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

DiSSCo Prepare WP3 – MS3.5: Digitisation Standard Operating Procedures

Author(s)

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

Identifier of the author(s)

Lisa French: https://orcid.org/0000-0001-7279-8582
Laurence Livermore: https://orcid.org/0000-0002-7341-1842
Elspeth Haston: https://orcid.org/0000-0001-9144-2848
Robyn Drinkwater: https://orcid.org/0000-0002-1820-9422
Pedro Arsénio: https://orcid.org/0000-0003-3860-9789
Rui Figueira: https://orcid.org/0000-0002-8351-4028
Frederik Berger: https://orcid.org/0000-0001-8400-3337
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
Helen Hardy: https://orcid.org/0000-0002-9206-8357
Sally King: https://orcid.org/0000-0002-5809-9811
Anne Koivunen: https://orcid.org/0000-0002-3475-7971
Esko Piirainen
Sabine von Mering: 0000-0003-2982-7792
John Zhengzhe Wu
Vincent Smith: https://orcid.org/0000-0001-5297-7452

Affiliation

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

Contributors

Robyn Crowther, Ana Raquel Cunha, Kate Holub-Young, Michael Jardine, Phaedra Kokkini, Krisztina Lohonya, Jennifer Pullar, Larissa Welton

Smedt

Publisher

Identifier of the publisher

Resource ID
https://doi.org/10.34960/wyhd-ef59

Publication year
2022

Related identifiers

Is it the first time you submit this outcome?
Yes

Creation date
14/01/2022

Version

Citation
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

Abstract
Aims
- Review existing approaches and research on digitisation workflow documentation.
- Develop standardised methodology for recording digitisation workflows.
- Pilot this methodology to create standard digitisation workflows from partner institutions
- Identify the areas where standard operating procedures (SOPs) would be most helpful for capacity building, through a review of existing resources and national node feedback. This will be used to inform next steps on the task, after the milestone, to be incorporated into the final Deliverable report.
- Make recommendations to support digitisation as part of the DiSSCo Plan.

Content keywords
organisational

Project reference
DiSSCo Prepare (GA-871043)

WP number
WP3

Project output
Milestone report

Deliverable/milestone number
3.5

Dissemination level
Public

Rights

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 – MS3.5: Digitisation Standard Operating Procedures

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

Aims

- Review existing approaches and research on digitisation workflow documentation.
- Develop standardised methodology for recording digitisation workflows.
- Pilot this methodology to create standard digitisation workflows from partner institutions
- Identify the areas where standard operating procedures (SOPs) would be most helpful for capacity building, through a review of existing resources and national node feedback. This will be used to inform next steps on the task, after the milestone, to be incorporated into the final Deliverable report.
- Make recommendations to support digitisation as part of the DiSSCo Plan.

Keywords

Digitisation, Standard Operating Procedures, Business Process Model and Notation (BPMN), Workflow, Digitisation Guides

Contents

01 INTRODUCTION

One of the aims of the Distributed System of Scientific Collections (DiSSCo) is to provide harmonised physical and digitisation-on-demand services as part of a wider services portfolio (Hardisty et al., 2020a); this has been effectively trialed as part of the Virtual Access work in the SYNTHESYS+ project (Hardy et al., 2020). For many natural science collections (NSCs) physical access is the main mechanism through which scientists interact with collections but this is shifting, with digital access becoming more important. It is also worth noting that mass digitisation, in the form of creating inventory or stub records, supports both physical and digitisation-on-demand services through increased discoverability and more efficient curation (e.g., tracking and processing using barcodes applied during digitisation). In order to provide standardised digitisation-on-demand services across multiple NSCs we require standard operating procedures (SOPs), and eventually service level agreements (SLAs) supported by a DiSSCo IT Service Management (ITSM) framework.

Developing standardised operating procedures for digitisation that can be used by a range of institutions poses multiple challenges. Both collections themselves and their supporting technical infrastructure (e.g., collection management systems, digital asset management systems, physical network infrastructure, local file management, and data standards) are heterogeneous. This makes it difficult to generalise any SOPs, or indeed any aspect of collections data mobilisation and curation, without some level of simplification. Another challenge is the maintenance and upkeep costs of the workflow documentation. While we only have anecdotal evidence, we are working on the assumption that most institutions frequently adjust workflows and their supporting processes, but capturing these changes and then generalising them so they are useful to others can impose a relatively high labour cost with little direct benefit to the institute sharing the workflow. Finally, there are regular technological changes to imaging and automation hardware and software approaches. These changes require modifications and updates to workflows.

Other approaches to describe and standardise digitisation workflows include Nelson et al's (2015) review of more than 30 herbaria-based workflows in the US, as part of the nationally coordinated iDigBio digitisation activities. The account proposed a modular approach to describing workflows, covering all phases from pre-digitisation to data management and publication. A similar exercise was performed for other types of collections, like pinned insects, wet collections or three dimensional objects (iDigBio, 2022).

One of our key questions was "What is an appropriate level of abstraction for workflows that are still informative but do not go into institutionally specific detail"? Is this level of abstraction useful? How can we test whether it is or is not?

Scope

We have deliberately focused on digitisation workflows but recognise there are other components that are required to run successful digitisation projects. These are discussed in the '**Recommendations and Next Steps**' section.

Project context

This project report was written as a formal Milestone (M3.5) of the DiSSCo Prepare Project.

The following text is the formal description (Subtask 3.2.1) from the DiSSCo Prepare project's Description of the Action (workplan):

This subtask will publish SOPs for major collection types, documenting digitisation workflows that include information on when different scales of operation demand different modes of digitisation. This will also cover entry point digitisation (i.e. for the acquisition and digitisation of new collections), as well as the digitisation of pre-existing collections. Many of these workflows have been well established through related projects but are poorly documented. This subtask will take into account the contrasting scales and needs of day-to-day databasing operations including targeted research focussed digitisation, 'on demand' digitisation and major institutional digitisation programmes.

02 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 (RGBE)

Universidade de Lisboa (ULISBOA)

03 Community Digitisation Manual

DiSSCo projects, including SYNTHESYS and ICEDIG, have all created best practices and workflows for digitisation (Table 1). These workflows are published in various places, including academic publications and project deliverables and milestones. These resources have the potential to enhance digitisation capacity across DiSSCo Partner institutions and national node members, but it can be difficult for institutions to access and find this information. Workflows also evolve as technology improves, whereas the outputs from these projects remain static. Task 3.2 (T3.2) aims to create a community edited digitisation manual, which will act as a source for DiSSCo digitisation standard operating procedures and best practices.

Table 1: Digitisation best practice resources from ICEDIG and SYNTHESYS projects

Project	Title	DOI
ICEDIG	D3.1 Quality Management Methodologies for Digitization Operations	https://doi.org/10.5281/zenodo.3469521
ICEDIG	D3.2 State of the art and perspectives on mass imaging of microscopic and other slides	https://doi.org/10.5281/zenodo.3364481
ICEDIG	D3.3 State of the art and perspectives on mass imaging of skins and other vertebrate material	https://doi.org/10.5281/zenodo.3364385
ICEDIG	D3.4 State of the art and perspectives on mass imaging of liquid samples	https://doi.org/10.5281/zenodo.3469547
ICEDIG	D3.5 State of the art and perspectives on mass imaging of pinned insects	https://doi.org/10.5281/zenodo.3520667
ICEDIG	D3.6 Best practice guidelines for imaging of herbarium specimens	https://doi.org/10.5281/zenodo.3524263
ICEDIG	D3.7 Rapid 3D capture methods in biological collections and related fields	https://doi.org/10.5281/zenodo.3469531
ICEDIG	D3.8 R&D in robotics with potential to automating handling of biological collections	https://doi.org/10.5281/zenodo.3719101
SYNTHESYS3	D4.1 Developing edge detection technology for natural history images	https://doi.org/10.34960/x67b-2314
SYNTHESYS3	D4.2 Automating data capture from natural history specimens	https://doi.org/10.34960/bm82-vx46

Audience

The purpose of the Community Digitisation Manual is to help enhance digitisation capacity across DiSSCo partners and the DiSSCo national nodes, and it is important to consider the target audience when developing this resource. This will inform design decisions, including the style, the level of detail required and the format of workflows.

Audience and Workflow Design: iCollections - Lepidoptera Digitisation

iCollections was a project at the NHM which digitised 181,545 lepidopteran pinned insect specimens. A digital image was taken of each specimen, and the species name, georeferenced location, collector and collection date was digitised (Paterson *et al.,* 2016). This project can help to illustrate the importance of defining an audience when designing workflows. The iCollections project workflows were published in a number of formats, each suitable for a unique audience

Public Website and Blog Posts

The iCollections pinned insect workflows were used to publicise the NHM's Digital Collections Programme on its public website. This was in the form of a webpage (Figure 1) and a blog describing the work of a digitiser (Figure 2). The audience for these workflows was primarily the general public, and the focus was therefore on including high quality images and diagrams alongside accessible explanations. The blog post also included a time-lapsed video of the process. Although designed with the general public in mind, these workflows would also be useful for those new to digitising pinned insect collections, and much of the content could be repurposed for training material.

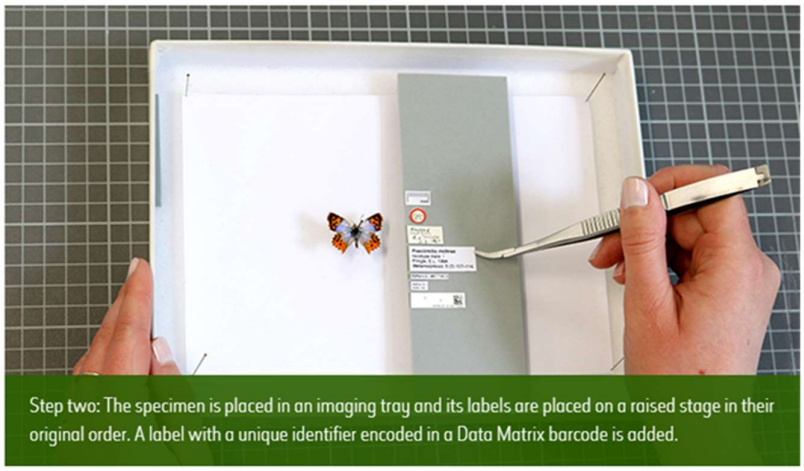

Figure 1: Pinned Insect Digitisation Workflow from NHM London website (Pullar, 2019)

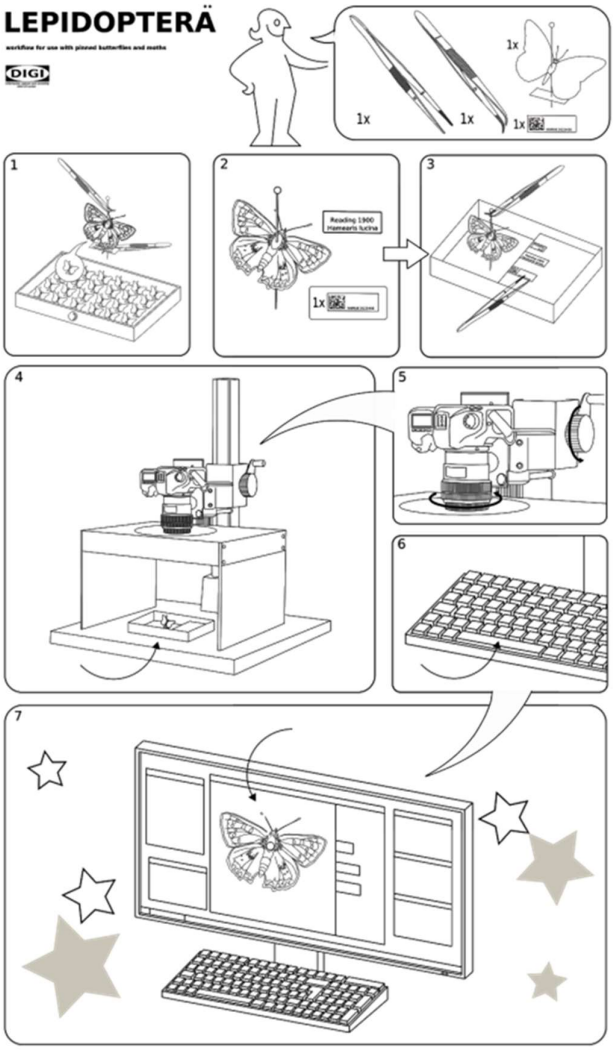

Figure 2: iCollections workflow for 'Day in the life of a digitiser' blog post (Devenish, 2019)

Academic Publication

The iCollections dataset was published as a data paper in the Biodiversity Data Journal. The audience for this work would be primarily researchers and other digitisers. The workflow in this paper (Figure 3) is more detailed than those shared on the public website, with a focus on the electronic data capture rather than the imaging process (although this is still included in the paper) (Paterson *et al.,* 2016).

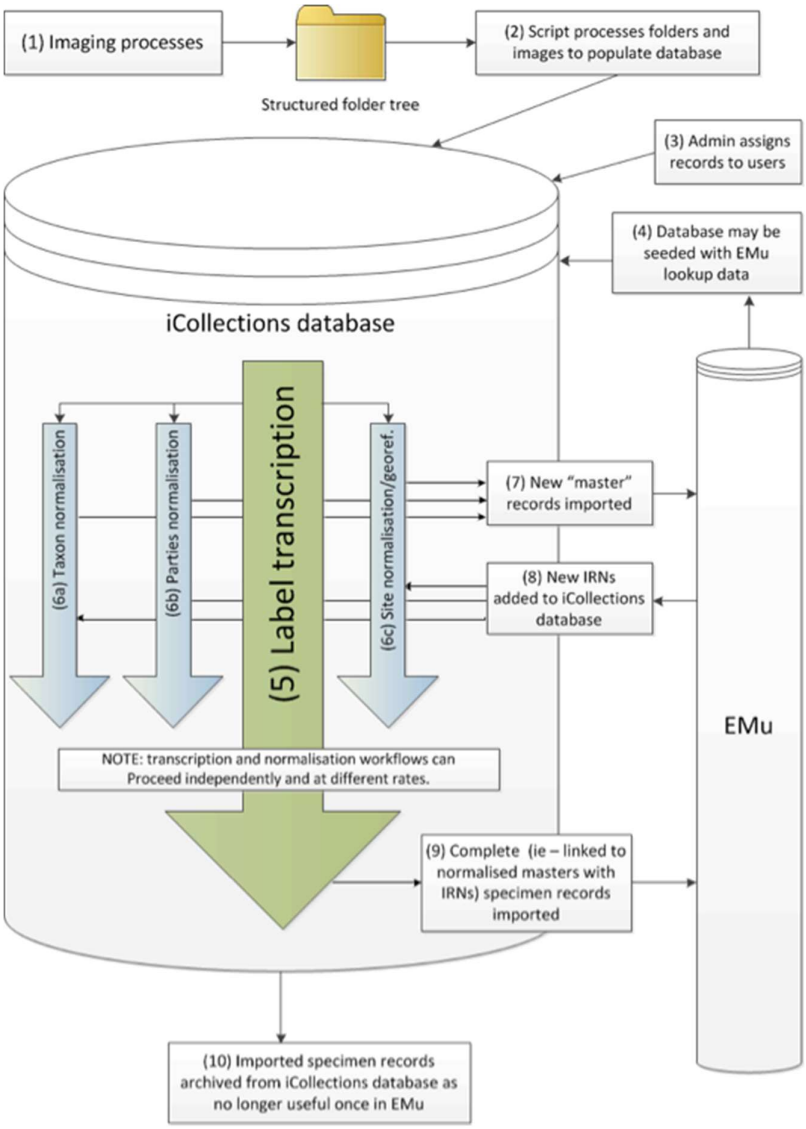

Figure 3: iCollections workflow (Paterson *et al.*, 2016)

Internal Workflow

The NHM's internal iCollections workflow is a step by step guide, containing screenshots and photographs (Figure 4). The intended audience is the museum's digitisers, and it is used as training material for new starters and as a reminder of the process for experienced digitisers. It would not be suitable for sharing externally, as it includes detail that is specific to the NHM - such as information which is specific to the layout of the digitisation laboratory and the NHM's collection management system.

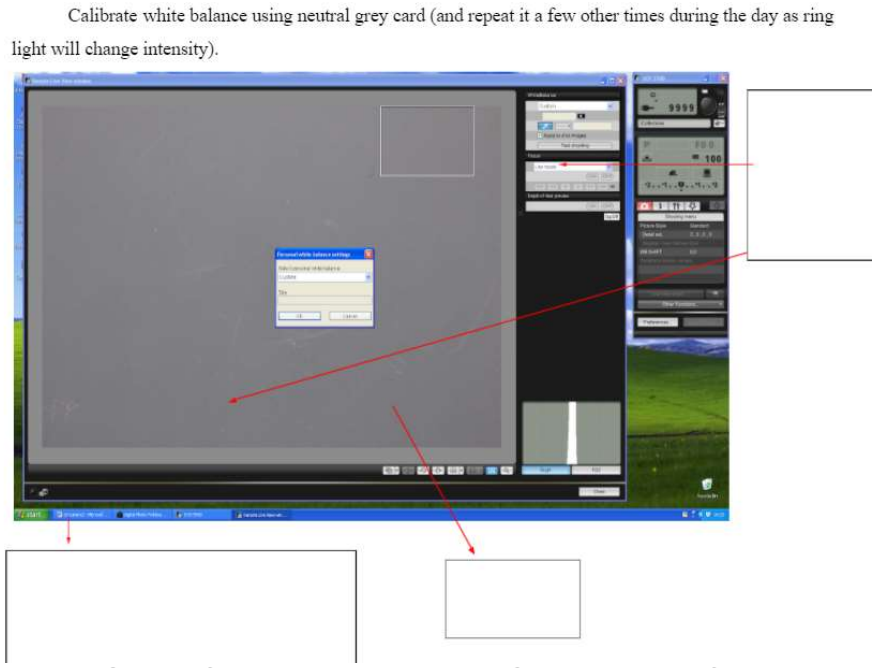

Figure 4: Section from the NHM's internal iCollections workflow, showing a detailed guide to the digitisation process.

These examples illustrate the importance of considering the audience when designing workflows, and this was used as a starting point for discussion of the audience for the Community Digitisation Manual.

Community Digitisation Manual: Audience

The target audience for the community digitisation manual was defined using the Atlas of Living Australia's (ALA) Digitisation Maturity model (Figure 5). This model was developed as a framework to help institutions assess their digitisation maturity and identify areas for improvement (Kalms, 2012). There are six levels of maturity, from 'Disorganised', where there is inconsistent practice and little governance, to level 5 'On the look out' where the organisation continuously improves processes and digitised data is managed as a strategic asset.

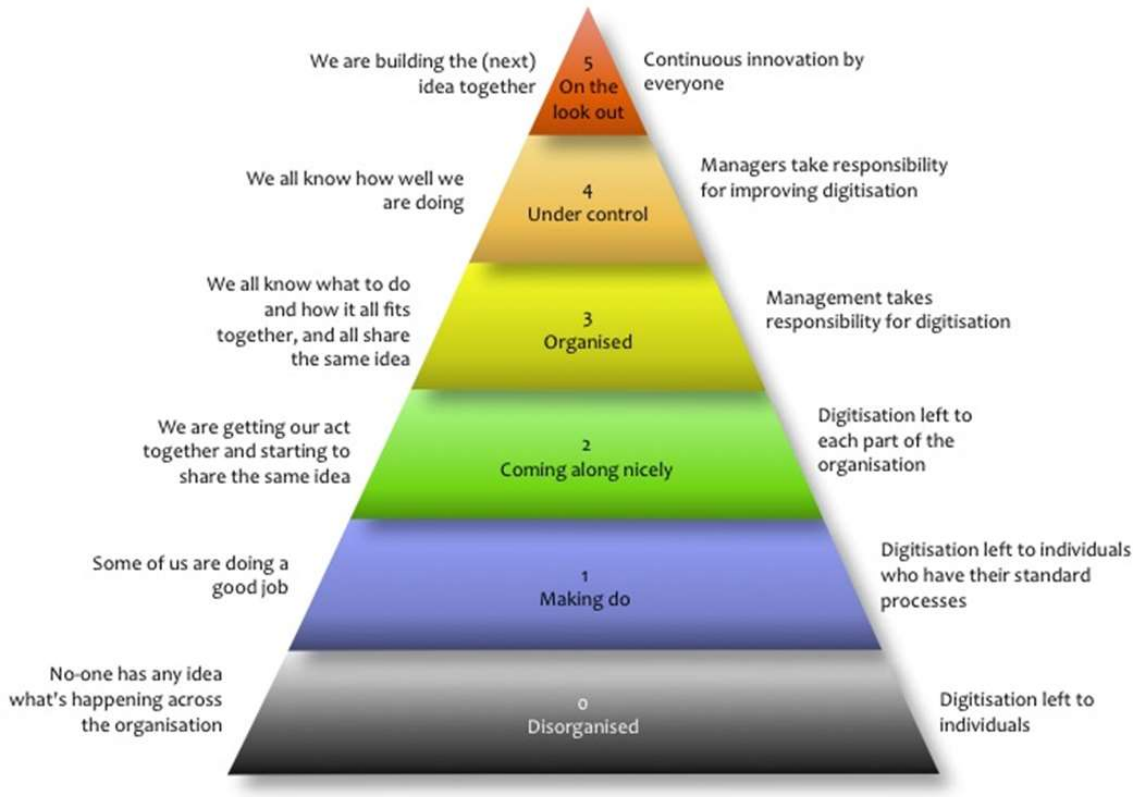

Figure 5: Atlas of Living Australia Digitisation Maturity Model (Kalms, 2012)

Level 0: 'Disorganised'

Digitisation Maturity Level 0 represents organisations which have limited digitisation. There may be individuals within the institutions that digitise and manage their own data. Digitisation practice will be inconsistent, and there is little governance within the organisation. These organisations are likely to require a lot of resources and guidance, on topics ranging from collections management systems, network resources and guidance on imaging techniques and equipment.

Level 1: 'Making Do'

Organisations at Maturity Level 1 show consistent digitisation practice at an individual level, rather than organisational level, often with inadequate 'make-do' equipment. Individuals will likely use their own personal storage for storing digital assets, with restrictive licences and data sharing based on personal requests. At an organisation level, policies and procedures are unlikely to be in place and there is no strategic recognition of the value of digitisation (Kalms, 2012). Organisations at this level are likely to find value in basic digitisation guidance, including equipment/hardware guidance, simple workflows and guidance on data sharing.

Level 2: 'Coming along nicely'

At this maturity level, a digitisation manager oversees digitisation activity, and some procedures exist for common digitisation activities. The organisation uses a central collections management system (CMS), and there are information management plans in place. There is executive support for a strategic approach to digitisation (Kalms, 2012). Organisations at this level may look for best practice

guidance to improve their own workflows, protocols for collection types which they have not yet begun to digitise, and guidance on data sharing.

There is also a potential audience for sharing more complex workflows between DiSSCo Partner institutions. This may be particularly beneficial for virtual access projects, where several institutions work on a digitisation project, each digitising part of their collection. The community digitisation manual could be used as a platform to share workflows, helping to facilitate discussion, improve processes and develop best practice. This use case was not used to drive the prototype developed for this milestone, but could be a consideration in the development of future virtual access calls. This is likely to benefit organisations at maturity levels 3-5.

Agreed Audience: The initial target audience for the community digitisation manual was agreed to be organisations at the ALA Digitisation Maturity Level 1 and 2. At this stage, Maturity Level 0 was considered out of scope, as organisations at this level would require detailed guidance and individualised support.

Prototype Development

Requirements

A short list of requirements were written to aid in the selection of a website to host the prototype community digitisation manual (Table 2). These requirements were developed based on discussions in task meetings.

Table 2: Requirements for Community Digitisation Manual

ID	User Story	Acceptance Criteria	Priority*
1	As a Digitisation Co-ordinator I want to access best practice examples of digitisation workflows so that I can adapt these to digitise my collection	1.1 User can view digitisation workflow pages	4
		1.2 User can search for a digitisation workflow pages	4
		1.3 Workflow pages can include images and text	4
		1.4 User can filter a list of digitisation workflow pages	2
2	As a DiSSCo Coordination and Support Office (CSO) user, I want to be able to curate the content in the community manual so that I can ensure the resource is useful for the community	2.1 DiSSCo CSO User can upload, edit, remove and replace all digitisation workflow pages	4

ID	User Story	Acceptance Criteria	Priority*
	As a DiSSCo Coordination and Support Office (CSO) user, I want to be able to curate the content in the community manual so that I can ensure the resource is useful for the community	2.2 DiSSCo CSO User can manage access permissions for users to submit new workflow pages and edits to existing pages	4
		2.3 DiSSCo CSO User can approve content from Institutional Users before it goes live on the site	4
		2.4 DiSSCo CSO user can tag workflows with keywords	2
3	As a Digitisation Co-ordinator I want to share my digitisation workflows so that other institutions can learn from our projects	3.1 Institutional User can submit digitisation workflows to the digitisation manual	4
		3.2 Institutional user can submit updates to previously completed workflows	4
		3.2 Institutional User can use a template to write their digitisation workflows	3
		3.3 Institutional User can tag workflow with keywords to help with search	2
		3.4 Workflow pages have a persistent identifier	2
4	As a Digitisation Manager, I want to be able to ask questions about the workflows so that I can apply them in my own institution	4.1 Users can ask questions about digitisation workflows	3
		4.2 Institution User can respond to questions from users	3

ID	User Story	Acceptance Criteria	Priority*
5	As a Digitisation Manager, I want to know when other institutions have shared new workflows	5.1 Users can sign up to notifications when new workflow pages are added	2
		5.2 Users can sign up for notifications for specific collection types	2

*(4=must have 3=should have 2=could have 1=won't have)

DiSSCo Knowledge Base

The DiSSCo Knowledge Base is a central repository for documentation related to DiSSCo, including research outputs from DiSSCo-linked projects, and is implemented with DSpace. It can also provide a place to store training materials, best practice, technical documentation and guidelines as well as publications relevant for the digitization process. The Knowledge Base was first investigated as a potential site to host the prototype manual.

The Knowledge Base would allow for workflows to be tagged with a basic metadata schema, for example by tagging workflows by digitisation stage and collection type, and would allow users to filter a list of workflows. However, there are limited options to format the pages documents are stored in and it is difficult to build user-friendly web pages.

GitHub Pages

GitHub Pages are public websites hosted on github.io. They allow users to build websites for projects from a github repository.

GitHub Pages allow for the creation of user-friendly web pages using the same tools in GitHub. GitHub Pages uses Jekyll, which is an open source tool to develop static websites, and allows for website pages to be customised. Users can contribute directly toward the website through the site's github repository, with central administrators able to give access to expert users and to manage when a new workflow goes live through pull and merge requests.

Prototype Design

The prototype website was initially developed using GitHub Pages, due to the higher level of customisation that was offered in comparison to the DiSSCo Knowledge Base. The ability to create and customise user-friendly pages is important given the target audience for the community manual. Discussions will continue with the Knowledge Base, as this may be a helpful tool to index the resources, and the GitHub Pages site will link to scientific articles and project outputs currently within the Knowledge Base repository. A dedicated directory could be created in the Knowledge Base to store material for the digitisation guidance.

The prototype was created using GitHub Pages with Jekyll, using the Just the Docs Jekyll theme. This theme is designed for documentation, and allows a simple navigation structure to be added to a GitHub Page, as well as a search bar (Figure 6).

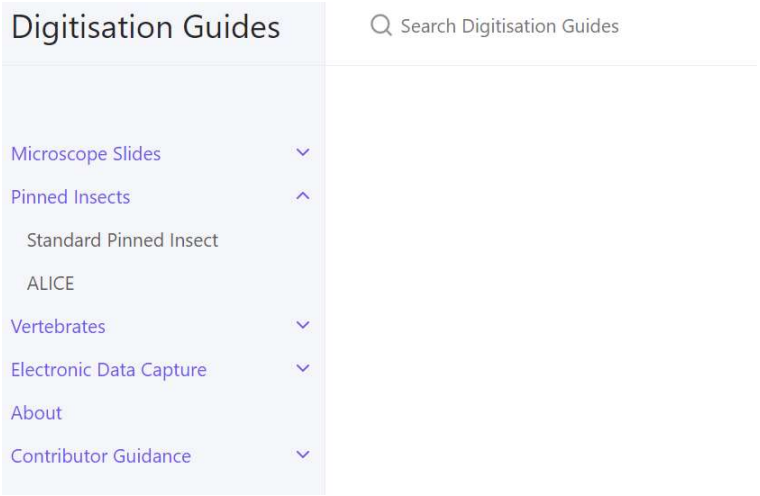

Figure 6: Navigation structure on the prototype website

The digitisation workflow pages on the site are created by adding a markdown file into the github repository. Markdown is a markup language for creating formatted text, and is relatively easy to use for non-technical users. An example of a digitisation page can be seen in Figure 7.

Figure 7: Example digitisation workflow page

The prototype website was used to store the first versions of the standard operating procedures for T3.2.

Digitisation Standard Operating Procedures

SOPs provide clear step-by-step instructions on how to complete a process. Task 3.2 aims to create SOPs for digitisation which can be used by natural history institutions to improve their digitisation capacity through the adoption of best practice.

Business Process Model and Notation 2.0 (BPMN)

The first step in developing the digitisation SOPs was to agree a common template which institutions can use to describe their digitisation workflows. SOPs often include workflow diagrams, which illustrate each step of a process. The use of a common workflow notation for use in DiSSCo digitisation procedures has a number of advantages:

- It helps users to interpret the workflows, as they need to only become familiar with one style of notation.
- Guidance can be produced to help users understand the workflow notation, and to help digitisation managers create their own workflow diagrams.
- Common notation can help users identify similarities and differences between workflows, and therefore see potential opportunities for process improvements.

Business Process Model and Notation 2.0 (BPMN) has been chosen for use in the workflows in the community digitisation manual. BPMN is a standard notation for business process diagrams, which is developed and maintained by the Object Management Group (OMG). OMG is a not-for-profit technology standards consortium. This notation aims to help people communicate business process information internally and externally, through the use of a standard set of workflow elements, and is designed to be understood by non-technical audiences (OMG, 2011).

BPMN is a method which can be used to create diagrams of business processes. It includes graphical elements, or symbols, which can be used to represent steps in a workflow (OMG, 2011). These workflows can be created at different levels of abstraction: from highly detailed step-by-step guides for internal use by employees following a process to more simplified workflows which can be helpful for sharing a process between organisations.

BPMN contains over 100 symbols, however, there are only a few which will be commonly used in digitisation workflows. Basic guidance has been written, which aims to help digitisers to write workflows and to help users interpret the workflows presented in the SOP template.

Guidance:
https://web.archive.org/web/20220114133416/https://lmfrench.github.io/Guidance/BPMN.html

It was agreed most workflows in the guide will be written at a relatively high level of abstraction, generalising detailed organisation processes to a level that can be applied across multiple organisations. For example, many internal workflows will have a high level of detail about data entry into the institution's Collections Management System, and this should be avoided for workflows in the guide as it is unlikely to be applicable across institutions.

Website Template

A template was created for the digitisation SOPs, which would help to ensure consistency in presentation across the website. This template aims to be flexible, so that it can be used across different collection types and digitisation stages.

The template suggests splitting workflows into the five task clusters defined in Nelson *et al.,* 2012 (Pre-Digitisation Curation, Specimen Image Capture, Electronic Data Capture, Georeferencing Specimen Data), with the addition of a 'Preserving and Publishing Data' section. These task clusters are optional, as many of the workflows will not cover all of these areas.

The template also has a section allowing the author to include examples of projects which have followed these workflows, so that they can give more information about how workflows have been applied in practice. Given that the workflows will be written at a high level of abstraction, this section gives an opportunity for a discussion of any challenges or considerations when applying it into a specific organisational context.

There is also a section to include hardware, software and camera setting requirements for the workflows, as well as space to include links to other sources, references, author attributions and version control.

Template:
https://web.archive.org/web/20220114133512/https://lmfrench.github.io/Template.html

Pilot SOPs

A series of pilot SOPs were developed using the BPMN notation and website template. This section gives a brief overview of what is included in each workflow, and links to an archival version of the prototype SOP. These will continue to be developed during T3.2 and the content will be tested with users and amended for clarity where required.

Pinned Insects - ALICE (NHM)

ALICE (Angled Label Image Capture and Extraction) is a custom-built multi-camera setup for high throughput pinned insect specimen digitisation. This method allows for label images to be captured without removing them from the pin.

Website:
https://web.archive.org/web/20220114133041/https://lmfrench.github.io/PinnedInsect/ALICE.html

Spirit and Vertebrate (Dry Preserved) - Bat/Chiroptera (NHM)

This workflow describes the digitisation of the bat collection at the Natural History Museum, London, (NHM) funded by a SYNTHESYS+ Virtual Access Project. The NHM Bat Collection includes skins, skulls and specimens preserved in spirit. Some of the collection was already partially digitised, and this workflow describes some of the challenges that digitisers can face when dealing with inconsistent data quality from past digitisation efforts. This project did not include image capture.

Website:
https://web.archive.org/web/20220114133322/https://lmfrench.github.io/Vertebrates/Bat.html

Microscope Slide (ICEDIG)

This page outlines the semi-automated mass digitisation workflow used by the Natural History Museum, London, to digitise its microscope slide collection. It provides a short summary of the workflow developed as part of the ICEDIG project, with more detail to be found in the Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides paper (Allan et al., 2019).

Website:
https://web.archive.org/web/20220114132918/https://lmfrench.github.io/MicroscopeSlides/MicroscopeSlideMassDig.html

Electronic Data Capture - Transcription (NHM)

Transcription is often the most time-consuming and resource intensive element of a digitisation workflow. This page provides a brief overview of manual transcription, with a focus on project planning.

Website:
https://web.archive.org/web/20220128132750/https://lmfrench.github.io/ElectronicDataCapture/Transcription.html

Herbarium Specimens (RBGE)

This workflow is designed for the digitisation of flat herbarium sheets, undertaken as part of an in-house mass digitisation programme at RBGE. The workflow is based on the concepts outlined in early publications for creating minimal data specimen records. The data element of the workflow results in minimal data records, equivalent to Minimum Information about a Digital Specimen (MIDS) 1, in this first stage of digitisation. The enhancement of these records will then be achieved as part of subsequent digitisation workflows. The physical curation element includes a level of specimen curation and conservation identified as a balance between achieving high throughput rates and maintaining best practice curation standards.

The workflow includes a level of automation to create the data records with associated metadata and to process the image files with associated metadata. The image processing pipeline includes Optical Character Recognition (OCR) which is carried out on all images.

Website:

https://web.archive.org/web/20220114132528/https://lmfrench.github.io/HerbariumSheets/RBGEHerbariumSheet.html

Herbarium Specimens (ULISBOA)

This workflow illustrates part of the procedures adopted by a small university herbarium (about 80.000 specimens) which started the complete digitization and imaging of its herbarium sheets just a few years ago. The team in charge of this project is also reduced, consisting of a curator (part time), IT specialist (part time), digitizer/database operator and herbarium technician. The previous collaboration of an additional technical assistant, who started a prototype database in *FileMaker Pro 3.x* is also noteworthy.

The collection database is presently managed using *Specify 6*. Regarding the hardware, the herbarium is equipped with an imaging station, consisting of an *all-in-one PC*, connected to a wireless *Zebra* barcode reader, one planetary scanner *IS2 eScan*, external drive (1Tb) and a *Zebra* carbon label

printer. The resulting dataset, including (at the moment) over 76.000 records and more than 8.600 images, is available at https://www.gbif.org/pt/dataset/835ac57e-f762-11e1-a439-00145eb45e9a.

The workflow is an adaptation to the local settings of a sequence of workflows published by Nelson *et al*. (2015), developed by the Digitization Group of iDigBio, and available through GitHub (https://github.com/iDigBioWorkflows) and was prepared in the framework of the participation of Instituto Superior de Agronomia in the Research Infrastructure PORBIOTA (www.porbiota.pt).

Website:
https://web.archive.org/web/20220128132647/https://lmfrench.github.io/HerbariumSheets/LISIULisboa.html

Pilot SOP Evaluation

The pilot workflows developed for this milestone will be evaluated in the next stages of the project. Initial feedback has been sought from T3.2 partners, with a number of recommendations on how to improve the SOPs going forward.

The explanatory text written underneath each workflow diagram could be improved. It may help the reader if the diagrams included numbering on key steps which were then explained in detail below the workflow. The format of this numbering should be standardised across workflows.

There is a need to add a section to the SOPs to explain the IT infrastructure requirements for each workflow: for example, the computing requirements and storage needed. This will not be relevant for all workflow diagrams (particularly those focussing on pre-digitisation curation), and could be included as an optional heading in the 'Requirements' section. This section would be particularly helpful for the target audience at ALA Digitisation Maturity Level 1. This need is likely to be addressed in the next milestone for this task (MS3.6), which will focus on Extract, Transform and Load (ETL) procedures.

The structuring of the website navigation and the labelling of workflows also needs consideration. There will likely be submissions of multiple workflows of the same collection type (e.g. herbarium sheet and pinned insect workflows). More information should be included on the top level pages for each collection and digitisation stage to help a user select the most appropriate workflow for their organisation, perhaps in the form of a flow chart.

Additional guidance will be required for authors of SOPs. The current guidance focuses on explaining the BPMN notation, and there will be a need for guidance which gives tips on how to write the diagrams. This includes guidance on the appropriate level of abstraction for workflow diagrams, best practice for the explanatory text, and what type of additional information might be useful to include in the SOP.

The pilot SOP pages will need to be tested with the target audience. Feedback from institutions at Maturity Level 1 and 2 will be invaluable to helping to improve the workflow diagrams and instructions, and this will be sought from national node institutions to inform the final deliverable for this task.

Recommendations and Next Steps

This milestone has described the development of a prototype community digitisation manual and a process for writing SOPs. These resources will continue to be developed during T3.2, and presented in the final deliverable.

Community Digitisation Manual

The digitisation manual developed for this milestone was created as a proof-of-concept in order to assess the viability of using GitHub Pages to host the T3.2 SOPs. This prototype allowed us to quickly develop template pages in an agile way, and to share the content with users. The prototype met all of the 'must have' requirements, and a discussion is now needed with the DiSSCO Technical Team to determine how best to continue to develop the website. DiSSCo has a GitHub account, and the website could be hosted here.

The user stories for the digitisation manual included "As a Digitisation Manager, I want to be able to ask questions about the workflows so that I can apply them in my own institution". The acceptance criteria for this user story could be met by GitHub, as each repository has a Discussion site. However, it may be more appropriate for this to be addressed by the DiSSCo Helpdesk, and T3.2 will work with T2.2 to consider suitable options.

Work is ongoing on a landscape analysis of digitisation workflows and procedures. This landscape analysis will examine journal articles and website resources to identify best practice. These best practices will be used to guide the prioritisation of SOP development, and to make recommendations on areas where there are gaps. It can also be used to help users find relevant published workflows, and will be included on the community digitisation manual website.

We will need to consider options for the maintenance of the digitisation manual once the DiSSCo Prepare task has completed. There will be some requirement for staff resources to maintain this website, both in terms of managing the administration and for providing help and guidance. It will be preferable to have a level of curation of the community-edited content, for example through the labelling of 'best practice' workflows and resources to help guide users towards the most appropriate workflow. Discoverability of the resource also needs to be considered, to help ensure users are both able to find the site itself, and to navigate and find the information they require within the website.

SOPs

This task will continue to develop SOPs, and this will be a particular focus for the next two milestones. MS3.6 will look at Extract, Transform and Load (ETL) procedures, and MS3.7 will create pre-digitisation curation SOPs.

The development of the pilot SOPs identified areas for improvement. Further consideration is required on the format of the step-by-step explanatory text, and the template should include a section on IT infrastructure requirements. More guidance is also required for SOP authors, which should include advice on the level of abstraction - the SOP should allow another organisation to follow the process successfully, but not include too much information that would not be relevant outside of a specific institutional context.

DiSSCo is a multi-lingual infrastructure and if these SOPs are to be implemented, used and adapted then we need to provide them in multiple languages. Crowdin, which is a localisation solution that has GitHub integration, may be a suitable option. This has been successfully used by Bionomia.

Additional guidance will be required for authors of SOPs. The current guidance focuses on explaining the BPMN notation, and there will be a need for guidance which gives tips on how to write the diagrams. This includes guidance on the appropriate level of abstraction for workflow diagrams, best practice for the explanatory text, and what type of additional information might be useful to include in the SOP.

Six SOPs have been presented in this milestone, and this will be user tested with the target audience. We will seek feedback from task partners and members of national nodes, with users invited to participate in semi-structured interviews. They will be asked how easy the workflows are to understand and what areas they need more information on to be able to implement the process. This will lead to a set of recommendations for improving the workflows, as well as helping to prioritise which SOPs should be developed next. We will also seek feedback from national nodes to understand what workflows to prioritise to help improve capacity building among these institutions.

Additional Considerations

There are a number of other components that need to be considered for digitisation projects, and digitisation programs:

- Overall project/programme management (not currently covered in DiSSCo Prepare or previous projects)
- Pre-Digitisation Curation (covered by Subtask 3.2.3)
- Standardised Extract Transform and Load (ETL) procedures (covered by Subtask 3.2.3)
- Digitisation monitoring (covered by Subtask 3.2.4)
- Quality control and assurance at different stages (Nieva de la Hidalga et al., 2019)
- Costs and digitisation rate information (covered by Work Package 4 and by Hardisty et al. (2020b) in the ICEDIG project)

We note that there will be significant differences in the level of documentation for small projects compared to programmes. We have received feedback from our prototype community manual that information related to generic digitisation setups, costs, rates, IT infrastructure, and data generation for each workflow would be useful (P. Brewer, personal communication, 6 January 2022). We will review the relevant recommendations from the DiSSCo Conceptual Design Blueprint (Hardisty *et al.* 2020a), which includes recommendations on digitisation rates and costs, and will work closely with T4.1 “Costbook for DiSSCo” to develop this documentation.

Author Contributions

Contribution types are drawn from CRediT - Contributor Roles Taxonomy

Conceptualization: Laurence Livermore

Formal analysis, Visualization: Lisa French, Pedro Arsénio, Robyn Drinkwater, Rui Figueira, Elspeth Haston

Methodology: Laurence Livermore, Lisa French

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

Writing - original draft: Lisa French, Laurence Livermore

Writing - reviewing & editing: Pedro Arsénio, Frederik Berger, Sofie De Smedt, Rui Figueira, Elspeth Haston, Helen Hardy, Vince Smith, Sabine Von Mering

Additional Contributors

Formal analysis: Robyn Crowther, Ana Raquel Cunha, Kate Holub-Young, Michael Jardine, Phaedra Kokkini, Krisztina Lohonya, Larissa Welton

Writing - reviewing & editing: Robyn Crowther, Phaedra Kokkini, Krisztina Lohonya, Jennifer Pullar, Larissa Welton

References

Allan, L. E., Price, B.W., Shchedrina, O., Dupont, S., Livermore, L., & Smith, V. S. (2019). Mass-imaging of microscopic and other slides. Zenodo. https://doi.org/10.5281/zenodo.336448

Devenish, L. (2019) Day in the life of a digitiser: Digital Collections Programme. Blogs from the Natural History Museum. Available at: https://naturalhistorymuseum.blog/2019/08/29/day-in-the-life-of-a-digitiser-digital-collections-programme/ [Accessed 03-12-2021]

Dupont S, Price BW (2019) ALICE, MALICE and VILE: High throughput insect specimen digitisation using angled imaging techniques. Biodiversity Information Science and Standards 3: e37141. https://doi.org/10.3897/biss.3.3714

Drinkwater R, Cubey R, Haston E (2014) The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels. PhytoKeys 38: 15-30. DOI: https://doi.org/10.3897/phytokeys.38.7168

Groom, Q., Dillen, M., Hardy, H., Phillips, S., Willemse, L. & Zhengzhe, W. (2019) Improved standardization of transcribed digital specimen data, Database, Volume 2019, baz129, https://doi.org/10.1093/database/baz129

Hardisty A., Saarenmaa, H., Casino, A., Dillen, M., Gödderz, K., Groom, Q., Hardy, H., Koureas, D., Nieva de la Hidalga, A., Paul, D.L., Runnel, V., Vermeersch, X., van Walsum, M. & Willemse, L. (2020a) 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

Hardisty A, Livermore L, Walton S, Woodburn M, Hardy H (2020b) Costbook of the digitisation infrastructure of DiSSCo. Research Ideas and Outcomes 6: e58915. https://doi.org/10.3897/rio.6.e58915

Hardy H, Knapp S, Allan EL, Berger F, Dixey K, Döme B, Gagnier P-Y, Frank J, Haston EM, Holstein J, Kiel S, Marschler M, Mergen P, Phillips S, Rabinovich R, Sanchez Chillón B, Sorensen MV, 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: e50354. https://doi.org/10.3897/rio.6.e50354

Nieva de la Hidalga, Abraham, van Walsun, Myriam, Rosin, Paul, Sun, Xianfang, & Wijers, Agnes. (2019). Quality Management Methodologies for Digitisation Operations. Zenodo. https://doi.org/10.5281/zenodo.3469521

Haston, E, Cubey, R & Harris, DJ (2012). Data concepts and their relevance for data capture in large scale digitisation of biological collections. International Journal of Humanities and Arts Computing, 6:1-2, 111-119. DOI: https://doi.org/10.3366/ijhac.2012.0042

iDigBio (2022). Workflow Modules and Task Lists https://www.idigbio.org/content/workflow-modules-and-task-lists [13 Jan 2022]

Kalms, B. (2012) Digitisation: A strategic approach for natural history collections https://www.ala.org.au/wp-content/uploads/2011/10/Digitisation-guide-120326.pdf

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

Nelson, G., P. Sweeney, L. E. Wallace, R. K. Rabeler, D. Allard, H. Brown, J. R. Carter, et al. 2015. Digitization workflows for flat sheets and packets of plants, algae, and fungi. *Applications in Plant Sciences* 3: 1500065.

Paterson, G. et al. (2016) iiCollections – Digitising the British and Irish Butterflies in the Natural History Museum, London. Biodiversity Data Journal 4: e9559. https://doi.org/10.3897/BDJ.4.e9559

Price, Benjamin W., Steen Dupont, Elizabeth L. Allan, Vladimir Blagoderov, Alice J. Butcher, James Durrant, Pieter Holtzhausen, et al. 2018. ALICE: Angled Label Image Capture and Extraction for High Throughput Insect Specimen Digitisation. OSF Preprints. November 5. https://doi.org/10.31219/osf.io/s2p73

Object Management Group (2011) Business Process Model and Notation Version 2.0. Available at: https://www.omg.org/spec/BPMN/2.0/PDF [Accessed 31-12/2021]

Walton S., Livermore L., Dillen M., De Smedt S., Groom Q., Koivunen A. & Phillips S. (2020) A cost analysis of transcription systems. Research Ideas and Outcomes 6: e56211. https://doi.org/10.3897/rio.6.e56211

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Pullar, J (2019) Pinned Insect Digitisation. Digital Collections Programme. Available at: https://www.nhm.ac.uk/our-science/our-work/digital-collections/digital-collections-programme/pinned-insect-digitisation.html [Accessed 03-12-2021]

APPENDIX

Meeting Date	Type	Outcomes
2021-01-18	All Hands	Overview of Task 3.2 Discussed Task 3.2 plan Questions and discussion session
2021-05-25	Task Meeting	Agreed subtask leads Agreed community manual/'recipe book' of workflows required Agreed to review of available resources and capture information in matrix of collection type vs digitisation process
2021-07-13	Task Meeting	Reviewed approach to capturing existing resources Agreed to share institutional workflows to inform decision on workflow format
2021-09-14	Task Meeting	Subtask leads shared work plans
2021-10-12	Task Meeting	Discussed DiSSCo Design Blueprint Recommendations (Hardisty *et al.,* 2020a) Agreed website resource required to share workflows
2021-10-20	Knowledgebase meeting	Discussed requirements for website resource/community digitisation manual. Agreed Knowledgebase would not be best placed to host the website, but could be used to store documents
2021-11-09	Task Meeting	Agreed audience for community digitisation manual Agreed BPMN would be used to create workflows Agreed most workflows should have a high level of abstraction, but might link through to other workflows showing detail. Agreed to create prototype website in GitHub Pages
2021-12-14	Task Meeting	Update on Milestone Progress Feedback on prototype website
2022-01-11	Task Meeting	Reviewed milestone Agreed next steps for SOPs and digitisation website Subtask task planning for MS3.6: ETL procedures.