Digitisation of natural history collections: criteria for prioritisation

Louise Isager Ahl[‡], Luca Bellucci[§], Philippa Brewer[‡], Pierre-Yves Gagnier[|], Helen M. Hardy[¶], Elspeth Margaret Haston[#], Laurence Livermore[¤], Sofie De Smedt[«], Helen M. Hardy[¶], Henrik Enghoff[‡]

‡ Natural History Museum of Denmark, Copenhagen, Denmark
§ Museo di Geologia e Paleontologia - Università di Firenze, Firenze, Italy
| Muséum national d'histoire naturelle, Paris, France
¶ Natural History Museum, London, United Kingdom
# Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom
¤ The Natural History Museum, London, United Kingdom
« Botanic Garden Meise, Meise, Belgium

Corresponding author: Louise Isager Ahl (louise.ahl@snm.ku.dk)

Abstract

There are approximately 1.5 billion specimens kept in European Natural History Collections. The mission for the Distributed System of Scientific Collections (DiSSCo) is to unite all these specimens into a one-stop e-science infrastructure of digital specimens. This is a monumental digitisation task and criteria for how to prioritise this effort are, therefore, crucial for the success of the project. In this report, we have reviewed the literature and designed and conducted surveys of the digitisation plans and criteria used by DiSSCo Partners to understand the prioritisation criteria used in the digitisation of natural history collections. As an attempt to provide some guidance for the digitisation of specimens, we suggest that an organisation (e.g. DiSSCo or an individual institution) that is planning to digitise natural history collections considers four categories of prioritisation criteria: Relevance, Data quality, Cost and Feasibility.

Keywords

DiSSCo Prepare, DiSSCo, natural history collections, natural science collections, digitisation, prioritisation, digitization, prioritization

Executive Summary

A core mission of the Distributed System of Scientific Collections (dissco.eu) is to unite the ~ 1.5 billion specimens kept in European Natural History Collections into a one-stop e-science infrastructure containing as many of these specimens as possible in the form of

digital specimens (Hardisty 2019). To achieve this, a massive digitisation effort is required and, to guide this effort, criteria for how to prioritise are needed.

This issue has been addressed in several previous publications, notably in a report from a GBIF taskforce (Krishtalka et al. 2016) and in a very comprehensive treatment resulting from the DiSSCo-related project ICEDIG*[1] (Bakker et al. 2018). We have reviewed these reports and conducted additional literature reviews in order to find any relevant publications post-dating Krishtalka et al. (2016) and Bakker et al. (2018). To address the issue in a different way, we surveyed DiSSCo partners, asking for their digitisation plans and for the criteria they have been using to prioritise digitisation of their own collections. We also obtained information on the actual cost of digitisation projects, striving to include all costs associated with such projects, something that is lacking from available publications on this subject.

The general picture emerging from previous studies (Krishtalka et al. 2016, Bakker et al. 2018) is that scientific and research relevance is rated as the most important criterion, but apart from that, the signal is unclear. Relevance in relation to management and stewardship of collections themselves, as well as funding opportunities, are acknowledged as important criteria, whereas societal relevance*[2] is regarded as a less important criterion. As an attempt to provide some guidance through the complex landscape of prioritisation criteria, we suggest that an organisation (e.g. DiSSCo or an individual institution) that is planning to digitise natural history collections considers four broad groups of criteria:

1. Relevance;
2. Data quality;
3. Cost;
4. Feasibility.

The four groups embrace all prioritisation criteria which have been previously proposed and are described in detail in this report.

Data quality is given particular attention since this aspect of digitisation has been somewhat neglected in previous works. We have split this criterion into two main components:

> **1) How much information is there in each digital specimen?** (Information level). This component has been addressed through the development of the MIDS concept (Minimum Information about a Digital Specimen, Haston et al. (2022));
>
> **2) How reliable is that information?** Reliability includes accuracy (the closeness of measured values, observations or estimates to the true value) and precision (e.g. of geographical information: latitude/longitude in degrees only, in degrees plus minutes or in degrees plus minutes plus seconds or of taxonomic information: identification to genus, species or subspecies level).

The quality of data also includes the potential for quality assessment and improvement, as well as its completeness in terms of taxonomic, geographical or collection coverage.

Cost is obviously a major consideration in any digitisation project. We emphasise that cost estimates should include all costs associated with the project, including pre-digitisation, digitisation *sensu strictu* and post-digitisation) as highlighted in two case-studies in which we have analysed all costs associated with the digitisation of a herbarium and a collection of fossils. Cost in relation to prioritisation includes both affordability (can the project be achieved within the resources available and in relation to any funding opportunities?) and value for money - whether the costs are reasonable in relation to the intended benefit or impact.

It has become obvious that there is no easy way to implement the multitude of criteria. The idea of an algorithm such as a “decision tree” seems unviable and we suggest that projects be evaluated/prioritised by a combination of a scoring method and a panel discussion, similar to what has been done in the series of SYNTHESYS projects*[3].

We strongly recommend collaboration, for example, at DiSSCo level, in order to optimise resources and we want to underline that, irrespective of which criteria are considered, there is no fit-all solution. Flexibility is essential, depending on the intended use of the digital specimens to be generated; the resources available; and in order to respond to opportunities.

We provide a list of questions to be considered in connection with the drafting or evaluation of digitisation projects.

Finally, we stress that digital specimens can never replace the physical specimens that exist in collections and that ensuring the long-term preservation of the collections remains a top priority.

Project context

This project report was written as a formal Deliverable (D1.3) of the DiSSCo Prepare Project (Koureas et al. 2023) and was previously made available to project partners and submitted to the European Commission as a report. While the differences between these versions are minor, the authors consider this the definitive version of the report.

The following text is the formal task description (Task 1.3) from the DiSSCo Prepare project's Description of the Action (workplan):

"Based on the analysis of previous studies, relevant criteria will be identified and developed into a basic model for the prioritisation of digitisation of objects held in NSCs. Criteria to be considered include scientific relevance, user needs, socioeconomic impact, specialisation, technical feasibility and cost".

Background

Natural history collections are treasure troves for scientists and, in order to safeguard and expand the use of these collections for the future, digitisation is pivotal. Attempts to digitise natural history collections throughout the world have already started. The Distributed System of Scientific Collections (DiSSCo) is a pan-European Research Infrastructure (RI) for natural science collections. The aim of this infrastructure initiative is to unify all European natural science assets under common access, curation, policies and practices. This approach and set-up will ensure that all the data is easily Findable, Accessible, Interoperable and Reusable (FAIR principles - see also Wilkinson et al. (2016)).

Digitisation in this context spans the spectrum from making basic information on a specimen (name, collecting locality etc.) digitally available, to including (or linking to) digital images (photographs, X-rays, scanning electron micrographs etc.), DNA sequences, chemical information and other data in the digitised information. These rich, linked specimen data have been referred to as the "Extended Specimen" (Lendemer et al. 2019) or the "open Digital Specimen" (Addink and Hardisty 2020).

Digitisation can be approached in different ways:

- **Mass digitisation** – large digitisation projects like the digitisation of an entire collection (usually of thousands up to hundreds of thousands of items)*[4];
- **Project-driven digitisation** – smaller defined projects focusing on particular specimens like those collected on a specific expedition or for a particular purpose;
- **Digitisation on demand** – digitisation of a limited number of specimens for a particular scientific study or project by external researchers, who approach the collection-holding institution;
- **Business-As-Usual** (BAU) digitisation – digitisation made in connection with everyday curation, for example, digitisation of specimens going out on loan, coming back from a loan or selected for an exhibition.

In Europe alone, there are an estimated 1.5 billion specimens stored in collections, representing nearly 80% of described species worldwide (Bakker et al. 2018). Today, more than 39 million specimen-related records have been uploaded by the DiSSCo network to GBIF*[5]. These specimens have become **digital specimens**, which means they are closer to the FAIR guiding principles. The DiSSCo RI (dissco.eu), which as of May 2023 has completed its Preparatory Phase and entering a Transitory Phase, aims to produce digitised specimens in a FAIR framework on a large scale.

Within institutions, prioritisation may need to take into account all of the four categories above, in a ‘balanced portfolio’ approach that, for instance, ensures mass digitisation projects are balanced against user-led services and the need for innovation or more bespoke pilots or the need to make equipment available for business as usual. For

DiSSCo, prioritisation of what to digitise is perhaps most critical in relation to the coordination of mass digitisation programmes and/or larger project-based digitisation, as these will primarily drive critical mass of content creation through the DiSSCo infrastructure. It is also likely that central coordination of on-demand approaches may be required; however, this is less a question of prioritisation - which, by definition, is user-led in these services - and more one of service design, funding etc. Mass or larger project digitisation activities are, therefore, the main (but not only) focus of this report. Technical approaches to digitisation are a related and overlapping subject, but this will not explicitly be dealt with here unless it is of direct relevance to the discussion.

The crucial question can briefly be framed as "*Where to start?*". Another crucial consideration is: *"to what extent should decisions be made at a European or global level, rather than in individual collection-holding institutions"?* A coordinated approach would allow us to focus more efficiently on solving specific problems that have a wide and significant impact on all of us, for example, by assembling critical mass of relevant data to address key societal challenges; or by enabling the most efficient and effective workflows to be deployed widely with maximum impact. Here, DiSSCo offers a unique opportunity for coordinating prioritisation, though it should also be recognised that each institution will have their own drivers and stakeholder requirements that will impact the prioritisation process (not least in that different institutions hold different types of collections and objects, which they will naturally see as their priorities).

Methodology

There are few descriptions and models available for prioritisation of digitisation targeting natural history collections. Many potential factors may influence the decision-making process regarding prioritisation and the present paper is to be seen as a help to " *establish relevant criteria to identify a prioritisation model for digitisation"* (DPP Description of Work). To obtain a better understanding of what has been done in the past and what is included in current digitisation programmes, we carried out the following:

- Performed a comprehensive review of the literature;
- Designed and conducted surveys of digitisation plans and criteria employed amongst all DiSSCo partners.

Additionally, we obtained detailed information of all costs associated with two digitisation projects that have been carried out in recent years.

Search for additional studies on digitisation criteria

At the onset of this project, two core studies were available on the topic of digitisation. The most recent work was carried out in the ICEDIG project and reported in the final deliverable "Inventory of criteria for prioritisation of digitisation of natural history collections" (Bakker et al. 2018). This work complemented the study by Krishtalka et al. (2016) on how to accelerate the discovery of biocollections data. The most important

points made in these studies have been summarised in Suppl. material 1 and they were the inspiration for our literature investigations. Two literature reviews were carried out, the first in 2021 and the second in 2022. Based on the results, a corpus of previous studies on prioritisation of digitisation was compiled, covering the period from 2018 until June 2022. The list of relevant references found during the 2021 survey was included in a previous report (Suppl. material 2) and those found during the 2022 survey are listed in Suppl. material 3.

For the 2021 survey, works deemed to be relevant were scored (1-3), based on relevance for the investigation with 1 being most relevant. The searches were carried out in Google Scholar with the following search parameters:

1. Search: ”natural history collections” ”prioritisation” since 2017;
2. Search: ”natural history collections” ”digitisation” since 2017;
3. Search: ”digitisation” ”prioritisation” since 2017;
4. Search: ”natural history collections” ”digitisation” ”prioritisation” since 2017.

In comparison to the results presented by Bakker et al. (2018), a total of 12 new publications deemed to be relevant were identified from the four search compilations (April 2021). In the additional analysis carried out in June 2022, a total of 14 new publications deemed to be relevant were identified from the four search compilations (see Table 1).

The 2022 survey was carried out under much broader criteria and resulted in a large number of publications (see Suppl. material 3).

Surveys

In addition to the literature study, two surveys were carried out amongst DiSSCo partners: one coverng their digitisation strategy if present and one covering the prioritisation criteria they used for digitisation completed or in progress.

Survey 1 - Essay-based questionnaire

DiSSCo partners were asked to provide information, in free text and preferably no more than 2 A4 pages, on:

1. Their digitisation strategy (if available, they were asked to provide a copy or link);
2. The prioritisation criteria employed for digitisation which has already been done or is in progress in the institution.

The following guiding questions were supplied to highlight relevant topics:

- Do you have a clear overview of the digitisation status of your institution (how many specimens databased, how many imaged, by which procedural standard etc.)?
- Are you monitoring it? How?

- What is your digitisation level: specimen level or higher collection unit level? What are your policies with respect to how much data are acquired (databasing/ transcription of specimen information and/or imaging)?
- Do you have a unique management software or more than one? What kind of protocol are you using for the data digitisation (e.g. ICEDIG guidelines)?
- Do you have a procedure for validating data (e.g. accuracy of identification and georeferenced)?
- What are you planning to digitise next and what projects are planned for further down the line and why?
- If you do not have a defined plan, what are the circumstances driving you to unplanned digitisation actions (e.g. specimens requested for loan, new accessions, specimens involved in an exhibition etc.)?

It was suggested that, in their answers, it could be useful to distinguish between:

- Mass digitisation or large scale where indeed the questions of prioritisation, feasibility etc. are very relevant;
- Digitisation on demand;
- Opportunistic digitisation;

This study was carried out in the autumn and early winter of 2021.

In Suppl. material 2, a list of all countries in DiSSCo and the institutions from each country that have replied to our questionnaire has been compiled. Institutions marked with (*) were partners in the task group for this report. A complete compilation of replies was submitted in a previous report "*Corpus of previous studies on prioritisation of digitisation compiled*" which has been included in Suppl. material 2.

Survey 2 – Multiple-choice questionnaire

The multitude of thoughts, approaches and results described by respondents to the essay-based questionnaire makes interesting reading although, as expected, the format makes it difficult to quantify or even to describe the results in a few paragraphs or diagrams. Therefore, we subsequently developed a short multiple-choice questionnaire focused on the digitisation activity, using a Google Form. The short questionnaire, after being reviewed by the task partners, was sent to all DiSSCo National Nodes who shared it with their own institutions in order to collect information from as many institutions as possible involved in DiSSCo. To facilitate the dissemination, the questionnaire was translated into different languages (English, Danish, French, Italian and Dutch). An overview of the questions and answers can be found in Suppl. material 4.

The structure of the questionnaire was as follows:

- Q1 – Q3: compiler's information (personal details, e-mail, role, country, institution);
- Q4 – Q5: information about collections (size and staff employed);
- Q6 – Q9: information about digitisation strategy (digitisation initiative, digitisation priorities classified in five main categories, Scientific Relevance, Institutional

Relevance, Economic Relevance, Educational Relevance, Technical feasibility and subcategories for each one of them);

- Q10 – Q12: information about the management of collections (overview and monitoring of the digitisation status, use of CMS-Collection management system);
- Q13 – Q16: information about digitised items (procedure for data validation, standards used for databasing, digitisation levels for databased items, images and 3D models);
- Q17: further remarks about digitisation strategy;

This study was carried out in spring of 2022.

Case studies on cost

In addition to the prior costbook work (Hardisty et al. 2020) and transcription cost work (Walton et al. 2020a) in the ICEDIG project, we asked all partners in the task for detailed and complete information on digitisation costs. Such information was not readily available for most projects, but we present two detailed case studies obtained from NHM D and UniFi. Both projects were externally funded and prioritised because there was internal research relevance (e.g. staff undertaking active research on the collections) and because they were considered relevant and impactful to the external funders.

Results

Literature review

The most significant results obtained through the literature review were reports carried out by GBIF (2016) and within the DiSSCo-related project ICEDIG*[1]. These two publications will, therefore, be summarised here (extended summary in Suppl. material 1); the additional relevant publications are listed in Suppl. material 3.

GBIF report

A task force was convened by GBIF “to help accelerate the discovery, digitisation and access to biocollections data”. One of the task force’s main objectives was to provide guidance on establishing priorities for digitising biocollections to serve institutional, national, and global needs and achieve the greatest economies of scale (Krishtalka et al. 2016). The GBIF task force undertook a large-scale, global survey amongst collection-holding institutions on the state and prioritisation of digitisation. A total of 519 respondents gave information on their priorities and these are presented in Fig. 1.

The most important priorities identified by the GBIF task force were reported to be:

1. Research;
2. Funding/grant opportunities;

3. Taxonomic priorities.

However, these findings are only in part compatible with the most important criteria found by ICEDIG (see below).

ICEDIG Report

ICEDIG was an EC-funded project under the Horizon 2020 Framework*[1]. In the report "Inventory of criteria for prioritization of digitization of Natural History Collections" (Bakker et al. 2018), a corpus and analysis of digitisation criteria was presented. It forms a very substantial part of the basis for the present report. The aim of the ICEDIG deliverable was to contribute to an "*easy and well-informed decision-making process in relation to prioritisation of digitisation of natural history collections*". In ICEDIG, it was decided to follow a multi-stage process to ensure that the solutions put forward were solid regarding the prioritisation of digitisation of natural history collections. Stages identified:

1. **A literature and reports inventory** was carried out to create an overview of the criteria of prioritisation of digitisation;
2. **Targeted survey.**

For the questions regarding prioritisation, Bakker et al. (2018) obtained 68 completed responses that were included in the depictions of the data shown in Fig. 1 and included in Suppl. material 1. Fig. 2 gives the overview of the ranking of the four areas of relevance identified: scientific, collection, social and economic. Included in Suppl. material 1 are figures (S1-S4) that show the ranking of the criteria used in the questionnaire identified for each of the four areas.

Based on the additional information added in free text, an extensive and revised list of criteria was assembled on six overarching topics:

1. Collection relevance;
2. Economic relevance;
3. Funding;
4. Practical criteria;
5. Scientific relevance;
6. Social relevance.

We note that there is some overlap between all of these topics.

Due to the broad range of criteria that were identified to be of importance in the process of prioritising digitisation efforts, three possible methods to determine the strategy for a digitisation project were proposed: 1) Decision tree; 2) Scoring method and 3) Panel review.

Although relevant publications were identified through the additional literature survey (Suppl. material 3), they did not add anything substantial that had not already been covered by Krishtalka et al. (2016) and Bakker et al. (2018).

Surveys

Two surveys were carried amongst DiSSCo partners on their digitisation strategy (if existing), as well as on which prioritisation criteria they employed for digitisation which had already been done or was in progress. The main findings have been summarised here and the complete responses can be found in Suppl. material 2 and Suppl. material 4 .

Survey 1 – Essay-based questionnaire

The natural history collections that replied to our questions are at different levels in their digitisation efforts. This means that the answers reflect whatever level they are at and are, therefore, hard to sum up in a coherent way as they varied from "*all our collections have been digitised*" to "*we have no official document outlining our digitisation priorities*". However, most seem to adhere to the criteria put forward by Bakker et al. (2018) by starting their digitisation process by capturing the data of their most important specimens (types, historic, fragile, cultural). Another strong driver of the collective digitisation efforts by DiSSCo members has been the opportunistic approach, i.e., a broad span of research and funding opportunities has determined the priorities. Finally, a lot of members are actively trying to digitise all new incoming specimens to some degree. Survey 1 was summarised and presented in a report included here as Suppl. material 2.

In terms of prioritisation criteria employed for digitisation efforts, many respondents had left this blank or indicated that internal work was in progress to define their approach. It is, therefore, not possible to extract general tendencies. Instead, we present, as a concrete example, the key criteria for digitisation efforts employed by the Natural History Museum of Denmark:

- National collection strength;
- Research and public relevance;
- Digitisation cost and volume;
- Established international policies and archival formats.

Survey 2 – Multiple-choice questionnaire

Of the 23 national nodes, only 10 answered, with a total of 79 answers. Most of the answers came from NH Museums or University Museums and Research Institutions. Thus, most respondents are curators, several are researchers or directors of the collections and a few are digital collection managers or similar (Suppl. material 4, Q1-3). Of the 79 institutions that replied to the questionnaire, 28 have a well-defined digitisation strategy (20 with small collections, four medium-size, two large and two very large collections), 13 were uncertain about this, but most (37) do not have any digitisation strategy.

In general, the size of team is proportional to the size of collection with some a few exceptions: five large or very large collections have a small team, six small collections have medium-sized teams and one very small collection has a large team (Suppl. material 4, Q4-5).

Digitisation seems to be primarily driven by "Projects (e.g. E-Recolnat, national lists of flora or fauna etc.)" and "Opportunistic digitisation (e.g. moving the collection into a new site, out-going loans, new specimens entering the collection, exhibition and other contingent events)". The "Digitisation on demand (i.e. *ad hoc* digitisation for specific research, as requested by external researchers, for example, through VA SYNTHESYS+)" is the third choice in the decision process described by Hardy et al. (2020). In any case, mass digitisation still occupies a small part in the digitisation activity and digitisation mainly by manual data entry is most frequent (Suppl. material 4, Q6-8). Amongst the few institutions that mainly applied mass digitisation (50-75%, up to 90% of the digitisation activity), three own very large or large collections, one holds a medium-size and one a small collection.

The short questionnaire highlighted that almost all the institutions share the same digitisation priorities as follows (see Suppl. material 4, Q.9-9e):

1. **Scientific relevance:**
    1. Focusing on taxonomic targets;
    2. Geographic targets;
    3. Museological targets;
    4. Global challenges activities.
2. **Institutional relevance:**
    1. Importance for the museum itself*[6];
    2. Strategic for national and/or regional programmes / projects / guidelines*[6].
3. **Educational relevance:**
    1. Education and training young people;
    2. Citizen-science initiatives;
    3. Other public engagement.
4. **Technical feasibility:***[7]
    1. Ease in specimens handling;
    2. Remote digitisation (e.g. from paper catalogues);
    3. Availability of dedicated technologies (e.g. conveyor belt for herbaria and pinned insects).
5. **Economic Relevance:***[7]
    1. Overall performance in respect to human resources and tools;
    2. Overall performance in respect to financial resources;
    3. Faster digitisation improving cost/volume rate.

Therefore, the "Scientific relevance" of a collection is the key element that drives digitisation, the taxonomic and the geographic relevance are the most important sub-

criteria in this category; if the collection has an institutional importance (maybe for funding programmes), the priority for its digitisation is boosted.

A total of 70% of the respondents declared that their institution has a clear overview of the digitisation status (how many specimens are in the database, how many imaged, open access database etc.), but for most, the database is not in open access. The digitisation status is monitored by automated means in less than 20%, while the remaining 80% are divided between “no monitoring in place” or “monitoring by extracting the needed information through different databases or sources”. A single CMS is used by a small percentage (28%), whereas 50% do not have a CMS, but use traditional databases (e.g. Access, Excel files) (Suppl. material 4, Q10-12). This result suggests that, even if it is more appropriate to have a single CMS to better manage all the collections, it is still very difficult to apply a unique CMS for different types of collections, from the geological to the biological ones.

Regarding information about digitised items (Suppl. material 4, Q 13-16), 70% of compilers answered that data are validated by the curator and/or by other specialists; of these, 50% answered that data are only partially validated, while the remaining 20% are totally validated. It is interesting that 23% declared they do not have a validation procedure in place. There are clearly needs and opportunities for creating more links amongst institutions to share expertise in data validation. As regards Minimum Information about Digital Specimen, four levels were defined in the questionnaire*[8]:

1. **MIDS0** - Bare: name + unique identifiers (inventory number);
2. **MIDS1** - Basic: MIDS0 + higher taxonomy (to family level) + higher geography (to country level);
3. **MIDS2** - Complete: MIDS1 + label information (collection locality, collector, date);
4. **MIDS3** - Integrated: MIDS2 + external data, not directly available from labels (e.g. bibliography).

The answers showed that MIDS3 level has the lowest percentage for almost all the collections (n = 41); while MIDS2 is the best «compromise» since it provides considerable information, while not being too demanding. The expected decreasing trend from MIDS0 to MIDS3 was not clear in the replies, probably because some respondents did not answer by following the suggested logic “MIDS0 ≥ MIDS1 ≥ MIDS2 ≥ MIDS3” in the question; observing the single answers, they probably reported the values by subtracting the number of digitised specimens at one level from the total digitised. There is a low percentage of imaged items and 3D models, this probably being due to lack of specific tools/technologies and a larger repository for data.

Finally, the replies have highlighted how funding, particularly for employed dedicated staff, is crucial for planning a digitisation strategy.

The multiple-choice questionnaire can be found in Suppl. material 4.

Case studies on cost

Cost is an important consideration in any digitisation project, it often constitutes a criterion overruling other considerations, either because projects are not considered to be affordable (they cannot be achieved within available resources) or, perhaps, because the value for money of pursuing them is not considered sufficient. We found that most of the published cost analyses of digitisation, including the in-depth analysis made in the context of the ICEDIG project (Hardisty et al. 2020) did not comprehensively consider all the costs involved in pre-digitisation, digitisation *sensu strictu* and post-digitisation. In the two examples summarised below, we have tried to include all stages in the process, from the moment a sample has left the cabinet until it has been safely returned. Perhaps the most important function of the examples is to serve as a checklist of cost items to keep in mind. See also the list of questions to be considered in the Conclusion and Recommendation section.

Costs associated with the digitisation of the Greenland Herbarium at the Natural History Museum of Denmark

This mass-digitisation project at the Natural History Museum of Denmark (NHMD) was initiated in 2019 and was completed in May 2023. The project was partly financed by a grant (2.2 million DKK ~ 295,000 euro) from the Aage V. Jensen Charity Foundation and NHMD invested considerable additional resources from its internal collection budget.

The aim of the project was to digitise the Greenlandic vascular plant herbarium, including transcription and georeferencing. The collection is significant as it is the large collection of plants from Greenland and includes a significant proportion of historical material. The project is summarised in more detailed by Iwanycki Ahlstrand (2023).

Table 2 presents an overview of the various expenses and Table 3 gives a detailed example of the data-cleaning process. This was data which were recorded part-way through the project in August 2022.

Costs associated with the 3D digitisation of the fossil holotypes housed at the Museum of Geology and Paleontology of the University of Florence (Italy)

This 3D digitisation was initiated in 2020 and finished in 2022 thanks to Tuscany Region Postdoc Grants in Cultural Heritage 2018 (“POR FSE 2014-2020 Asse A – Occupazione”). This project entitled “Virtual paleontology - a non-invasive approach for the fruition, diffusion and sharing of the paleontological heritage” (PalVirt) was carried out by Dr. Saverio Bartolini Lucenti and was the first example in Italy of the systematic and massive 3D digitisation of paleontological type-specimens, in particular 138 vertebrates (almost all) and 69 invertebrates and plants. Three partners were involved in the project: the Earth Science Dept. – Paleo[Fab]Lab, the Geology and Paleontology Museum and

Tbnet Soluzioni3d srl (Arezzo). For further information, see Bellucci et al. 2023. Table 4 presents an overview of the various expenses.

Discussion

Introduction

The results from both the essay-based and the multiple-choice questionnaire, like the results from the literature studies, highlighted the extreme complexity of prioritisation. Fulfilling the ambition of DiSSCo, to digitise millions of specimens in all possible shapes, sizes, origins, ages, state and value, is indeed a daunting task. The very high number of prioritisation criteria that have been suggested may appear as a barrier to progress for many institutions or may need to be balanced at an organisational level for example, to meet strategic or funding opportunities, while also carrying out projects to develop new digitisation workflows or to meet the needs of particular users. An organisation planning a digitisation project needs to consider whether, for example, scientific relevance should be a guiding principle (and define what this means in their specific case) and/or what the funding opportunities are and/or what data quality can be obtained with the resources at hand and/or what the societal interest in the digital specimens to be created is.

With the aim to facilitate decisions about prioritisation of digitisation to be taken by DiSSCo or by individual institutions, we here offer a classification of the multitude of possible criteria into four main categories. Based on our literature study and the results of our surveys, we propose the following four categories:

- Relevance;
- Data quality;
- Cost;
- Feasibility.

All criteria that have been suggested previously fall into one (or more) of the four groups which are, thus, not new criteria, but are meant as an aid to reduce the multi-dimensionality of the “criterion space” during the first steps in the prioritisation process.

The categories of criteria are not completely mutually exclusive. For example, “Cost” may be seen as a component of “Feasibility” an indeed, cost considerations often overrule other criteria. In spite of the somewhat simplistic classification of prioritisation criteria presented above, prioritisation remains a very complex task. It is important to bear in mind that considering just one criterion or just one category of criteria in isolation, will not result in a sound prioritisation. All categories need to be considered, as visualised in Fig. 3. It is also worth remembering that prioritisation is not an exact science, nor is prioritisation constant, but may vary over time, for example, as policies or funding opportunities change.

Relevance

Relevance may be seen as the primary criterion for prioritising digitisation. If the digitised specimens to be generated are of low relevance, i.e., will lead to no benefit or have no impact, other types of criteria (data quality, cost, feasibility) become almost irrelevant.

Different kinds of users have different needs: what is seen as most relevant for one may not be most relevant for another. According to the comprehensive ICEDIG study (Bakker et al. 2018), scientific relevance is deemed most important, at least amongst the respondents to the ICEDIG's survey, but collection relevance is also important, whereas social and economic relevance are less so. However, depending on the nature of the specimens to be digitised, on the funding possibilities etc., none of these categories of relevance can be neglected - and they are likely to overlap, if, for instance, scientific relevance is in a discipline which addresses societal and economic challenges, such as biodiversity loss. Concerning social and societal relevance, see the report by Figueira et al. (2023), as well as the "Discussion and outlook" chapter in Fitzgerald et al. (2021) and von Mering et al. (2021). The GBIF study (Krishtalka et al. 2016) agreed with ICEDIG in finding research most important, but disagreed in finding funding/grant opportunities,and taxonomic priorities second and third. Even "scientific relevance" is a complex concept. See Table 5 for an attempt to visualise the different needs of different scientific disciplines.

There are two further complexities in relation to using scientific relevance as a guide to prioritisation in DiSSCo. Firstly, it is likely that almost all collection objects where sufficient data are present have scientific relevance against one or more of the types of research mentioned above. Deciding which of these purposes are 'most' important or relevant is extremely challenging. Secondly, this relies on our current understanding of what is important, relevant and useful - but a key benefit sought through digitisation is to unlock new avenues and paradigms of research, for example joining up collections data to other data sources in ways which have not previously been explored. Again, this makes judgements of scientific relevance, based on today's evidence inherently flawed, although still worthwhile as one of the criteria to provide information on prioritisation. Irrespective of how carefully relevance criteria are analysed, nothing is immutable. Like prioritisation in general, scientific relevance may change over time as institutions and researchers change their focus.

Much of the existing research prioritisation focuses on scientific research. The low prioritisation of 'social-relevant criteria' or social relevance (Bakker et al. 2018) may seem surprising, but are at least, in part, the result of limitations in the current scale and scope of digitised material and existing patterns of usage. Bakker et al. (2020) describe the change in use of natural history collections from their original taxonomic focus to a much broader, interdisciplinary use, including climate change, human health and food security. Recent work by Popov et al. (2021) and Hardy et al. (2023) reports on the financial benefits of digitising collections and the growing demand for socially-relevant data with cross-domain approaches, such as using computer vision on natural history collections

for climate change research (Wilson et al. 2022) and supporting conservation assessments for wild relatives of important agricultural crops (Khoury et al. 2020).

Data quality

As a thought experiment, consider two digitised collection: one with 100,000 digitised specimens and a second with 1,000,000 digitised specimens. At first glance we might consider the latter more advanced in terms of quantity of digital specimens. However, what is the quality of the digital specimens in the two collections? When planning and assessing digitisation, data quality needs to be taken into consideration although this aspect has not been very much considered in previous studies. See Chapman (2005a) for a thorough treatment of the data quality concept.

There are two main dimensions of data quality:

- How much information is there in each digital specimen (Information level)?
- How reliable is that information?

A third essential aspect of data quality is potential for validation and improvement:

- How can we know how reliable is our data and how can we improve it?

Discussion of data quality is also not independent of the relevance criteria discussed above - the reason data quality is important has to do with whether data are 'research-ready' and impactful. There may be areas of data quality, such as high quality geo-referencing, that are relevant to widespread fields of research; but other areas of detail which are critical for particular studies, but less valuable to widespread users. It is also often the case that a few key data fields from a large volume of specimens may be more valuable than deep and detailed data on just a handful of objects - again, it depends on the potential uses and users. Ultimately, however, it is reasonable to say that if data about specimens are clearly poor or lacking (e.g. labels are missing, damaged etc.), those specimens are unlikely to achieve much impact through digitisation. These points are explored further below.

Information level

A digitised specimen may be anything from a textual record with minimal information (e.g. species name) to an extended digital specimen represented by full collection information, illustrations in the form of photos and CT scans, morphometric data, DNA sequences, sound recordings, chemical profiles and with links to related data and resources.

In order to quantify the information level of digital specimens, a digitisation standard has been developed. The Minimum Information about a Digital Specimen (MIDS) standard (Hardisty and Haston 2021) comprises three main levels of digitisation plus an initial 'pre-digitisation' level. These levels provide a framework for prioritising, planning, costing and monitoring a digitisation programme for collections. Using the MIDS standard, the

digitisation level of a collection can be scored and changes can be tracked. The four MIDS levels are shown in Table 6.

The level of information required varies significantly depending on what the data are being used for. Planning and costing a digitisation programme potentially requires a low level of information; some ‘big data’ analyses, including species distributions, require an additional set of data; whilst taxonomic research may require all the data that are available on the specimen. Mass digitisation programmes are commonly taking a staged approach to capturing information, starting at the basic level (MIDS, Level 1) and using a range of options, including outsourcing and crowdsourcing, to transcribe additional data and reach a higher digitisation level. The extended record (MIDS, Level 3) equates to the DiSSCo open Digital Specimen specification (Hardisty and Haston 2021).

An example of a digitised specimen with a very high information level can be considered a digital surrogate. This concept was described by Godfray (2007) for the digitisation of type specimens made available online. Considering that requests for access to type specimens constitute a significant fraction of requests for access to natural history specimens, these digital surrogates may save travel and shipment expenses, as well as time. For example, Akkari et al. (2015) described a new species of millipede and in addition to the physical type specimen, they published interactive CT scans of the same specimen (Fig. 4). The scans have subsequently been used by Naumann et al. (2019) for a study on millipede feeding mechanisms.

However, while digitisation of type specimens to a high level of detail has many benefits, it does not enable 'big data' type analyses, such as species distributions which are critical to understanding environmental change - it is likely that a balance is required in prioritisation between detailed data on some specimens and lower levels of data on many specimens.

Reliability

Reliability (data quality in the strict sense) was treated in detail by Chapman (2005a). The data that DiSSCo deals with to a high degree includes species-occurrence information, i.e., records of a particular species from a particular place. A typical species-occurrence data-point includes taxonomic/nomenclatural information (which species, subspecies or other taxon), geographical information, collector and collecting date information and often also other descriptive data, such as habitat, host plant etc.

For all these components of a data-point, but especially obvious for spatial data, their accuracy and precision need to be considered. Accuracy and precision are often confused: accuracy refers to the closeness of measured values, observations or estimates to the real or true value, whereas precision includes statistical precision (the closeness with which repeated observations conform to themselves) and numerical precision (the number of significant digits that, for example, decimal latitude/longitude is recorded in) (Chapman 2005a). The difference between accuracy and precision of species-occurrence data is shown in Fig. 5. The accuracy and precision can also be

applied to non-spatial data. For example, a collection may have an identification to subspecies level (i.e. have high precision), but be the wrong taxon (i.e. have low accuracy) or be [correctly] identified only to family level (high accuracy, but low precision) (Chapman 2005a).

Ideally, all data-points would have high accuracy and high precision. However, for some purposes, high precision is not necessary for the data to be "fit for use". This is illustrated in Fig. 5. The figure refers to spatial data, but "fitness for use" considerations also apply to other types of information. For example, for some purposes, identification to subspecies level is necessary, whereas for others, species level is sufficient. Additionally, for some purposes, year of collection is sufficient, whereas for others, the exact date or, at least, month is required.

Assessing and improving data quality

Irrespective of how carefully a dataset has been prepared, very few datasets – if any at all – are guaranteed error-free. Therefore, quality assessment and data cleaning are important aspects of digitisation.

For DiSSCo, four types of information are particularly relevant: 1) taxonomic and nomenclatural information, 2) spatial information (georeferencing), 3) collection date and 4) image quality. For fossils, 5) geological age is also essential. Concerning types 1–3, data cleaning was treated in detail by Chapman (2005b), with emphasis on 1) and 2). Just as the digitisation process itself needs prioritisation according to the four main categories of criteria, the data validation and cleaning process needs to be prioritised according to criteria of relevance, cost and feasibility.

Quality control should be done by experts with access to both the physical and digitised collections. When voucher specimens are kept in a collection, the accuracy and precision of the taxonomic/nomenclatural information can be checked by a specialist at any time, but this seldom applies to the accuracy and precision of data on location, date, collector, habitat etc. Hence a great responsibility for accuracy and precision in recording rests on the collectors themselves. An alternative approach is to use a range of online tools, such as the data quality control checks within aggregators, such as the Global Biodiversity Information Facility (GBIF) and SpeciesLink, which include checks on geocoordinates, taxon names and date formats. GBIF also provides a list of tools which include some that support assessing and improving biodiveristy data quality (https://www.gbif.org/resource/search?contentType=tool)*[10]. Bionomia is an online resource which has automated the process of parsing and cleaning names of collectors and determiners and finding associated specimens, using integrations with GBIF, Wikidata, ORCID and Zenodo. This enables the discovery of errors or inconsistencies in specimen data relating to collectors and determiners (https://bionomia.net/) (Shorthouse 2020).

Manual data cleaning, for example, by taxonomic specialists or curators, will continue to be important. For example, the identification of collectors' itineraries allows for checking

for possible errors if, for example, the date of collection does not fit the particular pattern of that collector (Chapman 2005b).

In the framework of the SYNTHESYS+ project, Walton et al. (2020b) made a "landscape analysis" for the Specimen Data Refinery that will become one of DiSSCo's e-services. Chapter 3 of Dillen et al. (2021) deals with the semantic enhancement of digital specimens, with emphasis on taxonomic names, geographical features of the specimen and names of persons (collectors, identifiers etc.) associated with the specimen.

Finally, as always, a balanced view is recommendable. It is better to release imperfect data than to hold data back in the pursuit of (impossible?) perfection. Releasing (imperfect) digital data can help to improve data quality, for example, by opening it up to comment from international experts remotely.

Cost

Cost considerations, including funding opportunities and the affordability of projects within available resources, will have a big impact as to what is prioritised in a digitisation project. The cost of digitisation has been the subject of many analyses – recent examples are Tegelberg et al. (2017), Hardisty et al. (2020), Medina et al. (2020), Walton et al. 2020a and also the costbook of DiSSCo (Landel et al. 2023. A general lesson from these analyses is that it is impossible to give a simple figure for "What does it cost to digitise a specimen"? The desired data quality, the level of infrastructure already available, as well as salary levels for different categories of people in different countries, all play a role in cost considerations.

Hardisty et al. (2020) analysed the different types of costs, based on information from seven natural history collection institutes in Europe and described the different types of costs to be considered:

- Capital costs, such as the purchase of equipment, buildings;
- Fixed operating costs (i.e. operating costs which are not dependent on the level of usage of the facility), such as maintenance contracts, some salaries, building/floor rental, heating and lighting etc.;
- Variable operating costs (i.e. operating costs which depend on the level of activity), such as per hour costs of staff carrying out digitisation tasks, barcode labels and other consumable materials.

Another useful classification described by Hardisty et al. (2020) divides costs into:

- Establishment costs, meaning the upfront costs of building and equipping a digitisation facility;
- Costs of digitising specimens;
- Costs of preserving the digitised data and making it findable, accessible, interoperable and re-usable (i.e. 'FAIR').

In particular, the costs of preserving digitised data are often neglected or underestimated, although they may constitute a very significant part of digitisation costs. See, for example, the case studies of costs in the present report. While cost, including funding opportunities, is likely to be critical to any decision to undertake digitisation, focusing on this cost alone is problematic if DiSSCo only prioritises specimens which are cheapest to digitise. Cost needs to be taken into account alongside the other criteria and is perhaps better expressed and understood as 'value for money' - the most advantageous combination of cost and quality (or likely impact) or, in other words, whether it is cost-effective to digitise certain things, because there is a feasible workflow; scientific or other relevance that will make the data impactful; sufficient data available; and funding to meet the expected costs. Cost data will be added to some of the workflows in DiSSCo's digitisation guides website (https://dissco.github.io/) and to the "digit-key" (https://digit.naturalheritage.be/digit-key) being developed by the Royal Belgian Institute of Natural Sciences.

Feasibility

The feasibility of a digitisation project is, of course, dependent on available funds. In other words, cost might be seen as one aspect of feasibility. However, cost considerations aside, there are other factors that determine a project's feasibility: Is the collection ready to be digitised? Are skilled staff available? Is the IT and other technical infrastructure geared to the task? Has a digitisation workflow been tested and established at a suitable scale?

De Smedt et al. (2022) provide a useful checklist for "pre-digitisation curation" as a contribution to the DiSSCo Digitisation Guides website (https://dissco.github.io/). "Skilled staff" not only refers to the people who do the digitisation. These people should, of course, know how to handle the sometimes fragile specimens; ideally, they would also possess some knowledge of the organisms they are digitising and of the collection in which the specimens reside. In addition to the "hands-on" digitisation staff, it is important that people with extensive knowledge of the organisms to be digitised are available, in order to ensure a high quality of the digitised data. For historical collections, knowledge on the relevant collections, collectors, expeditions etc. is also necessary.

"IT and other technical infrastructure" includes such things as cameras/scanners, conveyor belts etc., but also computing power, appropriate software, storage space and back-up options.

The human and other resources necessary for a successful project vary according to the type of specimen and the project scale. It has become known that digitisation (including at mass scale) of herbarium sheets is relatively easy. For collections of dried insects (which in terms of sheer specimen numbers constitute a very large, if not the largest part of DiSSCo's collections), methods are being developed for efficient mass digitisation of the specimens and the associated labels (Tegelberg et al. 2017, Price et al. 2018, Wu et al. 2019). Additionally, an automated mass digitisation workflow for microscope slides has been prepared (Allan et al. 2019). Wet-preserved specimens, such as invertebrates

stored in jars with alcohol or in glass tubes which are, in turn, stored in jars, pose a huge challenge in terms of human and other resources, but see Dupont et al. (2020).

The human and other resources necessary for a successful project also vary according to the desired level of data quality, including information level (e.g. MIDS), accuracy and precision.

Many, especially smaller, institutions will have difficulties mustering the necessary resources to make a digitisation project feasible. Collaboration may ameliorate this situation. DiSSCo provides a unique opportunity, not only for sharing and learning from best practice workflows which can improve feasibility, but also for direct collaboration on digitisation. The efficiency and potential impact of the digitisation of natural history collections will be immensely higher if DiSSCo-wide agreements can be made. At the DiSSCo level, it may also be possible to apply for European funds to carry out large-scale digitisation projects. DiSSCo-wide digitisation targets could be of the following types (hypothetical examples):

- X% of all herbarium sheets in DiSSCo collections databased and imaged before 20XX;
- All primary types of insects in DiSSCo collections databased to MIDS level X before 20XX;
- All African birds in DiSSCo collections databased and imaged before 20XX.

Implementing the criteria

Despite the complicated nature of the matter, the “academic” presentation of various types of criteria for prioritisation is relatively straightforward. In contrast, their practical implementation is anything but straightforward. All analyses show that there is no such thing as one primary criterion taking precedence over others. Bakker et al. (2018) outlined three methods to implement prioritisation criteria for digitisation:

1. A decision tree (not a tree, but an electronic multi-entry key), focusing on practical (feasibility) and funding (cost) criteria;
2. A scoring method;
3. A panel review.

Concerning the decision tree, Bakker et al. (2018) referred to an “Appendix 6” which, however, is not included in their report. We have had access to an incomplete draft of this appendix in the form of an extensive Excel sheet. It is obvious that constructing an operational decision tree or multi-entry key will be extremely complicated, if possible at all, even if the scope of the tree/key will be limited to feasibility and cost criteria. Therefore, we have focused on the scoring and panel methods. As pointed out by Bakker et al. (2018), these can be used one at a time or in combination and, based on the experience from the SYNTHESYS projects*[3], a combination does indeed look like the best solution.

Conclusion and Recommendations

When the DiSSCo RI becomes fully operational, it is expected that prioritisation of digitisation will, at least in part, take place at DiSSCo level. Whereas it is beyond the scope of the present report to suggest which specimens to digitise first, the preceding sections provide a background for making optimal decisions.

When choosing what to digitise and how to do it, consider:

- Where possible, collaboration on digitisation proposals, particularly within the DiSSCo framework. We support using the community itself and the rapid developments in approaches which are happening around the world as a solution in itself to help drive forward strategic prioritisation of digitisation activities. Communicating summaries of these and adding to these will have a dual role in helping others define or refine their strategies;
- Aiming to provide data that are sufficient for the use case within the project, whilst considering other likely use cases and paying attention to data quality. Biodiversity data quality is likely to affect downstream analyses, reports and decisions made based on the data and a consistent approach to assess and manage data quality will be required;
- Using a combined approach of scoring and panel review, allowing for a balanced and nuanced implementation of the prioritisation criteria.

More specifically, consider:

- Relevance, including
    - scientific relevance;
    - societal relevance.

- Data quality, including
    - level of information;
    - reliability;
    - potential for validation;
    - dataset completeness.

- Cost, including
    - pre-digitisation;
    - digitisation s.s.;
    - post-digitisation.

- Feasibility, including
    - possibilities for collaboration.

Questions to be asked

To gather the information required for prioritisation, whether for evaluation or preparation of project proposals or for preparing an internal strategy, the following questions are recommended:

RELEVANCE:

- What is the scientific relevance of the project? (Which types of research will be facilitated by the generated digital data)?
- What is the socio-economic relevance of the project? (Which economic and social benefits will result from the project? Will the project support national/European/ global political goals, including the 17 Sustainable Development Goals of the UN)?

COST:

- Is the cost/benefit ratio of the project reasonable ("value for money")?
- Are all steps in the digitisation process considered?
- Is sufficient funding available (affordability)?
- If not, is there a realistic plan for obtaining sufficient funding?

QUALITY:

- Is the level of information (e.g. MIDS) of the generated digital data sufficient for the purpose of the project?
- Is the accuracy and precision of the generated digital data sufficient for the purpose of the project?
- Is long-term storage and FAIR availability of the digital data ensured?
- Is there a plan for data validation/quality control/data enhancement?

FEASIBILITY:

- Is the necessary IT infrastructure available?
- If not, is there a realistic plan for gaining access to the necessary IT infrastructure?
- Is the necessary technical infrastructure (e.g. cameras, scanners, conveyor belts) available?
- If not, is there a realistic plan for gaining access to the necessary technical infrastructure?
- Is the necessary scientific (e.g. taxonomic experts, curators) and technical (e.g. IT) staff available?
- If not, is there a realistic plan for making such staff available?
- Is there scope for joining forces with other projects?

A final word

Finally, whereas prioritisation of digitisation is the subject of the present report, it is important to remember that the digital specimens that have been and will be created, still need links to the physical specimens since physical specimens always will be the ultimate (potential) validators (or 'vouchers') for digital data. Irrespective of the “digital revolution” in which DiSSCo takes part, physical collections, therefore, will need continued funding, including funding for skilled curators. This priority for digitisation of natural history collections is as high as any other.

Glossary

- DPP – DiSSCo Prepare Project, https://www.dissco.eu/dissco-prepare/
- DiSSCo – Distributed System of Scientific Collections, https://dissco.eu
- DiSSCo National Node - formal national representatives who form part of the DiSSCo governing body.
- FAIR – Findable, Accessible, Interoperable and Reusable, https://www.go-fair.org/fair-principles/
- GBIF – Global Biodiversity Information Facility, https://gbif.org
- ICEDIG – Innovation and Consolidation for large scale Digitisation of Natural Heritage*[1]
- NHMD – Natural History Museum of Denmark
- NSC – National Science Consortium
- RI – Research Infrastructure
- SYNTHESYS – Synthesis of Systematic Resources, https://www.synthesys.info/
- UniFi – University of Florence

Acknowledgements

We extend our thanks to all those persons who have provided information, either in the form of response to our surveys or in a more informal way. Particular thanks to Tim Robertson (ORCID: 0000-0001-6215-3617) from the GBIF secretariat and to Arthur Chapman (ORCID: 0000-0003-1700-6962) from the Australian Biodiversity Information Services for permission to use illustrations from GBIF reports.

Funding program

H2020-EU.1.4.1.1. - Developing new world-class research infrastructures

Grant title

Distributed System of Scientific Collections - Preparatory Phase Project (DiSSCo Prepare). Grant agreement ID: 871043.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Addink W, Hardisty A (2020) 'openDS' – Progress on the New Standard for Digital Specimens. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59338
- Akkari N, Enghoff H, Metscher B (2015) A New Dimension in Documenting New Species: High-Detail Imaging for Myriapod Taxonomy and First 3D Cybertype of a New Millipede Species (Diplopoda, Julida, Julidae). PLOS ONE 10 (8). https://doi.org/10.1371/journal.pone.0135243
- Allan EL, Livermore L, Price B, Shchedrina O, Smith V (2019) A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. Biodiversity Data Journal 7 https://doi.org/10.3897/bdj.7.e32342
- Bakker F, Antonelli A, Clarke J, Cook J, Edwards S, Ericson PP, Faurby S, Ferrand N, Gelang M, Gillespie R, Irestedt M, Lundin K, Larsson E, Matos-Maraví P, Müller J, von Proschwitz T, Roderick G, Schliep A, Wahlberg N, Wiedenhoeft J, Källersjö M (2020) The Global Museum: natural history collections and the future of evolutionary science and public education. PeerJ 8 https://doi.org/10.7717/peerj.8225
- Bakker HA, Willemse L, van Egmond E, Casino A, Gödderz K, Vermeersch X (2018) Inventory of criteria for prioritization of digitisation of collections focussed on scientific and societal needs. Zenodo https://doi.org/10.5281/zenodo.2579156
- Bellucci L, Bartolini-Lucenti S, Dominici S, Rook L, Cioppi E (2023) Digitalizzazione 3D delle collezioni paleontologiche del Museo di Geologia e Paleontologia di Firenze [in press]. Museologia Scientifica Memorie.
- Chapman A (2005a) Principles of Data Quality. Global Biodiversity Information Facility https://doi.org/10.15468/doc.jrgg-a190
- Chapman A (2005b) Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. https://www.gbif.org/document/80528
- De Smedt S, Bogaerts A, French L, Berger F, Cubey R, Koivunen A, Lohonya K, von Mering S, Wainwright T, Wing P, Livermore L (2022) DiSSCo Prepare WP3.2 – MS3.7 Pre-Digitisation Curation. https://know.dissco.eu/handle/item/491
- Dillen M, Groom Q, Cubey R, von Mering S, Hardisty A (2021) DiSSCo Prepare Deliverable D5.4 A best practice guide for semantic enhancement and improvement of semantic interoperability. DiSSCo Prepare https://doi.org/10.34960/ajxs-zr25

- Dupont S, Humphries J, Butcher A, Baker E, Balcells L, Price B (2020) Ahead of the curve: three approaches to mass digitisation of vials with a focus on label data capture. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e53606
- Figueira R, Fontainha E, De Smedt S, Casino A, Mergen P, Haston E (2023) DiSSCo Prepare Deliverable D1.4 - Report on socioeconomic impact indicators of DiSSCo and DiSSCo-enabled research and research applications. DiSSCo Prepare https://doi.org/10.34960/zg92-j758
- Fitzgerald H, Juslén A, von Mering S, Petersen M, Raes N, Islam S, Berger F, von Bonsdorff-Salminen TK, Figueira R, Haston E, Häffner E, Livermore L, Runnel V, De Smedt S, Vincent S, Weiland C (2021) DiSSCo Prepare Deliverable D1.1 Report on life sciences use cases and user stories. DiSSCo Prepare https://doi.org/10.34960/xhxw-cb79
- GBIF.org (2023) GBIF Occurrence Download. https://doi.org/10.15468/dl.d97dkv. Accessed on: 2023-9-01.
- Godfray HCJ (2007) Linnaeus in the information age. Nature 446 (7133): 259-260. https://doi.org/10.1038/446259a
- Hardisty A (2019) Provisional Data Management Plan for DiSSCo infrastructure. Deliverable D6.6. Zenodo https://doi.org/10.5281/zenodo.3532936
- Hardisty A, Livermore L, Walton S, Woodburn M, Hardy H (2020) Costbook of the digitisation infrastructure of DiSSCo. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e58915
- Hardisty A, Haston E (2021) Minimum Information about a Digital Specimen (MIDS). 0.14. TDWG. Release date: 2021-3-29. URL: https://github.com/tdwg/mids/blob/working-draft/old-drafts/MIDS-definition-v0.14-29Mar2021.md
- Hardy H, Knapp S, Allan EL, Berger F, Dixey K, Döme B, Gagnier P, Frank J, Haston E, Holstein J, Kiel S, Marschler M, Mergen P, Phillips S, Rabinovich R, Sanchez Chillón B, Sorensen M, Thines M, Trekels M, Vogt R, Wilson S, Wiltschke-Schrotta K (2020) SYNTHESYS+ Virtual Access - Report on the Ideas Call (October to November 2019). Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e50354
- Hardy H, Livermore L, Kersey P, Norris K, Smith V (2023) Understanding the users and uses of UK Natural History Collections. Research Ideas and Outcomes 9 https://doi.org/10.3897/rio.9.e113378
- Haston E, Hardisty A, Chapman C (2022) Minimum Information about a Digital Specimen (MIDS) - MIDS-1. https://github.com/tdwg/mids/blob/working-draft/current-drafts/MIDS-definition-v0.16-28May2022.md
- Iwanycki Ahlstrand N (2023) Digitization of the Greenland Vascular Plant Herbarium as a Unique Research Infrastructure to Study Arctic Climate Change and Inform Nature Management. Collections: A Journal for Museum and Archives Professionals https://doi.org/10.1177/15501906231159027
- Khoury C, Carver D, Greene S, Williams K, Achicanoy H, Schori M, León B, Wiersema J, Frances A (2020) Crop wild relatives of the United States require urgent conservation action. Proceedings of the National Academy of Sciences 117 (52): 33351-33357. https://doi.org/10.1073/pnas.2007029117
- Koureas D, Livermore L, Alonso E, Addink W, Alves MJ, Casino A, Curral L, Enghoff H, Guiraud M, Hardy H, Hoffmann J, Landel S, Paleco C, Petersen M, Scory S, Smith VS, Weiland C, Wesche K, Woodburn M (2023) DiSSCo Prepare Project: Increasing the

Implementation Readiness Levels of the European Research Infrastructure. Research Ideas and Outcomes 9 https://doi.org/10.3897/rio.9.e113906
- 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. URL: http://www.gbif.org/resource/83022
- Landel S, Guiraud M, Casino A (2023) DiSSCo Prepare Deliverable D4.1 The Cost Book for DiSSCo. DiSSCo Prepare https://doi.org/10.34960/kdkr-sf06
- Lendemer J, Thiers B, Monfils AK, Zaspel J, Ellwood ER, Bentley A, LeVan K, Bates J, Jennings D, Contreras D, Lagomarsino L, Mabee P, Ford LS, Guralnick R, Gropp RE, Revelez M, Cobb N, Seltmann K, Aime MC (2019) The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. BioScience 70 (1): 23-30. https://doi.org/10.1093/biosci/biz140
- Medina J, Maley J, Sannapareddy S, Medina N, Gilman C, McCormack J (2020) A rapid and cost-effective pipeline for digitization of museum specimens with 3D photogrammetry. PLOS ONE 15 (8). https://doi.org/10.1371/journal.pone.0236417
- Naumann B, Reip HS, Akkari N, Neubert D, Hammel JU (2019) Inside the head of a cybertype – three-dimensional reconstruction of the head muscles of *Ommatoiulus avatar* (Diplopoda: Juliformia: Julidae) reveals insights into the feeding movements of Juliformia. Zoological Journal of the Linnean Society 188 (4): 954-975. https://doi.org/10.1093/zoolinnean/zlz109
- Popov D, Roychoudhury P, Hardy H, Livermore L, Norris K (2021) The Value of Digitising Natural History Collections. Research Ideas and Outcomes 7 https://doi.org/10.3897/rio.7.e78844
- 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 digitisation. OSF Preprints https://doi.org/10.31219/osf.io/s2p73
- Shorthouse D (2020) Slinging With Four Giants on a Quest to Credit Natural Historians for our Museums and Collections. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59167
- Tegelberg R, Kahanpaa J, Karppinen J, Mononen T, Wu Z, Saarenmaa H (2017) Mass Digitization of Individual Pinned Insects Using Conveyor-Driven Imaging. 2017 IEEE 13th International Conference on e-Science (e-Science) https://doi.org/10.1109/escience.2017.85
- von Mering S, Petersen M, Fitzgerald H, Juslén A, Raes N, Islam S, Berger F, von Bonsdorff-Salminen TK, Figueira R, Haston E, Häffner E, Livermore L, Runnel V, De Smedt S, Vincent S, Weiland C (2021) D1.2 Report on Earth sciences use cases and user stories. DiSSCo Prepare https://doi.org/10.34960/n3dk-ds60
- Walton S, Livermore L, Dillen M, De Smedt S, Groom Q, Koivunen A, Phillips S (2020a) A cost analysis of transcription systems. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e56211
- Walton S, Livermore L, Bánki O, Cubey R, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos C, Scott B, Williams A, Wu Z (2020b) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e57602
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo

I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

- Wilson R, de Siqueira AF, Brooks S, Price B, Simon L, van der Walt S, Fenberg P (2022) Applying computer vision to digitised natural history collections for climate change research: Temperature-size responses in British butterflies. Methods in Ecology and Evolution 14 (2): 372-384. https://doi.org/10.1111/2041-210x.13844
- Wu Z, Koivunen A, Saarenmaa H, Van Walsum M, Wijers A, Willemse L, Ylinampa T (2019) State of the art and perspectives on mass imaging of pinned insects. Zenodo https://doi.org/10.5281/zenodo.3520667

Endnotes

*1 The EU-funded ICEDIG project – "Innovation and Consolidation for Large Scale Digitisation of Natural Heritage" - aimed to support the implementation phase of the new Research Infrastructure DiSSCo ("Distributed System of Scientific Collections") by designing and addressing the technical, financial, policy and governance aspects necessary to operate such a large distributed initiative for natural sciences collections across Europe. The ICEDIG project ran just over two years (January 2018 to March 2020).

*2 In Bakker et al. (2018), the categories of 'social relevance' included: contributing to public awareness, education or outreach;

contributing to conservation (policy); underpinning importance of collections to stakeholders and public; contributing to appearance and profile of institution; contributing to solving societal challenges and issues (health, agriculture, climate); extending networking and cooperation beyond traditional domain; complying with legal rules and regulations.

*3 SYNTHESYS (https://www.synthesys.info/about-synthesys.html) has run successfully from 2004 to 2023 and, as a core activity, has funded short transnational research visits to a considerable number of European collections. In the latest version of the project, SYNTHESYS+, a virtual access grant scheme to fund smaller digitisation projects of the collections, was included as well. Applications for transnational and virtual access in SYNTHESYS are prioritised and funded, based on a combination of scoring and panel review. Applications are submitted using a structured form and applications are evaluated and scored by a panel of experts. Importantly, prioritisation and funding are not decided on the basis of the panel scores alone, but are discussed at a panel meeting where aspects that cannot easily be assigned a numerical score can also be discussed and considered.

*4 NB: Especially, but not exclusively for mass digitisation, a pilot phase testing a new digitisation workflow and/or technology, is recommendable.

*5

A query of GBIF on 01-09-2023 for occurrence records from the "Data network=Distributed System of Scientific Collections (DiSSCo)" and "Basis of record=Preserved specimen" returned the following summary report:

Total: 39,679,015
Licence: CC BY-NC 4.0
Year range: 1501–2023
With year: 58 %
With coordinates: 33 %
With taxon match: 98 %
This query has been saved: GBIF.org (2023)

*6 These two subcategories had equal relevance.

*7 NB: Economic relevance ranked as equally important as educational relevance.

*8 These definitions of MIDS level differ from the more recent version of Haston et al. (2022) cited elsewhere in the document.

*9 Specimen label transcription included:

- NHMD barcode/specimen number
- Plant taxonomic data including associated author names for taxonomic rank (family, genus, specific epithet, intraspecific rank, hybrid status etc.)
- Collector(s)
- Collector number
- Collecting date (day, month, year)
- Location
- Type of sheet (single or multiple sheet)
- Multi specimen sheet (yes/no)
- Specimen in envelop (yes/no)

*10 As of 2023-08-29 there were 112 tools listed including a mix of general tools (like QGIS and R) to specific biodiversity data tools (like a Georeferencing Calculator and GBIF's scientific name parser).

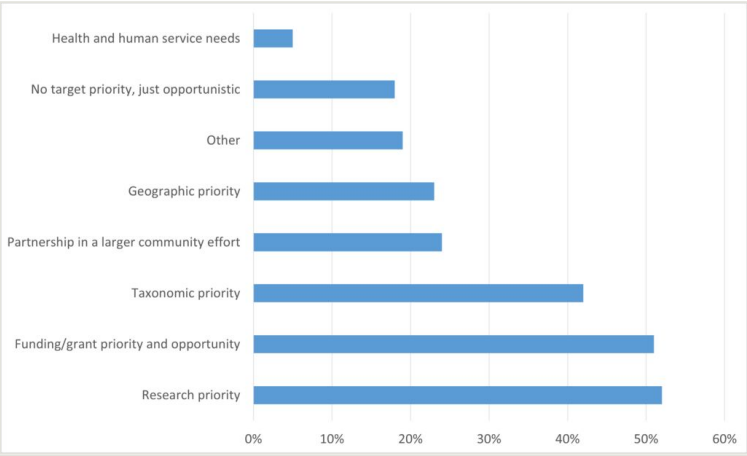

Figure 1.

Percentages of collections surveyed by GBIF applying various criteria for prioritisation of collections, from Krishtalka et al. (2016).

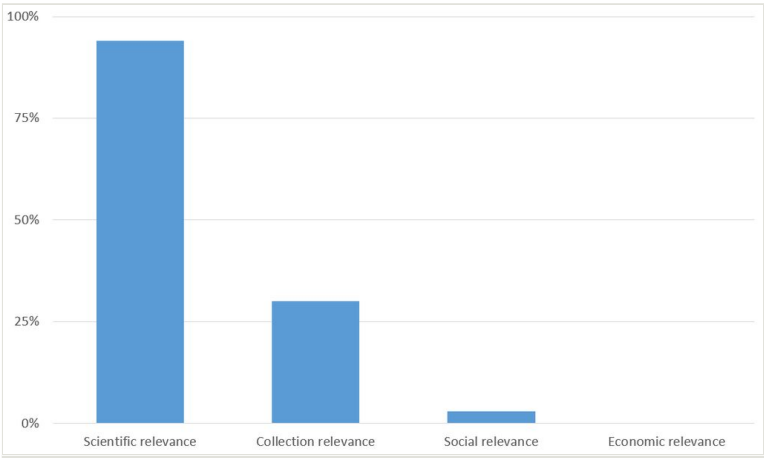

Figure 2.

Overview of the relative importance of the relevance areas identified regarding digitisation from Bakker et al. 2018 (Fig. 15).

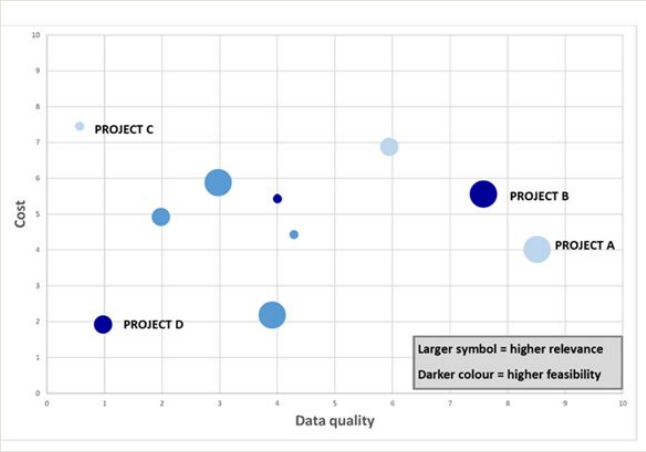

Figure 3.

Interrelation of the four main categories of criteria. Data quality and cost are represented on the horizontal and vertical axes (axis values are arbitrary). Relevance is represented by the size of the circles and feasibility by the intensity of their colour. Project A and B will both deliver data of high quality and high relevance. Although Project B data will be of slightly lower quality and slightly higher cost, this project may be chosen because of higher feasibility. Project C has little to recommend it, whereas Project D (low data quality, medium relevance and feasibility and low cost) might be prioritised depending on what the data will primarily be used for.

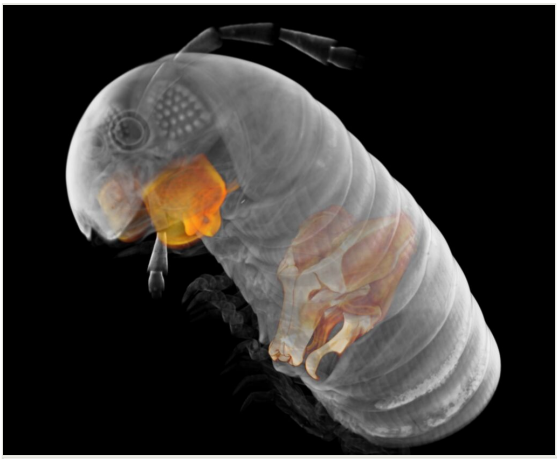

Figure 4.

A CT scan of the millipede described by Akkari et al. (2015) (© 2015 Akkari et al. used under the CC BY 4.0 license). The image shows the anterior part of the body with mouthparts and copulatory organs highlighted. The scan may be manipulated to show details important for, for example, taxonomy.

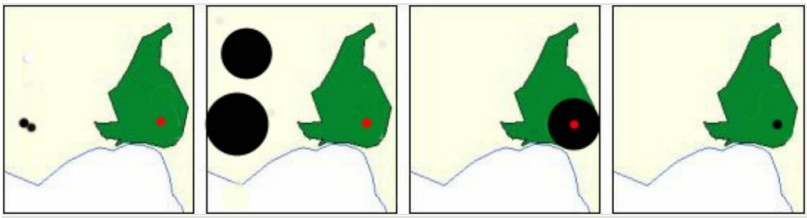

Figure 5.

The differences between accuracy and precision in a spatial context. The red spots show the true location, the black spots represent the locations as reported by a collector. *Far left* - High precision, low accuracy. *Middle left* - Low precision, low accuracy showing random error. *Middle right* - Low precision, high accuracy. *Far right* - High precision and high accuracy. From Chapman (2005a) (© 2005 Chapman et al. used under the CC BY 4.0 license).

Table 1.

Table 1. Results of the four search compilations undertaken in April 2021 and June 2022.

	April 2021		June 2022	
Search no.	**No. results**	**No. relevant**	**No. results**	**No. relevant**
1	143	4	223	6
2	775	4	1170	4
3	4460	2	4640	2
4	46	2	46	2

Table 2.

Expenses associated with the digitisation of the Greenland Herbarium at NHMD. Important: the cost for each item consists of cash costs plus time costs; conversion of time (hours) to cash (euro or other currency) has not been attempted. *71,879 out of 170,000 records had been transcribed, cleaned and imported into Specify as per August 2022; this required 128 hours. The figure in the Table, 303 = 128 × 170,000/71,879.

Process	**Cash Cost (EUR)**	**Duration (Hours)**	**Notes**
Imaging of 147,500 sheets and 15,900 folders	109,150	Not recorded	done by external contractor, paid by grant
Transcription of 170,000 labels*[9]	103,700	Not recorded	done by external contractor, paid by grant
Transport of specimens, materials and professional freezing services	12,500	Not recorded	done by external contractor, paid by grant
Project management	Not recorded	960	800 hours paid by grant, rest by NHMD
Packing of collection	Not recorded	160	paid by grant
Data management	Not recorded	303*	small part paid by grant, rest by NHMD
Collection management	Not recorded	175	paid by NHMD
Student assistance (data cleaning etc)	Not recorded	158	partly paid by grant, rest by NHMD
Total	**225,350**	**1581 hours**	**Total cost = cash (euro) plus time (hours)**

Table 3.

Example of Specify manager's work on a batch of 4019 sheets.

Item	**Time spent**	**Time upscaled to 170,000 specimens (rounded to hours)**	**Notes**
cleaning collector names – clustering	60 min	42 hours	
cleaning taxonomy – clustering	15 min	11 hours	
cleaning author names	10 min	7 hours	
cleaning infraspecific taxonomy - clustering	10 min	7 hours	
cleaning locality – clustering	90 min	63 hours	variable, depends on original data quality
uploading images	1 min	1 hour	usually scheduled to happen during night
Total	**3 hours 6 min**	**131 hours**	

Table 4.

Expenses associated with the PalVirt Project.

Item	**Cash cost (€)**	**Time cost (hours)**	**Notes**
3D models of 200 fossil specimens (acquisition and elaboration)	*56,000*	792	done by external contractor, paid by grant
Project coordinator	Not recorded	176	paid by NHM UniFi
Collection manager (Project Referent)	Not recorded	352	paid by NHM UniFi
Collection managers	Not recorded	176	paid by NHM UniFi
Total	**56,000**	**1496 hours**	**Total cost = cash (euro) plus time (hours)**

Table 5.

Types of information to be included in digital biological specimens depending on intended use.

	PRIMARY USE OF DIGITISED SPECIMENS				
TYPES OF INFORMATION INCLUDED	Taxonomic research	Other types of fundamental research (e.g. biogeographical, ecological)	Applied research (e.g. medical)	Conservation/ land use	Outreach
Taxonomy	+	+	+	+	+
Georeference	+	+		+	
Images	+				+
Habitat info	+	+		+	
Sequence data	+	+	+		

Table 6.

Four levels of MIDS (Minimum Information about a Digital Specimen). From Hardisty and Haston (2021).

MIDS level	Record extent	Purpose
1	Basic	A basic record of specimen information.
2	Regular	Key information fields that have been agreed over time as essential for most scientific purposes.
3	Extended	Other data present or information known about the specimen, including links to third-party sources.
0 (Note)	Bare	A bare or skeletal record making the association between an identifier of a physical specimen and its digital representation, allowing for unambiguous attachment of all other information.

Supplementary materials

Suppl. material 1: ICEDIG and GBIF report

Authors: Louise Isager Ahl, Luca Bellucci, Pip Brewer, Pierre-Yves Gagnier, Elspeth Haston, Sofie De Smedt, Laurence Livermore, Henrik Enghoff
Data type: Word document
Brief description: Summary of relevant data from the ICEDIG project and the GBIF task report.
Download file (590.32 kb)

Suppl. material 2: ApMS1.3 Corpus of previous studies on prioritisation of digitisation compiled

Authors: Louise Isager Ahl, Henrik Enghoff
Data type: Word document
Brief description: Analysis of previous studies, identify relevant criteria and develop them into a basic model for the prioritisation of digitisation of objects held in Natural Sciences Collections (NSCs).
Download file (675.12 kb)

Suppl. material 3: Additional literature searches

Authors: Louise Isager Ahl, Luca Bellucci, Pip Brewer, Pierre-Yves Gagnier, Elspeth Haston, Sofie De Smedt, Laurence Livermore, Henrik Enghoff
Data type: Word document
Brief description: The combined list of new and relevant studies found through two searches.
Download file (43.24 kb)

Suppl. material 4: Survey 2, Multiple-choice questionnaire

Authors: Louise Isager Ahl, Luca Bellucci, Pip Brewer, Pierre-Yves Gagnier, Elspeth Haston, Sofie De Smedt, Laurence Livermore, Henrik Enghoff
Data type: Word document
Download file (1.79 MB)