DiSSCo related output

This template collects the required metadata to reference the official Deliverables and Milestones of DiSSCo-related projects. More information on the mandatory and conditionally mandatory fields can be found in the supporting document 'Metadata for DiSSCo Knowledge base' that is shared among work package leads, and in Teamwork > Files. A short explanatory text is given for all metadata fields, thus allowing easy entry of the required information. If there are any questions, please contact us at info@dissco.eu.

Title

D3.2 DiSSCo Digitisation Guides Website - Consolidating Knowledge on Collections Mobilisation

Author(s)

Lisa French, Frederik Berger, Sabine von Mering, Pedro Arsénio, Elspeth Haston, Ann Bogaerts, Robert Cubey, Sofie De Smedt, Robyn Drinkwater, Rui Figueira, Helen Hardy, Anne Koivunen, Esko Piirainen, Vincent Smith, Peter Wing, Zhengzhe Wu, Laurence Livermore

Identifier of the author(s)

Lisa French https://orcid.org/0000-0001-7279-8582
Frederik Berger https://orcid.org/0000-0001-8400-3337
Sabine von Mering https://orcid.org/0000-0003-2982-7792
Pedro Arsénio https://orcid.org/0000-0003-3860-9789
Elspeth Haston https://orcid.org/0000-0001-9144-2848
Ann Bogaerts https://orcid.org/0000-0003-3435-2605
Robert Cubey https://orcid.org/0000-0001-7902-3843
Sofie De Smedt https://orcid.org/0000-0001-7690-0468
Robyn Drinkwater https://orcid.org/0000-0002-1820-9422
Rui Figueira https://orcid.org/0000-0002-8351-4028
Helen Hardy https://orcid.org/0000-0002-9206-8357
Anne Koivunen https://orcid.org/0000-0002-3475-7971
Esko Piirainen
Vincent Smith https://orcid.org/0000-0001-5297-7452
Peter Wing https://orcid.org/0000-0002-8634-8790
Zhengzhe Wu
Laurence Livermore https://orcid.org/0000-0002-7341-1842

Affiliation

Natural History Museum, London: Lisa French, Laurence Livermore, Vincent Smith, Peter Wing
Museum für Naturkunde, Berlin: Frederik Berger, Sabine von Mering
Universidade de Lisboa: Pedro Arsénio, Rui Figueira
Royal Botanic Garden Edinburgh: Elspeth Haston, Robert Cubey, Robyn Drinkwater
Meise Botanic Garden: Ann Bogaerts, Sofie De Smedt
Finnish Museum of Natural History (Luomus): Anne

Contributors

Koivunen, Esko Piirainen, Zhengzhe Wu

Publisher

Identifier of the publisher

Resource ID

Publication year
2022

Related identifiers

Is it the first time you submit this outcome?
Yes

Creation date

Version

Citation
French, L., Berger, F., von Mering, S., Arsénio, P., Haston, E., Bogaerts, A., Cubey, R., De Smedt, S., Drinkwater, R., Figueira, R., Hardy, H., Koivunen, A., Piirainen, E., Smith, V., Wing, P., Wu, Z., & Livermore, L. (2022) DiSSCo Digitisation Guides Website - Consolidating Knowledge on Collections Mobilisation. DiSSCo Prepare WP3 - D3.2

Abstract
In order to support the digitisation activities of DiSSCo, we have considered how best to prepare collections for digitisation, digitise them, curate their associated data, publish those data, and measure the outputs of projects and programmes. We have examined options and approaches for different types and sizes of collections, when outsourcing should be considered, and what different project management approaches are most appropriate in this range of circumstances.

This report describes the approach we have taken to developing an online community-edited manual, our guidelines, other relevant resources and platforms, and a set of recommendations on how to develop and this work to enhance future digitisation capacity across the DiSSCo collection-holding organisations.

Content keywords

Project reference
DiSSCo Prepare (GA-871043)

WP number
WP3

Project output

Deliverable/milestone number

Dissemination level
Public

Resource type
Text

Format

Funding Programme
H2020-INFRADEV-2019-2

Contact email
lisa.french@nhm.ac.uk

H2020-INFRADEV-2019-2
Grant Agreement No 871043

DiSSCo Prepare WP3– D3.2 DiSSCo Digitisation Guides Website - Consolidating Knowledge on Collections Mobilisation

Lisa French, Frederik Berger, Sabine von Mering, Pedro Arsénio, Elspeth Haston, Ann Bogaerts, Robert Cubey, Sofie De Smedt, Robyn Drinkwater, Rui Figueira, Helen Hardy, Anne Koivunen, Esko Piirainen, Vincent Smith, Peter Wing, Zhengzhe Wu, Laurence Livermore

Abstract

In order to support the digitisation activities of DiSSCo, we have considered how best to prepare collections for digitisation, digitise them, curate their associated data, publish those data, and measure the outputs of projects and programmes. We have examined options and approaches for different types and sizes of collections, when outsourcing should be considered, and what different project management approaches are most appropriate in this range of circumstances.

This report describes the approach we have taken to developing an online community-edited manual, our guidelines, other relevant resources and platforms, and a set of recommendations on how to develop and this work to enhance future digitisation capacity across DiSSCo collection-holding organisations.

Contribution to DiSSCo RI (FOR DELIVERABLES ONLY)

This work will help to enhance future digitisation capacity in DiSSCo collection-holding organisations.

Keywords

Digitisation, Guides, Workflows, Business Process Model and Notation (BPMN), Standard Operating Procedures, Best Practices

Index

Abstract 2
Contribution to DiSSCo RI 2
Keywords 2
INTRODUCTION 4
Project Context 4
Task Partners 5
DiSSCo Digitisation Guides Website 5
Summary of Work 5
Dissemination 7
Feedback 7
Digitisation Monitoring 9
1. Introduction to digitisation monitoring 9
2. Case studies 9
2.1 Monitoring digitisation rates 9
2.2 Case studies on monitoring processes for providing KPIs and downstream monitoring 13
3. Discussion and conclusion 16
Recommendations 20
Conclusion 22
Author Contributions 22
References 22

INTRODUCTION

One of the aims of DiSSCo is to provide harmonised physical and digitisation-on-demand services as part of a wider services portfolio in natural science collections (NSCs). Regardless of whether access to collections is digital or physical, both require standardised approaches and the sharing of best practices to work effectively as a multi-country, multi-institutional Research Infrastructure. As part of the scoping work undertaken in the ICEDIG Project, a series of recommendations were made regarding the arrangements, processes and practices for data mobilisation and associated tasks - key components in ensuring delivery of collections data a service (Hardisty et al, 2020). They included:

- Consolidating how we prepare collections for digitisation
- Consolidating our digitisation processes and keeping them as lean as possible
- Keeping an overview of ongoing development and innovations in digitisation workflows and technology

In this report and our online community edited manual, the DiSSCo Digitisation Guides Website (dissco.github.io) we have attempted to gather best practice examples and guides for digitisation and data mobilisation, starting from pre-digitisation curation (De Smedt *et al,* 2022), standard operating procedures for digitisation (French *et al.*, 2021), the data management and processes required for ingesting data into collections management systems and data portals (Piirainen *et al.,* 2022), and how to monitor digitisation activities and processes. While there are no existing resources that cover all of these topics in a single place, there are formal publications and project/programme resources that are relevant and useful. The most notable are resources from the nationally coordinated iDigBio digitisation activities in the USA e.g., iDigBio (2022). We have given more details of these other resources in the associated reports and the DiSSCo Digitisation Guides Website.

The scope of data mobilisation in NSCs is very broad but we hope this work is a starting point for further development, and can provide guidance for both digitisation at a national and institutional level.

Project Context

This project report, alongside the website and milestones cited below, form the Deliverable of Task 3.2 of the DiSSCo Prepare Project.

The following text is the formal description (Task 3.2) from the DiSSCo Prepare project's Description of the Action (edited for readability):

How do you best prepare collections for digitisation, digitise them, curate the associated data, publish this information and measure the outputs? What are the options and rationale for different types and sizes of collections, when should this be outsourced and what different project management approaches are most appropriate in this range of circumstances? This task seeks to address these questions, describing and refining best practices and building on a substantial investment from prior and current projects (MOBILISE COST Action, ICEDIG; SYNTHESYS+) [...]. Consolidating what is known into a community-edited manual, and other relevant platforms, [this work] will streamline the reuse and implementation of these procedures and enhance digitisation capacity across the DiSSCo collection-holding organisations.

Task Partners

Natural History Museum, London (NHM)
Finnish Museum of Natural History (Luomus)
Meise Botanic Garden (MeiseBG)
Museum für Naturkunde Berlin (MfN)
Royal Botanic Garden Edinburgh (RBGE)
Universidade de Lisboa (ULISBOA)

DiSSCo Digitisation Guides Website

Summary of Work

We have developed a DiSSCo Digitisation Guides website (dissco.github.io, Figure 1), which contains digitisation standard operating procedures (SOPs), guidelines and best practices. Milestone 3.5 describes the initial development of the site. It sets out the intended audience for the website, and outlines a template SOP (French *et al.*, 2021). This template was used to create a set of pilot workflows, including pinned insect, herbarium sheet and microscope slide SOPs.

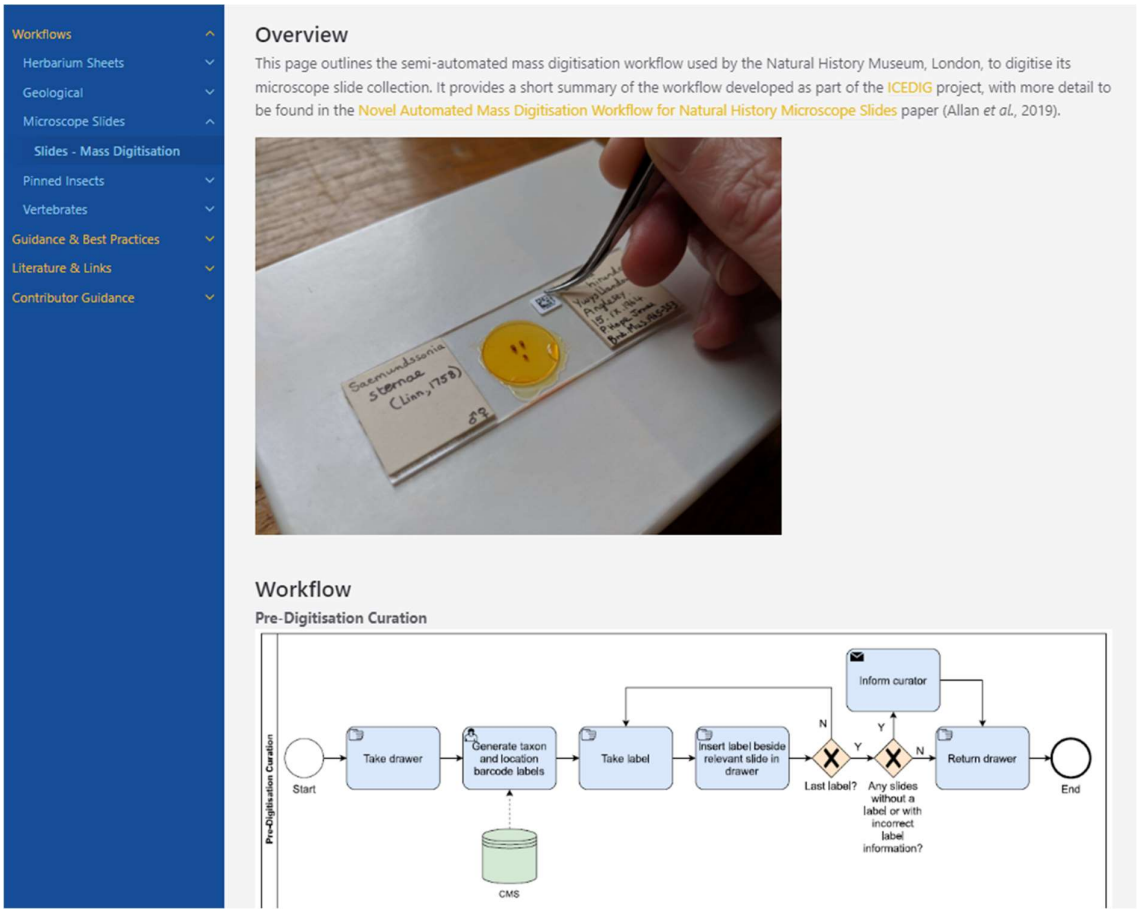

Figure 1: Screenshot of Microscope Slide workflow from DiSSCo Digitisation Guides website

The SOP template uses Business Process Model and Notation (BPMN, www.bpmn.org) to visualise the workflow steps (example in Figure 2). BPMN is used to communicate business process information to non-technical audiences, and includes a standard set of workflow elements.

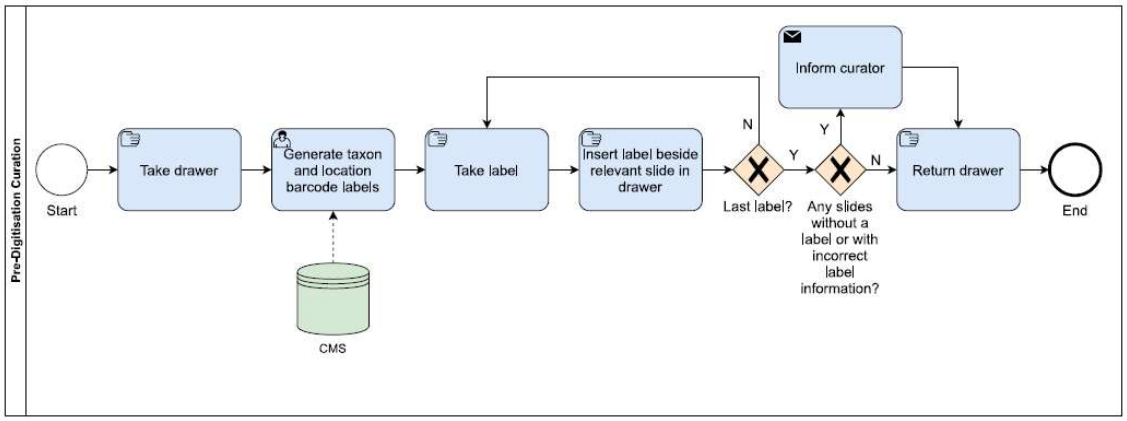

Figure 2: Example of BPMN, showing the pre-digitisation curation step from a Microscope Slide workflow.

The website is designed to be a community manual, allowing institutions to contribute their own workflows. Guidance is included to help people to write their own SOPs, and users can directly contribute through the GitHub repository.

Milestone 3.6 includes a set of best practice recommendations for Extract, Transform and Load (ETL) procedures, which were then later added to the website. Each best practice recommendation follows a standard template (Table 1), describing the level (basic/ advanced/ state-of-the-art), use case, recommendation and implementation examples. The milestone describes how these best practices should be reviewed and maintained (Piirainen *et al.,* 2022).

Table 1: Template for Best Practices

Id	***EXAMPLE1***
Level	*BASIC | ADVANCED | STATE-OF-ART*
Use case	***As*** *xxx* ***I want to*** *xxx* ***so that I can*** *xxx*
Best practice recommendation	*Procedure to follow/task to accomplish that fulfils the use case*
Discussion	*Rationale behind the recommendation*
Implementation example	*One or few references/examples on how the recommendation has been implemented in practice if applicable*
References	*Link, Ref*

A pre-digitisation curation checklist is included in Milestone 3.7. This checklist is designed to help institutions undertake a self-assessment of their collection to help prepare for digitisation. It links to useful references, and task partners contributed case studies to help explain some of the steps (De Smedt *et al.,* 2022).

We have described project management approaches to digitisation monitoring in this deliverable, and this work has been published on the DiSSCo Digitisation Guides website. We provide a series of case studies showing how digitisation processes can be monitored, including KPI development, and how downstream impact can be assessed.

In addition to the work described in the milestone documents, we have continued to improve the website design and add new SOPs and guidance. New pages added to the site since MS3.5 include:

- Herbarium Sheets SOP - MeiseBG
- Herbarium Sheets SOP - NHM
- Geological Thin Section SOP - British Geological Survey
- Specimen Image Capture Guidance

We have also added links to relevant digitisation resources to the Knowledgebase, tagging each article by collection type and digitisation stage (using the task clusters defined by Nelson *et al.* 2012).

Dissemination

The success of the digitisation guides website ultimately depends upon whether it is used by the intended audience: people working in institutions that are beginning to develop their digitisation capacity. This means that dissemination activities are critical, as many of this audience will not be directly involved in DiSSCo Prepare (although many will be members of the wider DiSSCo consortium).

We presented at the Society for the Preservation of Natural History Collections (SPNHC) Conference 2022, with our presentation providing an overview of the development of the website as well as encouraging contributions and feedback from the audience. We shared our work at the DiSSCo All Hands Meeting (AHM) 2022, again inviting the audience to share feedback.

The website has also been shared informally through our networks, for example with colleagues in DaSSCo (Danish System of Scientific Collections), DiSSCo Flanders (Belgium) and CSIRO (Commonwealth Scientific and Industrial Research Organisation) in Australia, as well as with the emerging DiSSCo UK network.

We plan to present this work at a DiSSCo National Nodes meeting, as well as sharing the website with the CETAF Digitisation Working Group. Identification of dissemination opportunities will need to continue during the construction and operation phases of DiSSCo, and this is discussed further under Recommendation 3.

Feedback

We have received feedback on the website design from participants at the DiSSCo All Hands Meeting 2022, SPNHC 2022 conference and Mobilise COST Action workshop. Feedback has also been received via e-mail and through issues on our GitHub repository. Much of this feedback has been acted upon, although further work will be required during the DiSSCo construction phase.

The DiSSCo Digitisation Guides website was created on GitHub Pages using the 'Just-the-docs' Jekyll theme. This approach meant we were able to build the website without a developer, however, it means some of the more complex UI improvements cannot be implemented. DiSSCo will need to consider whether developer resources should be assigned to this website during the construction phase, and this is discussed further in Recommendation 1 below.

Feedback has been positive, including comments that the SOPs are easy to understand, and simple to use and can be used by institutions to create internal guidance. Table 2 shows some of the suggested improvements to the site.

Table 2: Suggestions for improvements to the digitisation guides website

Feedback	Status	Notes
Include a resource section (staff, costs, digitisation rates) in SOPs	COMPLETE	Added as an optional section to the SOP template.
Add 'Changes since last version' heading	COMPLETE	Added to SOP template
Allow users to zoom into workflow diagrams	COMPLETE	See GitHub Issue #48
Provide a way for users to comment on workflows	COMPLETE	GitHub discussion forum created. In future, the DiSSCo helpdesk could be used.
Include example spreadsheets for transcription data entry	COMPLETE	We will also include examples of digitisation monitoring spreadsheets when this guidance is added to the website.
Consider colour coding workflows	OPEN	Suggestion to colour code workflows depending on digitisation stage. This should be considered further, e.g. through user feedback.
Create dois for workflows	OPEN	Zenodo provides an option to create a persistent identifier for GitHub, however this is for the full GitHub repository rather than individual pages. It may be possible to use the DiSSCo Knowledgebase, more discussion on how best to implement this is required. In the meantime, we have added 'Content Last Updated' and 'Changes Since Last Version' headings to the SOPs so users can see what has changed.

More engagement with the community is required to help prioritise the development of new SOPs. We have received the following suggestions for additional SOPs to add to the website:

- Stacking Photography of Radioactive Rocks/Minerals
- Spirit Collection Imaging
- Quality Control guidance
- Conveyor/pipeline workflows

Feedback from the community should continue to be gathered during the operation and construction phases of DiSSCo, and this is discussed further in Recommendation 3.

Digitisation Monitoring

1. Introduction to digitisation monitoring

Project monitoring and control are part of project management and run in parallel to the execution phase in a project life cycle. It enables measurement of the performance of processes, thus giving project management the opportunity to intervene in a purposeful manner and to act appropriately to adapt the process. The time effort for digitisation projects may vary greatly from a few weeks to several years, depending on the objectives, the size of collection, availability of staff and funding as well as other external factors. A structured monitoring will help to keep track of process and to fulfil the objectives on time and on budget (i.e. performance). However, it will not provide information on the quality of assets or processes. Accordingly, quality assurance and control will not be discussed in this section.

For digitisation projects, it may be distinguished between process, database and downstream monitoring. Process monitoring applies directly to the execution phase, where objects are handled and (digitally) processed. In this phase of digitisation there is usually little or no digital data available that would allow (semi-)automated monitoring. Database monitoring on the other hand refers to the phase after digitisation, where structured digital data are available. In that case regular database queries can provide the necessary information. Ideally, if the data pipeline is well established, process and database monitoring go hand in hand, but especially pre-digitisation tasks are often not reflected in databases. The following section 2.1 provides use cases from various institutions describing methods and workflows for process and database monitoring.

Downstream monitoring refers to the usage of data acquired during digitisation projects. Strictly speaking this is not part of a given digitisation project, but applies to departmental or institutional digitisation objectives. In the following it is referred to downstream monitoring as key performance indicator (KPI). A KPI is a type of performance measurement used to evaluate the success of an institution or organisation. More specifically, it can evaluate the progress and success of certain projects or initiatives the institution is involved in, according to the respective project goals. The success can also be defined as how the institution is making progress towards strategic goals (institutional KPIs). Because of the importance of KPIs for funding bodies and steering boards use cases on measuring and monitoring KPIs are included in section 2.2.

In the discussion and concluison of this section the main observations from the case studies are compiled, which leads to some key recommendations to consider, when establishing monitoring processes as part of digitisation projects.

2. Case studies

2.1 Monitoring digitisation rates

2.1.1 Natural History Museum London (NHM) Case Study: monitoring digitisation rates

Understanding digitisation rates is an essential element to our digitisation project management approach at the NHM. Each day, our digitisers record in a shared spreadsheet the number of specimens they have imaged, the number of database records created and the number of specimens transcribed. Each digitiser also includes the number of minutes spent on these tasks, which allows us to calculate the average time to digitise each specimen.

To monitor our digitisation rates, we use calculations from the Program Evaluation and Review Technique (PERT). For each digitisation project, the minimum (a), maximum (b) and median (m)

rates per person per hour are calculated. This is used to calculate a base performance rate (a + 4m + b/6) and standard deviation ((a-b) / 6).

These figures are used to schedule our digitisation projects. We take the base performance rates, alongside an estimate of specimens to be digitised in a particular project, and forecast how many person hours it will take to complete each project with additional contingency built in should estimates be found to be overly conservative. This forecast is particularly helpful if we have a project deadline coming up, as we can use this information to assign enough digitisers to ensure the specimens are digitised on time.

We monitor our digitisation rates for each project on a monthly and quarterly basis, comparing our forecasts to the actual rate. We can then adjust resources if our rates are higher or lower than expected. It also helps us to schedule new digitisation projects, as we have an estimate of when current projects will finish.

2.1.2 Museum fur Naturkunde Berlin (MfN) case study on monitoring in two collection digitisation processes

As part of the ongoing construction work, the historical bird hall of the MfN Berlin is being emptied. The hall contains approximately 11,000 mounted specimens, ranging from hummingbirds to ostriches. The process includes cleaning, repairing and labelling of the objects, followed by imaging and packing in transportation boxes. As the individual process steps require different amounts of time and expertise, the process runs asynchronously. After cleaning, specimens are temporarily stored and then collected by the digitisation team. After the imaging process, the specimens are temporarily stored again before being packed into transportation boxes. The imaging process takes place once a week and is dependent on the rate of cleaning and restoration. Therefore, the amount of birds being digitised greatly varies (between 16 and 276), with an average of 145 per week.

The start of the renovation work in the hall is scheduled and sets the deadline for the removal of the collection objects. Weekly statistics are collected to monitor the progress of the project. The digitisation step is most suitable for this, as it produces easily countable digital data. Images are being named with the specimen collection number and are stored in a daily folder. After checking the quality of the images as well as the accuracy of the file names, images are being sorted into taxonomic groups. A checklist for reference is being provided by collection management for each taxonomic group. For process monitoring, the number of processed specimens is relevant. It is irrelevant how many images are created during digitisation or how much time is required for processing individual preparations. The aim of the monitoring is merely to keep track of regular progress with regard to the deadline for emptying out the room.

The malacological collection of the Museum für Naturkunde Berlin comprises approximately 7 million individuals in 250,000 lots. Those lots will be digitized over the course of the next 4 years by transferring the collection data into a database and taking object images. The digital recording of the collection data is limited to the core elements taxon, collection number, type status and locality. If no catalogue number exists one will be assigned during this step. Data acquisition is done via a specially developed software that enables interoperability with the digitisation system. Since both process steps function completely independently of each other, monitoring is also carried out independently. Operators report the amount of processed objects (inventorized and photographed) into a spreadsheet on a weekly basis.

It is important to jointly agree on project goals, i.e. units to be processed in a certain time. Therefore various test phases were carried out in which operators and managers could determine the realistically achievable throughput. Project monitoring aims at identifying problems at an early stage

in the step up of a new workflow. However, due to staff availability, strong daily fluctuations may be observed in the throughput rates. Thus, small-scale monitoring on a daily basis may easily lead to overreactions by project controlling. The digitisation process in the mollusc collection is therefore monitored on a monthly basis, which is sufficient to identify potential workflow disturbances and to level out daily fluctuations.

As mentioned above, the process includes two independent steps that are also monitored independently by the project management. At the moment, however, only the digitisation statistics are reported to the upper project management. The reason for this is that the same specimens are being handled and monitored twice, for databasing (inventorization) and imaging (photographing). At the MfN, all monitoring results from digitisation processes are aggregated to a database in order to observe the progress of the collection discovery and development project. This includes the removal of the mounted bird preparations and the digitisation of the mollusc collection among others. Including the same specimen twice in those statistics would suggest a much faster and unrealistic progress.

2.1.3 Royal Botanic Garden Edinburgh (RBGE) Case Study: monitoring digitisation rates

During the life of the Herbarium digitisation programme, digitisation rates have changed significantly through changes in level of digitisation, equipment upgrades and software development. The mass digitisation programme currently uses bespoke software for data capture and image processing, allowing a more automated workflow. We have recently migrated to a new Herbarium collection management system (Specify) for which a rapid data entry application has been developed by an external contractor. We have also installed new imaging equipment with custom-built lightboxes and adjustable desks. We are now in the process of assessing the impact of these improvements on the digitisation rates.

Databasing rates are calculated by querying the Herbarium specimens database (Specify). This allows us to get both weekly and monthly rates for each digitiser based on the timestamp of the records they have created.

Imaging rates are calculated by querying the imaging database (in house). As with the databasing rates this allows us to record both weekly and monthly rates for each digitiser.

We currently transfer the weekly rates for imaging and databasing into an excel spreadsheet and use conditional formatting to help us visualise whether each digitiser is meeting a weekly target for databasing and imaging.

2.1.4 LISI Case Study: monitoring digitisation rates

The «João de Carvalho e Vasconcellos» Herbarium (LISI), from University of Lisbon, is a small university herbarium (about 80.000 specimens) within its School of Agriculture (Instituto Superior de Agronomia), primarily aimed to support research and education. The team involved in the digitisation operations described in this section is restricted to a curator (part time), an IT specialist (part time), and a digitiser/database operator, who is assisted in tasks related with 'Pre-digitisation Curation' by a herbarium technician.

A previous sampling had been performed in the Iberian Peninsula vascular plants collection (in 11% of the total number of shelves) in order to estimate:

a) total number of specimens (estimated specimen number in November 2016: 66,400 specimens, less than the expected number);

b) rate of total/severe damage to specimens (about 1% of the specimens were totally/severely damaged and 16% of the sampled specimens presented medium to light damage); and
c) average number of specimens by shelf (70 specimens/shelf).

The collection database was catalogued in MsAccess through several projects in different phases/years, to a total of about 75,000 records. Before the migration to a new Herbarium collection management system (Specify), it was assessed the need for revision of specimen data for consistency and correction. This task was performed using Excel and Openrefine, and implied a 6 PM effort, and did not involve data checking through specimen visualisation, which was left to the digitisation phase. Possible monitoring indicators of this stage can be:

- PM time effort
- Number of operations in Openrefine

For the image digitisation task, specimens were inspected for cleaning/remounting (if needed), before being taken to the digitisation station. A new numbering system was also devised, which implied adding to each specimen a vinyl sticker with a Barcode (Data Matrix format), as well as the corresponding text line (to ensure human and machine readability), representing the new catalogNumber (*e.g.* LISIXXXXXX, where X represents a number in the range [0-9]).

Digital imaging of specimens was then carried out, involving:
a) Image capture and file renaming (according to the specimen new catalogNumber) of each specimen;
b) Locating the specimen record in Specify;
c) Checking minimal data correctness **or** adding a new record (with minimal data) to Specify. Often, an additional verification of the collection locality correctness was also performed, mainly on previously existing location data (since new specimen georeferencing is a high time-consuming, tool and source data demanding job that is better performed as a specific, optimised task at a later moment).
Monitoring of digitisation rates was performed informally, by registering day-by-day imaging rates on a spreadsheet. Log data in Specify database (e.g. create or update timestamps) are an option, but in a limited way, because record creation date relates to the date of data imports to Specify, and update date can vary if record is edited multiple times or by multiple database users.

The numbers obtained are highly conditioned by a number of factors, which vary dynamically and for person to person. Some of the most significant influencing factors seem to be:

- Digitiser/database operator's experience level and familiarity with equipment and its setup;
- Up-to-date level of information in the database. Groups of specimens recently reviewed require fewer effort to update data (e.g. taxonomic updates in the database);
- Image station acquisition speed (*e.g.* flatbed scanners are much slower then planetary scanners);
- Handwritten labels readability, detail level (especially regarding specimen collection location data) and correctness;
- Frequency of 'new record' operations, which are more time-consuming then 'minimal data checking' operations;
- The need for 'resting breaks' from the operator, since the tasks performed demand a high level of concentration, which is hard to maintain over long periods of time.

Over a period of 169 working days (about 8 working months, and considering a 7 hour/day journey) overall figures showed an average of 50 and a median of 48 specimens digitised and database checked, per day. Considering just the imaging tasks, this number can easily rise to about 150 specimens processed by day. Figure 3 provides a quick view of the spread of rates for the whole period.

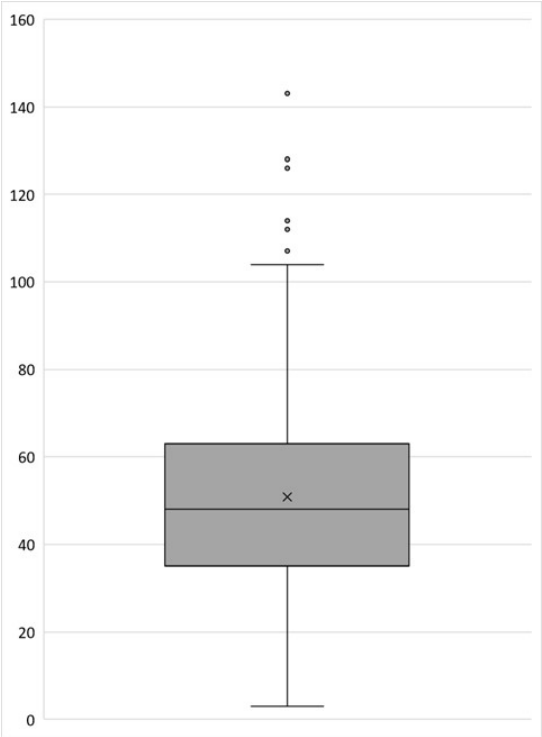

Figure 3. Box-plot representation of the daily specimen digitisation rates at LISI Herbarium. The x symbol represents the average.

All operations involved in three distinct digitization phases (Pre-digitisation curation - tasks occurring prior to databasing or imaging; Digital Image workflow - scanner setup, image capture, image processing and image archiving; Post-digitisation curation - image quality control and data quality control) are described in detail in LISI Herbarium Digitisation Workflow, as well as represented in the diagrams available at DiSSCo Digitisation Guides website (see LISI ULisboa Herbarium Sheet Mass Digitisation section).

2.2 Case studies on monitoring processes for providing KPIs and downstream monitoring

2.2.1 NHM Case Study: Institutional KPIs

The NHM has an annual Operating Plan that underpins our vision and strategy. This includes 'Page zero' targets which are a short set of targets measuring the key aspects of our business. Two page zero targets monitor digitisation, both related to the strategic goal of 'securing the future of our collection'. One focuses on digital discoverability of collections, measured by the number of new specimen records added to our data portal each quarter (the majority - but not all - of which are added by our central digitisation team), and is expressed as an annual and quarterly target. It is recognised that this target cannot be the only driver for digitisation, however - new records are balanced with other projects that enhance records or offer innovation benefits. The other relevant

page zero measure focuses on access to collections and includes download events for our data from our data portal and GBIF - this is not expressed as a target as we do not directly control it, but as a 'minimum' level based on previous averages. This is monitored quarterly and if download events fell below the minimum set further work would be undertaken to understand why and to correct this.

In addition to the page zero metrics, the Museum's strategic objective to 'transform the study of natural history' also has a number of thematic science deliverables for the year in the Operating Plan, including for the Data, Digital and Informatics team and theme. These are where we balance the page zero new records target by covering the need to deliver a wider balanced portfolio of digitisation work. This is also where we record other key aspects of work for example in relation to EU and UK projects at a high level. Typically these deliverables are time bound and we report quarterly against them but they are not expressed as quantified targets more as milestones to be reached by certain dates.

2.2.2 RBGE Case Study: Institutional KPIs

The Royal Botanic Garden (RBGE) has an annual Operational Delivery Plan (ODP) which prioritises and organises the work required to deliver the RBGE Strategy. The ODP contains deliverables for departments across the organisation, with a set of KPIs to monitor progress.

For the Herbarium collections, the following KPIs are monitored:

- Number of downloads from RBGE online catalogue
- Total Herbarium specimen records databased
- Total % of Herbarium specimen records databased
- Total Herbarium specimen images digitised and put on-line
- Total accessions recorded in the Silica-dried Collection

Each KPI has an annual target, with monthly, quarterly and annual reporting. The figures are calculated using scripts to query a set of databases: the Herbarium specimens database (Specify), the image database (In-house), the downloads database (In-house) and the silica-dried collections database (In-house).

2.2.3 MfN case study on metadata monitoring

Successful monitoring processes require clear objectives. Performance indicators can be collected regularly, but they only become interesting as a steering mechanism when they allow conclusions to be drawn about whether goals can be achieved. Digitisation rates mainly depend on the depth of information acquired during the process and are usually restricted by the availability of staff and the technological setup. For example, the acquisition of three images will require more time than just one and a high-resolution multi-focus image requires a more complex workflow than a standard definition overview image. However, not only the number and resolution of images should be defined at the beginning of each digitisation project. It is equally important to agree on which metadata can and should be acquired.

For this purpose MfN developed the framework "Minimum Extent of Information and Purpose Oriented Specimen Description" (MIPOD). As a collaborative effort with collection managers and heads of collections, all information relevant for the management and publication of object-related collection data was collected across the collection, independently of the collection management system used. In addition, lists of controlled vocabularies and links to existing external references were also taken into account where appropriate. In a bottom-up process a minimum level of information has been defined for all objects across geological, paleontological, zoological collections

as well as the library and archive. The minimum data to be acquired for each object are defined by a unique identifier, a title (e.g. taxonomic name), an object type name (e.g. PreservedSpecimen, based on ABCD RecordBasis), a collection name, the institution name and the type status (is it a type specimen or not, if applicable).

In the context of digitisation monitoring the MIPOD framework allows to define targeted metadata in a transparent and consistent way. At the beginning of each project all stakeholders can agree on a set of metadata that must be acquired. There is no limitation as to the size of the defined subset, but it means that additional data will not be acquired during that digitisation process. This allows us to manage expectations and to better control the digitisation time spent on each object. Further it enables us to monitor the completeness and thus the quality of the process.

2.2.4 NHM Case Study: Economic Benefits

Like many collections, the NHM has often relied on case studies to make the case for the impact of digitisation, alongside statistics about records downloaded, download events and citations (via GBIF).

In 2021, the Museum worked with Frontier Economics (www.frontier-economics.com) to build on this by developing estimates of the economic benefit of digitised collections. This report (Popov *et al.,* 2021) identified benefits of some £2bn over 30 years, a seven- to ten- times return on investment. Three methodologies were used to understand the economic benefits; two looking at the range of return on investment in scientific research from the literature, and applying this either to an investment figure or to an estimate of reinvested research efficiencies from digitised data (based on physical visit costs and numbers of digital download events). Most importantly, the report examined five pathways to value based on particular economic sectors or activities, estimating the value of these activities and the difference that access to digitised collections data could make. The areas examined were invasive species; agricultural research & development; medicines discovery; biodiversity conservation and mineral exploration, each showing a clear benefit through digitisation (Figure 4).

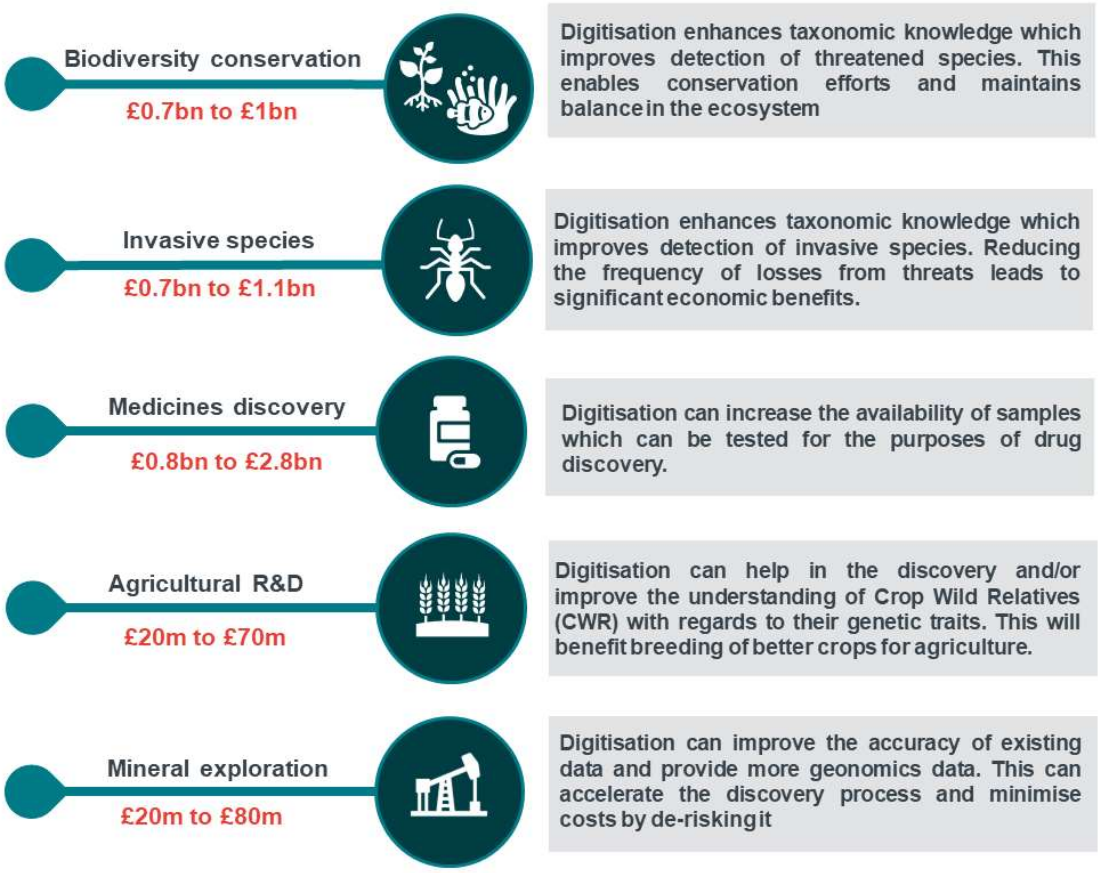

Figure 4: Five pathways to value from digitisation (Popov *et al.* 2021).

This report has been a valuable new way to discuss investment in digitisation with government or other potential funders. The NHM also hope to build on this with additional future research e.g. more detailed understanding of the users and uses of digitised specimen data, to continue developing the case.

3. Discussion and conclusion

As described in the use cases partner institutions monitoring applies primarily to digitisation rates, which allow to measure the performance, i.e. the number of objects processed in a given time frame. The use case of LISI (2.1.4) mentions in addition a monitoring step applied to data preparation, where PM time effort and operations in OpenRefine are the main indicators.

The main objective of monitoring digitisation rates is to identify potential issues of the project design, which become apparent or relevant during the execution phase. In order to interpret digitisation rates correctly it is crucial to specify the expected performance rates before project start, ideally by allowing an appropriate time for workflow testing. Based on the results of a testing phase and based on the specified performance rates it will be easily possible to refine the objectives, to calculate the costs and to specify the targets for the digitisation staff.

During the execution phase, information derived from monitoring may generally help to improve digitisation projects on the level of work organisation, of staff skills and of technology.

- Regarding work organisation, monitoring identifies elements of digitisation workflow needing improving. For a stable and consistent digitisation rate it is important that the individual steps of a workflow run at the same speed. Workflows may involve curatorial tasks to prepare objects for digitisation, metadata capture, imaging, transcription, database integration, publishing of assets and many more. Those tasks may involve different people and technologies, which may lead to high complexities and interdependencies. Monitoring helps to prevent backlog and to assure that operators have sufficient objects available to work continuously

- Training requirements become apparent, leading to the development of skills of the staff involved. Stable digitisation rates do not only require a good workflow design, but also operators need to have the right skills to fulfil the assigned tasks. Especially if projects run over a long time or involve many people (e.g. volunteers or student helpers), training is key to success.

- During the execution phase of a digitisation project technical improvements are less likely, because they may lead to inconsistencies in quality of the digital product. However, especially software improvements may speed up digitisation processes considerably. Also additional computing resources may lead to an overall improvement of the digitisation rates.

Changes on any of these three levels will affect the workflow. Monitoring then again helps to understand and to measure the impact of the applied measure.

Technical details of methods for monitoring digitisation rates are not discussed in detail in this report. Many factors may influence the choice of a specific method. RGBE established an automated monitoring pipeline using bespoke software and their collection management system, while LISI, NHM and MfN collect monitoring data in spreadsheets. Whatever method is chosen, it is important to monitor in consistent intervals, be it daily, weekly, monthly or quarterly (see 2.1.1).

Monitoring processes can reveal potential issues in the project and workflow design. The following table enumerates a number of factors that influence digitisation rates and possible solutions to react.

Table 2: Factors that influence digitisation rates

Influence	**Possible solution**
Size of collection	Assess the number of objects to be digitised accurately
Operator's experience level and familiarity with equipment and its setup	Use a competency framework to determine the required skills and train accordingly
Granularity of available database information (e.g. groups of specimens recently reviewed require fewer effort to update)	Assess the depth of available database information and define the objective of metadata acquisition (see 2.2.3) before starting the project

Influence	Possible solution
Frequency of 'new record' operations, which are more time-consuming then 'minimal data checking' operations	Assess the depth of available database information and define the objective of metadata acquisition (see 2.2.3) before starting the project
Efficiency of the technological setup image station acquisition speed (*e.g.* flatbed scanners are much slower then planetary scanners);	Measure efficiency (see 2.1.1) and make improvements before starting the project. Hardware changes should be avoided during the project
Handwritten labels readability is an important factor for transcription tasks	Make sure to have the necessary skills in the team to read a wide spectrum of handwritings. If transcription is based on digital images, make sure that smallest relevant detail is resolved (usually > 400ppi)
The need for 'resting breaks' from the operator, since the tasks performed demand a high level of concentration, which is hard to maintain over long period	Establish a performance rate before the start of a project (see 2.1.1) and allow up to an extra 20% of time effort for the execution phase in comparison to the testing phase
Acquisition of as much information as possible	It is impossible to acquire all information. To avoid fluctuations in information depth, make sure that all stakeholders agree on the outcome of the digitisation process (whole pipeline from pre-digitisation curation until data delivery and publication) before starting the project

Establishing a process for monitoring, in particular for measuring digitisation rates, is a very useful tool for managing digitisation projects. In addition, those statistics can and should be used for promoting and monitoring the performance of larger units, such as departments or institutions. When used at organisational level, it is common to refer to those monitoring data as Key Performance Indicators (KPIs). They allow assuring that individual digitisation projects are properly embedded into the institutional strategy. While data for establishing KPIs can be derived from process and database monitoring, they are usually aggregated for regular reporting on a monthly, quarterly or annual base, not in parallel to the execution phase of digitisation projects.

Common KPIs for promoting the performance of digitisation activities in institutions are:

- Total number of specimen records databased
- Total number of specimen images digitised
- Total number of specimen records and digital assets put on-line
- Percentage of specimens from the collection digitised (requires a good assessment of the total of specimens in the collection)
- Number of running and concluded digitisation projects
- Staff time spent on digitisation projects
- Actual costs for digitisation projects
- Number of thematic science deliverables

Reporting KPIs will follow institutional requirements in order to be compliant with KPIs from other departments or areas such as research or collection management. Recently, one standard emerged allowing cross-institutional comparison of digitisation activities that may be used as an indicator: Minimum Information about a Digital Specimen (MIDS, https://www.tdwg.org/community/cd/mids). MIDS reports digital objects and can cover the first three of the above-mentioned indicators. As there is still no community standard for counting specimens in natural history collections, MIDS is less suited to provide information on the percentage of specimens digitised across institutions. However, MIDS may be useful at institutional level for providing this kind of relative data.

All monitoring data mentioned so far focusses on digital data acquired from collection objects or processes for data acquisition. From a strategic perspective for institutions, but also for the larger community, it is important to understand, if and how the data from digitisation projects is used. This KPI has also been referred to as 'downstream monitoring'. So far downstream monitoring is limited to the measurement of downloads from online catalogues or GBIF (see above 2.2.1 and 2.2.2). With the progress of digitisation in collection holding institutions and the usage of stable identifiers it will be possible in future to better track the usage of collection specimens in research and beyond. As shown in case study 2.2.4 above, the economic benefit of digitally available collection objects from Natural History Institutions is huge. It will be a task for the DiSSCo community to develop the right tools and processes for measuring this impact through digitisation monitoring.

Recommendations

Recommendation 1: Maintain the digitisation guides website during the construction and operation phases of DiSSCo

The digitisation guides website should be maintained during the construction and operation of DiSSCo. At a minimum, this will require input from someone who can manage pull requests in GitHub and is able to use markdown. Business analysis skills would be beneficial, to help advise new contributors how to write standard operating procedures and produce workflow diagrams. Community engagement is also important, to promote the website, encourage contributions and respond to feedback.

Further development of the website is dependent on the direction that DiSSCo would like to take. The website was created using the GitHub Pages Jekyll theme Just-the-Docs. This approach was taken for two primary reasons:

1. GitHub is a version control system, and allows for community contributions to the site through pull requests.
2. Limited web development skills and technical resources were required to set up and maintain the website

GitHub Pages has allowed task partners to contribute directly to the website, with SOP authors able to create their own workflow page following a short tutorial on how to use GitHub pull requests and markdown. This model allows for the website to be maintained with limited resources from the DiSSCo Coordination and Support Office (CSO). However, the lack of web development and design skills within T3.2 means it has not been possible to implement all of the suggestions to improve the UI.

The website could continue under the current model, with a community manager and/or business analyst responsible for maintaining the website. It would be beneficial for a software developer to improve the UI design, e.g. to implement open suggestions in Table 2, perhaps during the construction phase of DiSSCo. However, we recommend that the site continues to use GitHub Pages (or a similar option that allows for community contributions e.g. a wiki) so that developer resources are not required to update and add new SOPs.

Recommendation 2: Establish a process for reviewing of digitisation standard operating procedures and best practices

A process to review the digitisation standard operating procedures and best practices needs to be agreed. This includes reviewing new guidance, as well as a regular review period for published content. It is possible that a collaborative process could be established with the CETAF Digitisation Working Group.

It is recommended that best practices have a more rigorous review process than SOPs, and are reviewed by relevant experts before publication. This website is intended to be a community manual for digitisation, and a long review process for SOPs might discourage contributions to the site. However, DiSSCo could consider establishing a process to label some SOPs as a ‘best practice SOP’.

Recommendation 3: Disseminate the digitisation guides website to natural science collection-holding institutions

Dissemination of the digitisation guides website could be led by the national nodes. Each national node will understand the needs of their community and can develop a communications plan. The DiSSCo CSO could help to support these activities. This could include an element of 'hands on' training sessions with digitisation teams and could be considered as part of the DiSSCo Training Strategy in Task 2.1. These dissemination activities are also opportunities to seek feedback on the user interface and workflow design, and this should continue to inform the development of the website.

For example, the UK national node (DiSSCo UK) funded the development of some of the SOPs through a grant from the UK Arts and Humanities Research Council (AHRC). Future plans to share this work by DiSSCoUK include presentations at DiSSCoUK meetings, blogs on institutional websites (e.g. NHM blog), news stories on the DiSSCoUK website, and posts on social media (Twitter/Linkedin).

DiSSCo can also help dissemination by adding a link to the site on the main DiSSCo website, by sharing on its social media channels, and including presentations on the agenda of DiSSCo-linked conferences.

Recommendation 4: Encourage institutions involved in virtual access projects to submit their digitisation SOPs to the digitisation guides website

Institutions involved with DiSSCo-related virtual access projects, such as the recent SYNTHESYS+ calls, could be encouraged to submit their workflows from these projects to the digitisation guides website. This would help institutions to share knowledge, and act as a record of the digitisation procedures used within these projects.

Recommendation 5: Add tags to workflows to help direct users to SOPs that are relevant to their institutional circumstances

There are currently multiple SOPs for herbarium sheets on the digitisation guides website, and it is likely that other collection types will see an increase in workflows (e.g. pinned insect, microscope slides). This will make it more challenging for users to find an appropriate SOP, and some guidance will need to be provided. Tags could be added that describe the SOP, e.g. 'high throughput'', 'low-cost', 'out-sourced', which could then allow the user to filter workflows.

Recommendation 6: Identify areas where investment in digitisation software, hardware and processes can have high impact

From the work and discussions in this task, it is clear that there are many areas where investment in digitisation and data mobilisation would increase rates and the efficiency of processes. For subsequent phases DiSSCo we strongly recommend an impact-led approach to investigating and investing in digitisation. There are still many large collection/preservation types without scaleable workflows, and where targeted investment could have a large impact.

Conclusion

The DiSSCo Digitisation Guides website has been designed as a community-based resource for digitisation standard operating procedures, best practices and guidance. The milestones from this task describe the development of this resource, and the website includes workflows for commonly digitised collections, including pinned insects, herbarium sheets and microscope slides. Guidance has been provided on pre-digitisation curation, ETL procedures and digitisation monitoring. This deliverable has outlined a set of recommendations for the development and maintenance of this website during the construction and operation phases of DiSSCo. Continuous feedback from the community will be important to ensure the resource remains relevant and up to date.

Author Contributions

Contribution types are drawn from CRediT - Contributor Roles Taxonomy

Conceptualization: Laurence Livermore

Investigation: Pedro Arsénio, Frederik Berger, Ann Bogaerts, Robert Cubey, Sofie De Smedt, Robyn Drinkwater, Lisa French, Rui Figueira, Elspeth Haston, Helen Hardy, Anne Koivunen, Esko Piirainen, Vincent Smith, Sabine von Mering, Zhengzhe Wu, Laurence Livermore

Methodology, Visualization, Project Administration: Lisa French

Writing - original draft: Lisa French, Pedro Arsénio, Frederik Berger, Elspeth Haston, Helen Hardy, Laurence Livermore, Sabine von Mering

Writing - review & editing: Sofie De Smedt, Helen Hardy, Rui Figueira, Laurence Livermore, Peter Wing

References

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) Pre-Digitisation Curation Checklist. DiSSCo Prepare WP3 - MS3.7. Available at: https://know.dissco.eu/handle/item/491 [Accessed 28 July 2022]

French, L., Livermore, L., Haston, E., Drinkwater, R.. Arsénio, P., Figueira, R., Berger, F., Bogaerts, A., Cubey, R., De Smedt, S., Hardy, H., King, S., Koivunen, A., Piirainen, E., von Mering, S., Wu, Z., Smith, V. (2022) Digitisation Standard Operating Procedures. DiSSCo Prepare WP3 – MS3.5. Available at: https://know.dissco.eu/handle/item/468 [Accessed 12 July 2022]

Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul DL, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6: e54280. https://doi.org/10.3897/rio.6.e54280

iDigBio (2022). Workflow Modules and Task Lists https://www.idigbio.org/content/workflow-modules-and-task-lists [28 July 2022]

Nelson, G., Paul, D., Riccardi, G. & Mast, A.R. (2012) Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 2019: 19-45. http://dx.doi.org/10.3897/zookeys.209.3135

Piirainen, E, Wu, Z., French, L., De Smedt, S., Figueira, R., Arsénio, P., Haston, E. & Livermore, L. (2022) Best Practice Standardised Extract, Transform and Load (ETL) procedures. DiSSCo Prepare WP3 - MS3.6. Available at: https://know.dissco.eu/handle/item/468 [Accessed 12 July 2022]

Popov D, Roychoudhury P, Hardy H, Livermore L, Norris K (2021) The Value of Digitising Natural History Collections. Research Ideas and Outcomes 7: e78844. https://doi.org/10.3897/rio.7.e78844