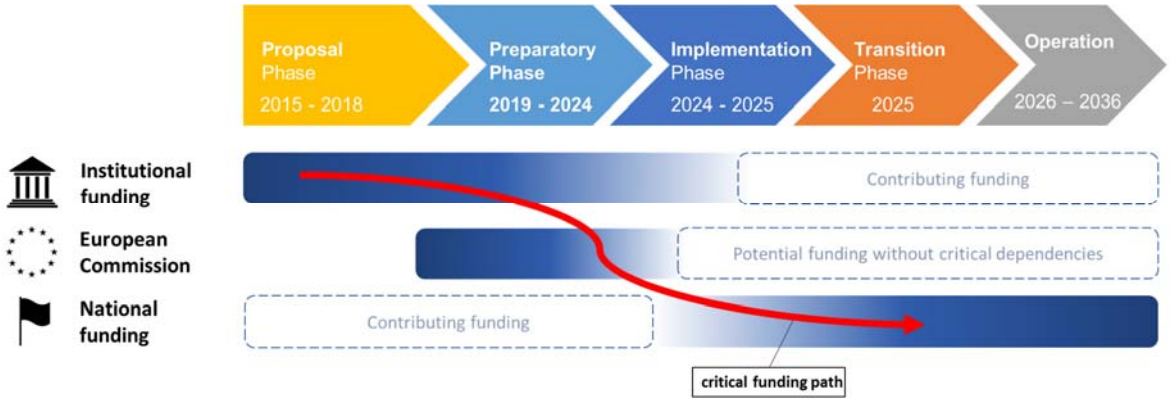


Figure 9: Funding sources as they correspond to the different development phases of the DiSSCo RI

Like other European research infrastructures before it, DiSSCo transitions between the funding models A and B[108], before being required to secure longer-term funding by national governments, C. At this point the decision of national bodies towards a long-term commitment to the RI hinges on multiple criteria.

6.3.2 Criteria influencing national funding commitment towards DiSSCo

After discussions with national contact points and analysis of national decision-making processes towards RI funding commitments, a list of key criteria as perceived at the present time (spring 2020) can be summarised as the following:

1. RI value proposition in relation to nationally set agendas in research, innovation and science policy.
2. RI value proposition in relation to national Smart Specialisation Strategies.
3. Previous national level investments in the specialisation area of the RI, including operation of national facilities.
4. Number (as percentage of total researchers nationally) of scientific users of the RI and level of maturity of science linked to the European RI facilities.
5. Level of prior understanding of national decision-makers of mission and key objectives of the RI.
6. Operation at national level of a roadmap for the development of research infrastructures and level of alignment of this roadmap to the European priorities (i.e., ESFRI).
7. Perceived national need for further strengthening of the science and science-policy collaboration level with other countries in the specialisation area of the RI.
8. Potential for the European RI to improve the scientific or infrastructural performance of science linked to the RI national facilities
9. Contribution of the RI to monitoring and international reporting obligations of the national government (e.g. towards achieving internationally agreed targets).
10. Level of clarity of mission and objectives of the RI and uniqueness of RI in wider landscape of RIs.
11. Level of perceived urgency for the scientific scope of the RI.
12. Clarity of the governance model of the RI and level of prior contributions of the governments to the formulation of governance structures.
13. The legal entity form that the RI chooses to adopt.
14. Clarity of the business model and conformity with national and international best practices.
15. Perceived public opinion on the value of the RI and the operation of the linked national facilities.

Circumstantial criteria, such as internal political volatility or national fiscal capacity might temporarily also affect the position of a national government towards committing to the construction and operation of a European RI.

Considering the complexity and diversity of the criteria that might affect the decision for future funding of a RI, as well as, the lack of long-term financial commitment of countries towards RIs, it becomes apparent that the sustainability of a RI cannot be easily secured. Continuous adjustment of key operational and organisational parameters of the RI might be needed in order to retain the interest and commitment of national governments.

A crucial factor is the diversity of national funding flows towards research infrastructures, and specifically the lack of flexibility from governmental budgets in adjusting funding flows to the changing development requirements of the infrastructures.

The funding instruments and funding flow channels, which national governments opt for in support of the construction and operation of a European RI might be an important factor for the longevity of such commitments.

[108] Commencement of the DiSSCo Prepare Preparatory Phase project in January 2020.

6.3.3 Direct funding model option

In this model, national governments directly commit resources to the RI. Commitments towards the RI usually last between from two to five years, depending on the development phase of the RI and the funding evaluation cycles of the national government. National governments can delegate the funding responsibility towards European RIs to corresponding governmental departments or scientific councils. This model (Figure 10) has been the cornerstone of the financial model of RIs. National governments directly opt to financially support the implementation and/or operation of a European RI. National governments commit this way in a series of RIs, which many times address similar national-level scientific or reporting needs.

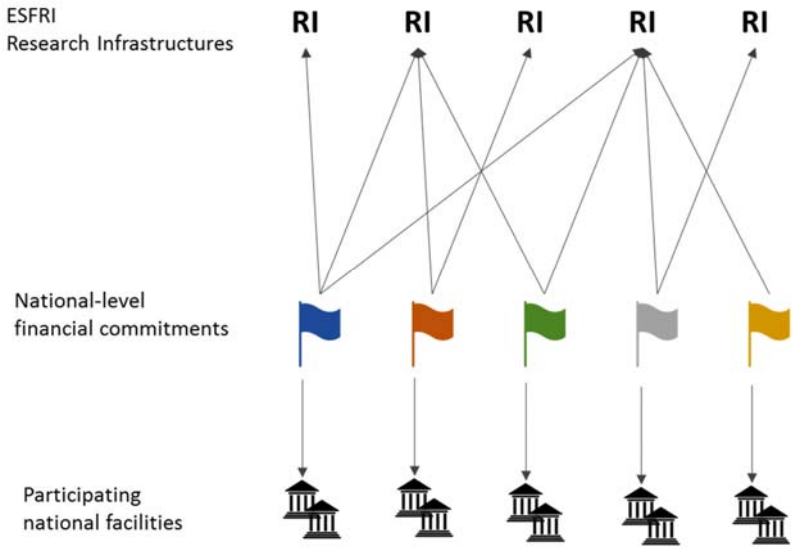


Figure 10: Illustrating the direct funding model for RIs
Arrows depict funding flows.

Despite that this model continues to be the predominant one for the funding of RIs, a SWOT analysis (Figure 11) illustrates that its limitations and caveats seem to become more apparent as the number of RIs in the landscape increases.

Figure 11: SWOT analysis of the direct funding model for RIs

6.4 The need for new funding instruments

Infrastructure funding is heavily dependent on the capacity of RIs to present strong science-policy narratives aligned with corresponding national-level agendas. To mitigate the risk of the volatile nature of government-level funding, RIs need either: i) to be able to diversify their funding streams, complementing national-level resources with other income; and/or ii) improve the stability of government-derived funding, securing funding cycles of a minimum of five years. Such longer-term funding cycles would enable the RIs to implement multi-year action plans, in accordance with typical 5-year science strategies.

6.4.1 Diversification of funding streams

The non-profit and public nature of European RIs can be perceived as a limitation to the capability of the RI to offer commercial services to relevant industry users. Additional limitations, regarding the source of income from commercial activities, also often derives from the legal vehicle that a RI has chosen to adopt. RIs that opt, for example for the adoption of the legal entity of the European Research Infrastructure Consortium (ERIC) are bound by certain limitations regarding the percentage of total income that derives from commercial activities. Similar limitations might be sometimes self-imposed by the governing bodies of the RIs, especially given that those governing bodies are usually comprised by representatives of the national funders that support the RI.

The above create a situation where RIs are perceived to have limited capabilities for commercial exploitation of their services, limiting the options to significantly diversify their income sources.

6.4.2 National funding frameworks

The increasing number of European RIs puts pressure to governments to invest in a multitude of RIs. The overhead for national governments linked to these commitments is not only relevant to the financial impact on national budgets but also to the administrative costs of monitoring and evaluating these investments and participating in their respective governance structures.

Consequently, the RIs often find themselves in a position in which they need to compete to secure funding. Furthermore, on many occasions, such competition is more prominent between RIs that operate within the same thematic area (e.g., environment, in the case of DiSSCo). That is because such RIs usually appeal to budgets owned by the same governmental departments or science councils. Such intra-domain competition for funding can, in turn, limit the potential for cross-infrastructure collaboration and subsequent interoperability.

A further factor influencing success in securing funding is strong national emphasis on specific areas of science as being most in the national economic interest. If an RI looking for funding is not addressing one of those areas, the opportunities to secure investment can be lower.

6.4.3 Consolidating national funding – The hourglass model

The increasing number of European RIs that apply for national level funding puts additional financial and administrative strain on national governments. This establishes distinct and usually isolated funding streams between funder and an RI. Such funding streams are difficult to establish and once in place they largely remain unchanged, in terms of funding size, unless the funder agrees to re-evaluate its contributions. Research infrastructures, however, are dynamic organisations that undergo significant changes in their operational and investment plans, as they evolve through their development phases. An RI at the beginning of its implementation phase typically requires significantly more financial resources (construction costs) than a RI already in its operational phase. Furthermore, these distinct funding flows do restrict governments from better evaluating how they can better balance funding allocated to a specific strategic direction or international reporting obligation (e.g., biodiversity monitoring) across the RIs relevant to a specific strategy or obligation. It becomes apparent that distinct and isolated funding flows from national governments directly towards RIs can reduce the level of funding responsiveness needed to meet the changing needs of the RIs.

To address this issue, several countries are now introducing an intermediate layer of coordination of scientifically related activities. In this model, a national bureau coordinates the funding allocations and flows towards the underlying international RIs and across the national participating facilities. This enables countries to retain a more agile way through which they can allocate funding towards relevant national efforts on one hand, and on the other hand, to the corresponding international infrastructures (Figure 12). This is needed, for example in the domain of biodiversity monitoring where one can recognise the role of several European and international Environmental RIs (e.g., DiSSCo, LifeWatch, eLTER, GBIF, etc.). Here, coherent interoperation of the national nodes of these infrastructures is imperative to meet national targets and global obligations in biodiversity monitoring and conservation.

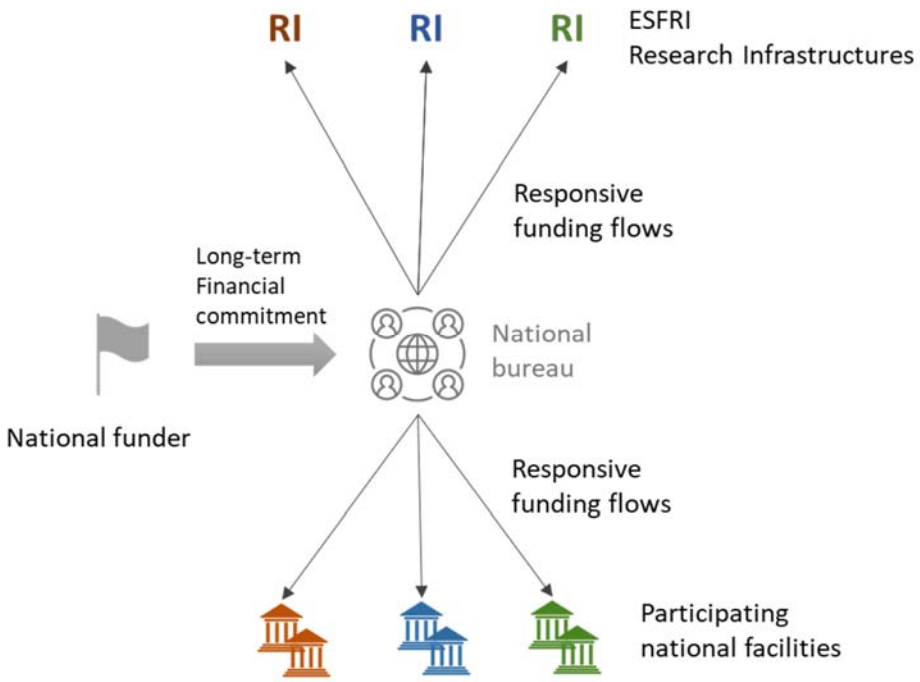


Figure 12: Introducing a funding thematic coordination layer between national facilities and RIs

There are already successful examples that follow this model of hourglass funding for RIs.

In the Netherlands, NLBIF[109] has gradually transformed from a GBIF national node to a more inclusive centre for biodiversity information in the country. In this respect, NLBIF is now a comprehensive organisation that looks after the Dutch participation of relevant facilities to the overarching European and international infrastructures, including DiSSCo and GBIF. A similar model is already in place in Finland with the FinBIF organisation[110]. Other countries are examining this model as a mechanism to improve the efficiency and effectiveness of the national level commitments to infrastructure development and operation at national and European level.

6.4.4 RI cluster funding

For more than 10 years, the European Commission has funded RI cluster projects linked to the identified major thematic domains in the ESFRI roadmap[111]. In the environmental domain, the cluster projects ENVRI (2011-2013), ENVRIplus (2014-2018) and ENVRI FAIR (2018-2022) aim to build a set of commons across the participating RIs and to improve the level at which RIs cooperate together and provide FAIR data and services. Despite significant results deriving from the investments in cluster projects, RIs remain still largely disconnected and with limited achievements in their cross-RI interoperability. Furthermore, interoperability needs are present between RIs from different ESFRI domains. For instance, DiSSCo is heavily concerned with achieving better links between physical specimens and deposited genomic information. In this respect, DiSSCo (environmental domain) and ELIXIR (health and food domain) have already started working together towards this goal. Similar examples can be identified between DiSSCo and E-RIHS (heritage interpretation)[112] where the two infrastructures investigate how to exchange experience and technology for mass scale digitization of objects. These examples demonstrate that the need for investments in cross-RI collaboration and interoperability must be predominantly driven by the

[109] https://nlbif.nl.
[110] https://laji.fi/en.
[111] https://esfri.eu.
[112] http://www.e-rihs.eu/.

scientific, operational and strategic needs, rather than being based on the de facto categorisation of the RI into an ESFRI domain.

Such focused and demand-driven investments can be more productive and support RIs to achieve the level of needed cross-RI data and system interoperability. They will, furthermore, enable RIs to provide better, more focused and tailored to the needs, scientific services to their audiences and specifically support new users that investigate multi-disciplinary research objectives.

6.4.5 Governmental securities – shared liability

Frontloaded investments are a typical need for developing RIs, usually required either at the beginning of the implementation phase (construction) or during planned re-investment cycles for renewing equipment and enhancing software). Start-up and construction costs can add up to multiple times the projected annual operational costs of an RI and can represent a considerable commitment. However, the typical financing committed by national funders aims to mostly cover operational budgets of the RIs.

The challenge is especially acute for those RIs that typically do not procure or construct tangible assets (property, plant, equipment) that can be used as collaterals for securing financing. RIs such as DiSSCo are both distributed in nature and heavily reliant on process and software infrastructure to deliver their value. They require investments that do not traditionally lead to strong (tangible) asset value from the accounting point of view. Furthermore, the operation of distributed RIs is predicated on investments made across multiple legal entities that together are contributing towards achieving the DiSSCo vision. This further perplexes the ability of a RI legal entity to access finance.

To address this issue, RIs can investigate the opportunity to leverage complementary financial instruments, such as European Structural and Investment Funds and funds from the European Investment Bank (EIB). EIB has relevant programmes (e.g., EIB InnovFin and InnovFin Advisory programmes) that enable public organisations, and particularly research and innovation organisations, to benefit from accessing large scale financial instruments. InnovFin Advisory, for example assists RIs to improve their overall bankability i.e., their ability to access financing through commercial or public financial institutions. The bankability aspects of RIs is still a topic that is not well investigated and further work in this area would allow us to better understand how DiSSCo can make wider use of financing options available, in order to improve financial capacity.

Recommendation 103: DiSSCo should investigate what is needed to improve bankability (likelihood of financial success) against the range of financial/investment instruments (e.g., European structural and investment funds, European Investment Bank programmes) available to complement national government funding.

Further work is needed to understand what models need to come in place to enable RIs to access financing instruments. Such models can include for instance the extension of the financial liability of funders beyond the annual financial commitment. Shared liability models can significantly improve the overall bankability of the RI and allow for more effective financial planning. They require, however, a new approach in the way national funders see their responsibility towards these RIs at national and at European level.

Such work may lead to future consideration of alternative business models, especially in the way that software development, enhancement and maintenance and infrastructure operations are performed.

Recommendation 104: DiSSCo should investigate shared liability models for more effective financial planning and how these may lead to alternative business models for DiSSCo.

7 Conclusions

Collection digitization efforts have reached most collection-holding institutions across Europe. Much of the leadership and many of the people involved in digitization and working with digital collections wish to take steps forward and expand the efforts to benefit further from the already noticeable positive effects. The collective results of examining technical, financial, policy and governance aspects (summarised in the present document as the results of the EU ICEDIG project, grant agreement No. 777483) show the way forward to operating a large distributed initiative i.e., the Distributed System of Scientific Collections (DiSSCo) for natural science collections across Europe. Ample examples, opportunities and need for innovation and consolidation for large scale digitization of natural heritage have been described. Numerous (104) recommendations have been made to be considered by other elements of the DiSSCo Programme of linked projects i.e., SYNTHESYS+, COST MOBILISE, DiSSCo Prepare, and others to follow, and the DiSSCo Programme leadership as the journey towards organisational, technical, scientific, data and financial readiness continues.

However, several significant obstacles must be overcome as a matter of priority if DiSSCo is to move beyond its Design and Preparatory Phases during 2024. Specifically, these include:

Organisational:

- Strengthen common purpose by adopting a common framework for policy harmonisation and capacity enhancement across broad areas, especially in respect of digitisation strategy and prioritisation, digitisation processes and techniques, data and digital media publication and open access, protection of and access to sensitive data, and administration of access and benefit sharing.
- Pursue the joint ventures and other relationships necessary to the successful delivery of the DiSSCo mission, especially ventures with GBIF and other international and regional digitisation and data aggregation organisations, in the context of infrastructure policy frameworks, such as EOSC. Proceed with the explicit aim of avoiding divergences of approach in global natural science collections data management and research.

Technical:

- Adopt and enhance the DiSSCo Digital Specimen Architecture and, specifically as a matter of urgency, establish the persistent identifier scheme to be used by DiSSCo and (ideally) other comparable regional initiatives.
- Establish (software) engineering development and (infrastructure) operations team and direction essential to the delivery of services and functionalities expected from DiSSCo such that earnest engineering can lead to an early start of DiSSCo operations.

Scientific:

- Establish a common digital research agenda leveraging Digital (extended) Specimens as anchoring points for all specimen-associated and -derived information, demonstrating to research institutions and policy/decision-makers the new possibilities, opportunities and value of participating in the DiSSCo research infrastructure.

Data:

- Adopt the FAIR Digital Object Framework and the International Image Interoperability Framework as the low entropy means to achieving uniform access to rich data (image and non-image) that is findable, accessible, interoperable and reusable (FAIR).
- Develop and promote best practice approaches towards achieving the best digitisation results in terms of quality (best, according to agreed minimum information and other specifications), time (highest throughput, fast), and cost (lowest, minimal per specimen).

Financial

- Broaden attractiveness (i.e., improve bankability) of DiSSCo as an infrastructure to invest in.
- Plan for finding ways to bridge the funding gap to avoid disruptions in the critical funding path that risks interrupting core operations; especially when the gap opens between the end of preparations and beginning of implementation due to unsolved political difficulties.

Strategically, it is vital to balance these multiple factors – organisational and political, technical and engineering, scientific and data, financial and legal, operational and governance – against one another to achieve the desired goals of the DiSSCo programme. Decisions cannot be taken on one aspect alone without considering other aspects, and here the various governance structures of DiSSCo (General Assembly, advisory boards, and stakeholder forums) have a critical role to play over the coming years.

8 References

8.1 ICEDIG project deliverables

The following ICEDIG project deliverables have been referred to at various points in the present document via footnotes. They are tabulated here in their entirety for convenience.

Number	Title	Date	DOI
D1.2	Final report	April 2020	
D2.1	Inventory of criteria for prioritization of digitization of collections focussed on scientific and societal needs	July 2018	10.5281/zenodo.2579156
D2.2	Prioritizing scientific and societal needs for data using small and private collections, Appendix 2, Appendix 5	October 2018	10.5281/zenodo.2582995
D2.3	Design of a collection digitization dashboard	March 2019	10.5281/zenodo.2621055
D2.3	Design of a collection digitization dashboard	March 2019	10.5281/zenodo.2621055
D3.1	Quality control methodology for digitization operations	April 2019	10.5281/zenodo.3469521
D3.2	State of the art and perspectives on mass imaging of microscopic and other slides	April 2019	10.5281/zenodo.3364481
D3.3	State of the art and perspectives on mass imaging of skins and other vertebrate material	May 2019	10.5281/zenodo.3364385
D3.4	State of the art and perspectives on mass imaging of liquid samples	June 2019	10.5281/zenodo.3469547
D3.5	State of the art and perspectives on mass imaging of pinned insects	July 2019	10.5281/zenodo.3520667
D3.6	Best practice guidelines for bulk imaging of herbarium specimens	August 2019	10.5281/zenodo.3524263
D3.7	Rapid 3D capture methods in biological collections and related fields	September 2019	10.5281/zenodo.3469531
D3.8	R&D in robotics with potential to automating handling of biological collections	January 2020	10.5281/zenodo.3719101
D4.1	Methods for automated text digitization	January 2019	10.5281/zenodo.3364502
D4.2	Data quality in transcription	January 2019	10.5281/zenodo.3364509
D4.3	Data standards in transcription	July 2019	10.1093/database/baz129
D4.4	Interoperability with institutional collection management systems	April 2019	10.5281/zenodo.3361598

Number	Title	Date	DOI
D4.5	Cost analysis of transcription methods	December 2019	10.5281/zenodo.3724327
D5.1	Recommendations for volunteer transcription systems and a source repository	April 2019	10.5281/zenodo.3552318
D5.2	Improving the detection of collection-based citizen science projects	May 2019	10.5281/zenodo.3364519
D5.3	Natural history collections and digital skills of citizens	June 2019	10.5281/zenodo.3364541
D5.4	Digitization of private collections	January 2020	10.5281/zenodo.3598303
D6.1	Data management plan of the ICEDIG project	March 2018	10.5281/zenodo.3364523
D6.2	ICEDIG digitization infrastructure design for EUDAT/CINES	February 2019	10.5281/zenodo.3364533
D6.3	Digitization infrastructure design for Zenodo	July 2019	10.5281/zenodo.3346782
D6.4	Digitization infrastructure design for national open science clouds	August 2019	10.5281/zenodo.3469490
D6.5	Open access implementation guidelines for DiSSCo	September 2019	10.5281/zenodo.3465285
D6.6	Provisional Data Management Plan for the DiSSCo infrastructure	October 2019	10.5281/zenodo.3532937
D7.1	Policy component of ICEDIG project website	July 2019	10.5281/zenodo.3366656
D7.2	Common digital research agenda	February 2020	10.5281/zenodo.3724329
D8.1	Conceptual design blueprint for the digitization infrastructure of DiSSCo	March 2020	DOI to be assigned on publication by Pensoft
D8.2	Cost Book of the digitization infrastructure of DiSSCo	March 2020	DOI to be assigned on publication by Pensoft
D9.1	Communication and dissemination plan	March 2018	10.5281/zenodo.3539164
D9.2	Linking cultural heritage of natural sciences and humanities	February 2020	10.5281/zenodo.3685634
D9.3	Stakeholder round tables	January 2020	10.5281/zenodo.3632535
D9.4	Positioning DiSSCo among other research infrastructures	January 2020	10.5281/zenodo.3724307

8.2 Other references

[Allan 2018] Allan E, Price B, Shchedrina O, Dupont S, Livermore L, Smith V (2018) A Low Cost Approach to Specimen Level Imaging of Natural History Microscope Slides Using a DSLR System. OSF Preprints. doi: 10.31219/osf.io/dvmsh.

[Allan 2019] Allan E, Livermore L, Price B, Shchedrina O, Smith V (2019) A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. Biodiversity Data Journal 7: e32342. doi: 10.3897/BDJ.7.e32342.

[Ang 2013] Ang Y, Puniamoorthy J, Pont AC, Bartak M, Blanckenhorn WU, Eberhard WG, Puniamoorthy N, Silva VC, Munari L, Meier R (2013) A plea for digital reference collections and other science-based digitization initiatives in taxonomy: Sepsidnet as exemplar. Systematic Entomology 38(3): 637-644. doi: 10.1111/syen.12015.

[Balke 2013] Balke M, Schmidt S, Hausmann A, Toussaint EF, Bergsten J, Buffington M, Häuser CL, Kroupa A, Hagedorn G, Riedel A, Polaszek A (2013) Biodiversity into your hands -- A call for a virtual global natural history 'metacollection'. Frontiers in zoology 10(1):55. doi: 10.1186/1742-9994-10-55.

[Balke 2013] Barber A, Lafferty D, Landrum LR (2013) The SALIX method: A semi-automated workflow for herbarium specimen digitization. Taxon 62(3): 581-590. doi: 10.12705/623.16.

[Beaman 2012] Beaman R, Cellinese N (2012) Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys 209: 7-17. doi: 10.3897/zookeys.209.3313.

[Bethancourt 2017] Bethencourt J, Sahai A, Waters B (2007-05) Ciphertext-policy attribute-based encryption. In 2007 IEEE symposium on security and privacy (SP'07) (pp. 321-334). IEEE. doi: 10.1109/SP.2007.11.

[Blagoderov 2012] Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, Smith VS (2012) No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133-146. doi: 10.3897/zookeys.209.3178.

[Burnap 2012] Burnap PR, Spasić I, Gray WA, Hilton JC, Rana OF, Elwyn G (2012-05) Protecting patient privacy in distributed collaborative healthcare environments by retaining access control of shared information. In 2012 International Conference on Collaboration Technologies and Systems (CTS) (pp. 490-497). IEEE. doi: 10.1109/CTS.2012.6261095.

[Chapman 2005] Principles of Data Quality, version 1.0. GBIF Secretariat, Copenhagen, 61 pp. [In English]. URL: http://www.gbif.org/document/80509. ISBN 87-92020-03-8.

[Chapman 2020] Chapman AD, Belbin L, Zermoglio PF, Wieczorek J, Morris PJ, Nicholls M, Rees ER, Veiga AK, Thompson A, Saraiva AM, James SA, Gendreau C, Benson A, Schigel D (2020) Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data. Biodiversity Information Science and Standards 4: e50889. doi: 10.3897/biss.4.50889.

[DiSSCo DS 2017] DiSSCo Design Study Report Prepared for the European Strategy Forum on Research Infrastructures, May 2017.

[DMP 2019] Provisional Data Management Plan for DiSSCo infrastructure, deliverable D6.6 version 1.0 final. ICEDIG project. doi: 10.5281/zenodo.3532937.

[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.

[Drinkwater 2014] Drinkwater RE, Cubey RWN, Haston EM (2014) The use of optical character recognition (OCR) in the digitization of herbarium specimens labels. PhytoKeys 38: 15-30. doi: 10.3897/phytokeys.38.7168.

[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/.

[EC 2013] European Commission (2013) Digital science in Horizon 2020. url: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=2124. Accessed on: 2020/02/03.

[EC 2015] European Commission (2015) Validation of the results of the public consultation on Science 2.0: Science in Transition. url: https://www.eesc.europa.eu/resources/docs/validation-of-the-results-of-the-public-consultation-on-science-20.pdf. Accessed on: 2020/02/03.

[Ford 2017] Ford N, Parsons R, Kua P (2017) Building evolutionary architectures. O'Reilly Media Inc., Sebastopol CA, USA.

[Heerlien 2013] Heerlien M, van Leusen J, Schnörr S, van Hulsen K (2013) The natural history production line. In: Digital Heritage International Congress, Oct. 28 – Nov. 1. Vol. 2, pp. 289-294. Marseille, France. doi: 10.1145/2644822.

[Hereld 2017] Hereld M, Ferrier NJ, Agarwal N, Sierwald P (2017) Designing a high-throughput pipeline for digitizing pinned insects. 2017 IEEE 13th International Conference on e-Science (e-Science), Auckland, 2017, pp. 542-550. doi: 10.1109/eScience.2017.

[ICEDIG MS44] Cocks N and al. (2019-03-31) Technical capacities of digitization centres within ICEDIG participating institutions. ICEDIG milestone M44 report, 16 p.

[ICEDIG MS45] Woodburn M, Hardy H, Smith V, Dixey K (2019-06-30) Identification of provisional Centres of Excellence for digitization. ICEDIG milestone MS45 report, 10 p.

[ICEDIG MS46] Smith V, Goodson H. (2020-01-22) International landscape analysis of related international research infrastructures, supporting collections digitization. ICEDIG milestone MS46 report, 27 p.

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

[Kalms 2012] Kalms B (2012) Digitization: A strategic approach for natural science collections. Atlas of Living Australia, CSIRO Ecosystem Sciences, Canberra, ACT, Australia. http://www.ala.org.au/wp-content/uploads/2011/10/Digitization-guide-120326.pdf.

[Krishtalka 2016] Krishtalka L, Dalcin E, Ellis S, Ganglo JC, Hosoya T, Nakae M, Owens I, Paul D, Pignal M, Thiers B (2016) Accelerating the discovery of biocollections data. GBIF Secretariat, Copenhagen. http://www.gbif.org/resource/83022.

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

[Lehtonen 2011] Lehtonen J, Heiska S, Pajari M, Tegelberg R, Saarenmaa H (2011) The process of digitizing natural history collection specimens at Digitarium. In: Jones MB, Gries C (Eds) Proceedings of the Environmental Information Management Conference 2011 (EIM 2011). September 28-29, 2011. Santa Barbara, CA. University of California, pp. 87-91. doi: 10.5060/D2NC5Z4X.

[Lendemer 2019] Lendemer J, Thiers B, Monfils AK, Zaspel J, Ellwood ER, Bentley A, LeVan K, Bates J, Jennings D, Contreras D, Lagomarsino L (2019) The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. *BioScience*, biz140. doi: 10.1093/biosci/biz140.

[Mantle 2012] Mantle BL, La Salle J, Fisher N (2012) Whole-drawer imaging for digital management and curation of a large entomological collection. ZooKeys 209: 147-163. doi: 10.3897/zookeys.209.3169.

[Mendez 2018] Mendez PK, Lee S, Venter CE (2018) Imaging natural science museum collections from the bottom up: 3D print technology facilitates imaging of fluid-stored arthropods with flatbed scanners. ZooKeys, (795), 49. doi: 10.3897/zookeys.795.28416.

[McConnell 2006] McConnell S (2006) Software estimation: demystifying the black art. Microsoft press. ISBN-13: 978-0735605350.

[McCulloch 2013] McCulloch ES (2013) Harnessing the power of big data in biological research. BioScience, 63(9):715-716. doi: 10.1525/bio.2013.63.9.4.

[Mononen 2014] Mononen T, Tegelberg R, Sääskilahti M, Huttunen MA, Tähtinen M, Saarenmaa H (2014) DigiWeb – a workflow environment for quality assurance of transcription in digitization of natural history collections. Biodiversity Informatics 9: 18-29. doi: 10.17161/bi.v9i1.4748.

[Mons 2017] Mons B, Neylon CD, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson M (2017) Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services and Use, vol. 37, no. 1, pp. 49-56, 2017. 10.3233/ISU-170824.

[Nelson 2012] Nelson G, Paul D, Riccardi G, Mast AR (2012). Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 209: 19-45. doi: 10.3897zookeys.209.3135.

[Nieva 2016] Nieva de la Hidalga A, Hardisty A, and Jones AC. (2016) SCRAM-CK: A collaborative requirements engineering process for designing a web based e-science toolkit. Requirements Eng 21(1):107–129. doi: 10.1007/s00766-014-0212-0.

[Nieva *in prep.*] Nieva de la Hidalga A, Rosin PL, Sun X, Livermore L, Durrant J, Turner J, Dillen M, Musson A, Phillips S, Raes N, Groom Q and Hardisty A. (*in prep.*) Semantic segmentation network cross-validation on natural history collection specimen images.

[Oever 2012] Oever JP, Gofferje M. (2012) 'From pilot to production': Large scale digitization project at Naturalis Biodiversity Center. ZooKeys 209: 87-92. doi: 10.3897/zookeys.209.3609.

[Price 2018] Price BW, Dupont S, Allan EL, Blagoderov V, Butcher AJ, Durrant J, Holtzhausen P, Kokkini P, Livermore L, Hardy H, Smith V (2018) ALICE: Angled Label Image Capture and Extraction for high throughput insect specimen digitization. OSFpreprints. doi: 10.31219/osf.io/s2p73.

[Rogers 2016] Rogers N (2016) Museum drawers go digital. Science 352: 762-765. doi: 10.1126/science.352.6287.762.

[De Smedt 2020] De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From data pieces to actionable knowledge units. Preprints 2020, 2020030073. doi: 10.20944/preprints202003.0073.v1.

[Suarez 2004] Suarez AV, Tsutsui ND (2004) The value of museum collections for research and society. BioScience 54(1): 66-74. doi: 10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2.

[Tegelberg 2012] Tegelberg R, Haapala J, Mononen T, Pajari M, Saarenmaa H (2012) The development of a digitizing service centre for natural history collections. ZooKeys 209: 75-86. doi: 10.3897/zookeys.209.3119.

[Tegelberg 2014] Tegelberg R, Mononen T, Saarenmaa H (2014) High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. Taxon, 63(6), pp.1307-1313. doi: 10.12705/636.13.

[Tegelberg 2017] Tegelberg R, Kahanpää J, Karppinen J, Mononen T, Wu Z, Saarenmaa H (2017-10) Mass digitization of individual pinned insects using conveyor-driven imaging. In 2017 IEEE 13th International Conference on e-Science (e-Science) (pp. 523-527). IEEE. doi: 10.1109/eScience.2017.85.

[Tulig 2102] Tulig M, Tarnowsky N, Bevans M, Kirchgessner A, Thiers BM (2012). Increasing the efficiency of digitization workflows for herbarium specimens. ZooKeys 209: 103-113. doi: 10.3897/zookeys.209.3125.

[Valan 2019] Valan M, Makonyi K, Maki A, Vondráček D, Ronquist F. (2019) Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. Systematic Biology 68(6): pp.876-895. doi: 10.1093/sysbio/syz014.

[Veiga 2017] Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, Robertson TJ. (2017) A conceptual framework for quality assessment and management of biodiversity data. PLOS ONE 12 (6): e0178731. doi: 10.1371/journal.pone.0178731.

[W3C PROV 2013] Groth P, 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/.

[Weigel 2017] Weigel T, Wittenburg P, and al. (2017) Recommendations for Implementing a Virtual Layer for Management of the Complete Life Cycle of Scientific Data. doi: 10.15497/RDA00026. *File containing long version of the recommendations: https://www.rd-alliance.org/system/files/DataFabric_SupportingOutput_recommendation-aug-2017-v10.pdf. Viewed on 6th January 2020.*

[Wilkinson 2016] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3. 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.

[Ylinampa 2019] Ylinampa T, Saarenmaa H (2019) ENTODIG-3D. Biodiversity Information Science and Standards 3: e37473. doi: 10.3897/biss.3.37473.

9 Glossary of terms and abbreviations

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

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

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

Collection Digitization 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 digitization already 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 on behalf of and for the benefit of its community of users i.e., not provided and managed by the collection-holding institution itself for itself.

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

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. The term 'content data' is sometimes also used to refer to this kind of data; for example, to avoid confusion with metadata.

Note: 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.

Digital collection: A digital representation (surrogate) corresponding to a collection of identifiable natural science specimens. Cf. digital specimen

Digital collection object type (DCO): A collection of property definitions about a digital collection, the structure of which conforms to the requirements of the openDS specification (see 4.3.3).

Digital object (DO): A bit sequence with a persistent identifier (pid) and a type.

Note: This definition is provided by the DONA Foundation (https://www.dona.net/) as custodian of the Digital Object Architecture. The long definition is: "*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]

For all practical purposes, 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 an identifiable physical specimen in a natural science collection. Cf. digital collection.

Digital specimen object type (DSO): A collection of property definitions about a digital specimen, the structure of which conforms to the requirements of the openDS specification (see 4.3.3).

Digitization: The process of converting analog information about physical specimens to digital format, which includes electronic text, images and other representations. See also mass digitization.

Digitization line/factory: The facilities (premises, personnel, equipment (hardware, software)), processes and procedures necessary for large-scale, mass digitization.

Digitization-on-demand (DoD) (also known as demand-driven digitization): An activity where selected specimens are digitized, often based on specific requests. Cf. mass digitization.

DiSSCo: See Distributed System of Scientific Collections.

DiSSCo Centre of Excellence (DCE): A designated DiSSCo Facility specialised in one or more of researching, innovating, developing and operating/performing techniques and/or processes of digitization or other related facets, and disseminating information on same.

DiSSCo Facility(ies): The geographically distributed collection-holding organisation(s) (i.e., natural science/history collection(s)) and related third-party organisations (such as DiSSCo Centre of Excellence) 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.

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 digitization of and access to the data bound up in those collections. Cf. definition of DiSSCo Facility(ies).

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 Curation and Annotation System (ECAS): 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.

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.).

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 2016].

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

Mass digitization: An activity where entire collections, or their distinct major parts are digitized from one end to the other, without selecting individual specimens. Mass digitization is characterised by technological and procedural frameworks based on automation (hardware and software) and enrichment (link-building), with workflows at industrial scale, i.e., processing millions of objects at low cost. Cf. digitization-on-demand.

Metadata: Metadata is additional data that establishes a context for the content data to which it relates i.e., it is data about data. One of the main data types managed by DiSSCo.

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 digitization.

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 digitization.

Natural Science Identifier (NSId): A kind of persistent identifier for uniquely and universally identifying digitized natural science specimens (i.e., Digital Specimens) and other associated object types.

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 metadata state information about the identified digital object. In the case of DiSSCo, Natural Science Identifiers (NSId) are the principal persistent identifiers used.

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

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.

Appendix A: FAIR Digital Object Framework (FDOF)

This appendix contains a facsimile of the text of version 1.02, November 2019 of the FAIR Digital Object Framework (FDOF) Technical Implementation Guideline (TIG). Full details and the latest version of the TIG can be obtained at http://bit.ly/FAIRDO.

A.1 FDOF Technical Implementation Guideline

"We need a set of principles that are sufficiently specific to be useful but sufficiently abstract to exclude specific software stacks, i.e., a document that will still make sense and still be useful ten years from now."

This document includes some generic guidelines to be met (section A.3), a normative part defining the FAIR Digital Object Framework (FDOF) at an abstract level which will develop over time (section A.4) and a glossary of terms (section A.5). Related documents such as implementation examples can be found at the Github site[113].

A.2 Change history

Version	Date	Intention	Actors
Version 1.0	October 2019	prepared for the consensus meetings in Washington and Paris in October 2019	created by Luiz Bonino and Peter Wittenburg
Version 1.01	17.11. 2019	created after the consensus meeting in Paris at 28/29.10.2019	changes by Luiz Bonino and Peter Wittenburg
Version 1.02	22.11. 2019	created after various comments incl. polishing and adding clarity	changes by Peter Wittenburg, Bonnie Carroll, Alex Hardisty, Mark Leggott, Carlo Zwölf

Changes from V1.0 to V1.01

- Restructuring the Document and improving some formulations.
- Leaving out concretization footnotes from the normative part.
- Leaving out footnotes about matters explained in the glossary.
- Making more statements about metadata to indicate their importance.

Changes from V1.01 to V1.02

- The illustration examples of possible implementations were separated from the FDOF core document.
- Metadata Statements were added to address the importance of metadata.
- The first editing group did improvements on the original text (polishing, clarity)
    - Throughout, tidy up of grammar and punctuation to improve clarity.
    - Definitions added in glossary for terms 'FAIR-DO', 'FAIRness' and 'semantic assertion'.
    - Generic guideline G9 on using standards added.
    - Removed reference to RDA work in G3 and introduced the idea of indicators being measurable.
    - Moved definition of referential integrity out of G4, into the glossary.
    - In G5, replaced 'layer' by 'level' and qualified the 'management level' as meaning the level of managing objects.
    - The statement saying the FDOF requirements will evolve with experience is now a note.
    - FDOF03 clarified to make clear i) that the separation of metadata into a DO different from the FAIR-DO, with it's own pointer in the structured record of the resolved PID is an optional element; and ii) that the type definition is also accessed via a PID.
    - FDOF04: made it mandatory that PID records with additional attributes beyond the standard ones must be registered in a type registry.
    - Throughout section 3, introduced the term 'PID record' as the correct term for the structured record returned by PID resolution.

[113] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF.

A.3 Generic guidelines

Some overall guidelines need to be met by the FAIR DO Framework (FDOF).

G1: Show a path for infrastructure investments for **many decades.**
G2: Demonstrate **trustworthiness** to researchers and developers to become engaged.
G3: Offer compliance with the **FAIR principles** through measurable indicators of FAIRness.
G4: Support **machine actionability,** which includes referential integrity and explicitness of semantic relationships.
G5: Support the **abstraction principle**, i.e., abstract away from the details that are not needed at a specific level. At the object management level there is no difference to be made between data, metadata, software, semantic assertions, etc.
G6: Support **stable binding** between all informational entities that are required for machines to act.
G7: Support **encapsulation,** which means that specific operations can be associated with different types of FDOs.
G8: Support **technology independence,** allowing implementations using different technologies
G9: Comply with agreed **standards** (e.g., for exchange of FDOs between systems, for interacting with FDOs, etc.) so that machine-machine interoperability can be achieved across heterogeneous systems.

A.4 Requirements for FDOF

The requirements for FDOF describe rules that must be met by all implementation of the FDO framwork.
Note: Requirements will evolve, dependent on insights obtained from implementation experience.

FDOF1: A PID, standing for a globally unique, persistent and resolvable identifier, is assumed to be the basis of the Internet of FAIR Data and Services.
FDOF2: A PID resolves to a structured record (PID record) with attributes that are semantically defined within a type ontology (which can have different forms).
FDOF3: The structured PID record includes at least a reference to the location(s) where the bit-sequences encoding the content of a FAIR-DO (FDO) and the type definition of the FDO can be accessed. The structured record may also contain a PID pointing to a metadata DO (itself an FDO) describing properties of the target FDO.
FDOF4: The PID record may include other attributes that are important to characterize specific types of FDO or that are required by applications. Additional attributes being used in PID records must be registered in a type registry.
FDOF5: Each FDO identified by a PID can be accessed or operated on using an interface protocol by specifying the PID of a registered operation and the PID of the access point.
FDOF6: This protocol offers standard Create, Read, Update, Delete (CRUD) operations on FDOs and a possibility to use extended/domain operations for specific applications.
FDOF7: The relations between FDO Types and operations are maintained in a type ontology.
FDOF8: Metadata descriptions being themselves FDOs and describing the properties of the FDO must be made available as semantic assertions, enabling machines to act.
FDOF9: Metadata assertions can be of different types such as descriptive, deep scientific, provenance, system, access permissions, transactions, etc.
FDOF10: Metadata schemas are maintained by communities of practice. FDOF requires that such metadata are FAIR.
FDOF11: A collection of FDOs is also an FDO and semantic assertions must be used to describe their construction, i.e., the relationships of their constituents.
FDOF12: Deletion of a FDO must lead to standardised and thus machine interpretable tombstone notes in metadata and PID records, i.e., PIDs, PID records and metadata should normally not be deleted, but should be modified to indicate that the FDO associated with a particular PID no longer exists.

A.5 FDOF glossary

A short glossary with explanations about crucial terms such as "repository", "encapsulation" etc. will help in clarifications, since some terms may be interpreted differently by the participants.

Term	Explanation
abstraction	Abstraction is a conceptual process where general rules and concepts are derived from the usage and classification of specific examples. literal signifiers, first principles or other methods (Wikipedia)
binding	With binding we mean the possibility for humans and machines to find other relevant entities of a DO when being exposed to another, i.e., when an actor receives a PID of a DO it must find the PID of the corresponding metadata DO and the access rights information, since otherwise interpretation and access is impossible
collection	A collection is a complex DO consisting of other DOs, which have a PID and metadata.
CRUD operations	These are the usual primary type of operations such as create, read/retrieve, update and delete
encapsulation	Encapsulation is known from abstract data types and object oriented programming where internals of data objects are hidden to the user and where the user can only influence the internal state by using defined methods
	Note: in the FDO case DO types can be associated with registered operations that can be used to operate on DO's content
FAIR Digital Object (FAIR-DO)	FAIR Digital Objects can represent data, software, protocols or other research resources. They are accompanied by persistent identifiers (PID) and metadata rich enough to enable them to be reliably found, used and cited. (FAIR Implementation Report: doi: 10.2777/1524, and Wittenburg and Strawn 2019 doi: 10.23728/b2share.2317b12321764f669c92ebbcf7518164)
FAIRness	FAIRness is a characteristic exhibited by an infrastructure component when it maintains compliance with the principles of FAIR. Achievement of FAIRness is demonstrated, for example by achieving a score (passing a threshold) in an assessment against an agreed set of maturity indicators.
machine actionability	machine actionability means the capacity of computational systems to find, access, interoperate and reuse data and services without human intervention (GOFAIR)
metadata	Metadata descriptions of DOs are sets of assertions describing properties of DOs content which are required for finding, accessing, interpreting and reusing, these assertions can cover a wide range such as descriptive to support finding, deep scientific to support science, systemic to support management, rights to prevent unauthorized access, etc.
	Note: Yet the domain of metadata is not structured very well, i.e., terminology is not well-defined.
	Note: Basic interoperability assumptions are that the schemas are registered and the concepts defined and registered.
referential integrity	The idea that all PID references must resolve and be valid without temporal limitation
semantic assertion	The attachment (perhaps by reference to a defined vocabulary) of a specific meaning to a resource, attribute, property, etc.
repository	**DO View**: from the perspective of Digital Objects repositories are nothing else than a complex DO associated with a PID, metadata of different kinds and functions to offer DOs

Term	Explanation
	Common View: from the most common point of view repositories are entities that host data, metadata etc., apply trustworthy management procedures, offer a search and access interface, have a team of experts taking care and have a sustainability plan
	Note: repositories can be associated with research organisations, communities or projects, they can be small or big in terms of the collections they hold.
type	"Type" is an attribute of digital objects which tells computational actors how the content of the DO needs to be parsed, i.e., it defines the operations that can be done on the data, the meaning of the data, and the way values of that type can be interpreted
	Note: A MIME type is a standard that indicates the nature and format of a document, file, or assortment of bytes, i.e., it is a restricted concept of type.
	Note: A type of a DO implies a summary of otherwise complex metadata assertions describing the format, encoding etc. of a content.

Appendix B: DiSSCo PID Requirements

This appendix describes the key requirements for persistently identifying Digital Specimen and other object types.

The persistent identifier (PID) system to be adopted by DiSSCo will sit alongside other identifier schemes, such as institution specific ones and CETAF Stable Identifiers[114] to identify digital collections, digital specimens, and other object classes within the seamless European virtual collections that DiSSCo aims for[115]. 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 the DiSSCo PID scheme must be to virtualise natural science collections and associated services. Thus, the principle aim in DiSSCo is that each digital object instance handled by the DiSSCo infrastructure must be universally and persistently identified by a PID that is assigned when a digital object is first created and unambiguously linked to an identifier of a physical object (i.e., to a physical specimen). Each new version of that digital object must also be identified. It should be possible to resolve an assigned PID by any available and well-known Handle service, because these are well-known for other purposes as well.

From these aims, specific requirements flow. Somewhat in priority order, these are:

1. **Brand identity and marketing**. An adopted PID scheme must be applicable across the broad natural sciences community and must not appeal to one sub-section (such as earth/geo or biodiversity) less than to any other. The brand identity must be neutral in name and applicability such that branded PIDs can be used to identify any category of (digital) specimen (i.e., plants, animals, fossils, rocks, minerals, meteorites, sediment/ice cores, etc.). DiSSCo presently prefers the 'Natural Science Identifier' brand with its acronym 'NSId'.
2. **Scope of things to be identified**. The scope of things to be identified is very broad, extending beyond persistently identifying specimens and their containers, storage and the collections to which they belong, to include assigned interpretations, annotations, records of loans and visits, as well as identifiers for temporary purposes. Any individual specimen can have multiple interpretations, annotations and records of loans and visits associated with it. The scope of things to be identified includes physical objects (but see 7 below), digital representations and concepts i.e., abstract objects or classes.
3. **Resilience.** With a need for an estimated 30 billion identifiers, the PID scheme to be adopted by DiSSCo must be resilient for 30 years or more, and potentially beyond 100 years. Once such a scheme has been superseded by another mechanism, it must be possible to continue to resolve PIDs to the specimens they refer to in the collections where they are physically located; even after re-organisations of collections and institutions have taken place. This implies the need to continue to maintain resolution services and up-to-date metadata over the very long-term.
4. **Governance and membership.** Governance arrangements must be transparent with wide stakeholder representation, allowing all collection-holding organisations a fair and equal say in how the scheme is administered and managed. A sustainable membership model must cover operating, maintenance, and improvement costs, with membership requiring active contribution and participation from collection-holding organisations towards sustaining the scheme and the infrastructure necessary for its robust operation. Note that infrastructure can be decentralised.
5. **Global uniqueness and other requirements**. Persistent identifiers for natural science and related objects must be globally unique, case insensitive, and implementation neutral (does not mandate specific protocols, such as http). They should be language and character set neutral, although for ease of adoption, implementation with a url and filename safe subset of ASCII is preferred.

[114] https://cetafidentifiers.biowikifarm.net/wiki/.

[115] Access via a 'European Collection Objects Index' (ECOI).

6. **Versioning**. Unlike physical specimens, the information content of a digital specimen changes over time. Not only is it added to, but it can be modified, leading to a new version of a digital specimen. Hence, a versioning mechanism is needed that keeps the original PID of the specimen intact and current i.e., pointing to the latest version, whilst also allowing any specific earlier version to continue to be referred to (but not altered). (Note: Zenodo archiving solves this problem by issuing two PIDs for each original deposit; one of which is used for the deposit itself and the other and subsequent ones being used for the specific versions within a deposit. Note also that, unlike Zenodo, which is a repository for archiving artefacts, digital specimens are dynamic and expected to change over time.)
7. **Identifiers for physical and digital specimens.** The identifiers of physical and digital specimens are not the same. There are many identifier schemes for physical specimens already in use and these must be accommodated by maintaining a '*linkage*' between an identifier of the digital representation of the specimen and the identifier of the physical specimen itself. This linkage must not be at the identifier level but via the metadata. Indeed, there are already persistent identifier schemes in use for local (as opposed to global) digital representations of specimens and these also must not be rendered void for resolution.
8. **Registry metadata**. Registry metadata schemas (including kernel information profiles) must be flexible enough to support the above requirements, as well as supporting:
    - Attribution (link an object to an institution or contributor working in an institution);
    - Assertion (who asserted that the metadata for the object is correct and current);
    - Provision to update/correct or annotate metadata;
    - No proprietary metadata without a mechanism for saying how it should be handled when not recognised;
    - Standard open licence for metadata; and,
    - Capability for organizations to manage their own metadata.

    The biodiversity science and geoscience domains have established vocabularies and metadata schemes that must be accommodated as the basis of registry metadata schemes. Kernel information profiles must support the kinds of pre-resolution services foreseen as necessary for the domain, including for example: finding all specimens related to specific collecting/sampling event, finding all sub-samples and preparations from any specific specimen, finding all specimens of a specific kind, etc.
9. **Length of identifiers**. There must be enough flexibility in the scheme to permit identification of all kinds of digital object associated with natural sciences (see 2 above) BUT the identifiers of the objects central to the scheme's success (i.e., digital specimen and collection objects) must be short and easily read/used/remembered. There is ample evidence that using short identifiers[116] contributes greatly to dissemination and sharing. When a talk presenter includes a short identifier in powerpoint slides, it's easy for those to be used/copied/accessed by audience participants. Similarly, in communications – emails, conference chats, twitter, etc. The DiSSCo requirement is to standardise and adopt a short-form for specimens and collections (e.g., 8 characters) but allow more-or-less anything else (typically, upto 32 characters) for other object types.
10. **Prefix form and identifying registrars**. People are familiar with DOIs for journal articles beginning "doi: 10.". The term 'doi: ten dot' is significant because it is memorable and indicative of (usually!) a journal article or dataset. Uptake of identifiers by the natural sciences community can be enhanced when they are equally memorable, short and indicative. There is thus a strong case for having a recognisable short scheme name with a 2-digit top-level prefix indicating a natural sciences object (for

[116] As in short DOIs, shortened URLs e.g., bit.ly, compact identifiers (doi: 10.1038/sdata.2018.29) or readable letter combinations like 'sofa-nice-face'.

example, 'nsid: 12.'), followed by a second-level prefix of some kind[117,118]. A key open issue is the extent to which opacity is not an important requirement; perhaps more accurately expressed as: Which parts of an identifier must remain opaque? This is still being studied.

11. **Resolution**. Robust, persistent infrastructure that allows everyone to register and resolve persistent identifiers is needed. It should be possible to resolve an assigned PID by any available and well-known Handle service, such as http://hdl.handle.net/ or https://doi.org/ because these are well-known for other purposes[119].
12. **Adherence to FAIR metrics**. Natural science objects are expected to remain findable, accessible, interoperable and reusable (i.e., 'FAIR') throughout their lifetime. To help achieve this, the adopted PID scheme must adhere to FAIR metrics[120]; specifically, machine-readable metadata (FM-F2, FM-F3), described identifier management (FM-F1B), metadata longevity (FM-A2), public registration of the identifier scheme (FM-F1A), and provenance specification (FM-R1.2).

END.

[117] Alternatives for second-level prefixes must be further studied. Here are two possibilities: i) In many cases, the registrar organisation will be the collection-holding institution (e.g., Natural History Museum London). So, using the Global Registry of Scientific Collections (https://www.gbif.org/grscicoll), a second-level prefix could be "NHMUK", yielding (for example) 'nsid: 12.nhmuk/'; and ii) Rather than registrar institution, a further classification of the digital specimen type is reflected at the second-level of the prefix. For scientists, being able to quickly understand the (taxonomic) group of specimens e.g. 12.B/ (for algae, fungi and plants) or 12.Z/ (for zoology) can be helpful.

[118] Careful design of the NSId prefix scheme might provide the opportunity to include the current syntax of IGSNs, retaining backwards compatibility whilst moving them into a new top-level handle prefix. We understand that today IGSNs syntax mandates a concatenation of the namespace of the allocating agent and a unique in that agent identifier e.g. for NHM that would be NHMXXXXXXX. Retaining that syntax we can then have a nsid: 12.igsn/NHMXXXXXXX. IGSN retains its identity but within a larger universe of collections/samples handles.

[119] We note, however the DOI Foundation strategy discussion, commenced earlier this year to terminate resolution of non-DOI handles by the doi.org resolver. We think this would be a retrograde step.

[120] doi: 10.1038/sdata.2018.118