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

MS3.6 Best Practice Standardised Extract, Transform and Load (ETL) procedures

Author(s)

Esko Piirainen, Zhengzhe Wu, Lisa French, Sofie De Smedt, Rui Figueira, Pedro Arsénio, Elspeth Haston, Laurence Livermore

Identifier of the author(s)

Esko Piirainen
Zhengzhe Wu
Lisa French: https://orcid.org/0000-0001-7279-8582
Sofie De Smedt: https://orcid.org/0000-0001-7690-0468
Rui Figueira: https://orcid.org/0000-0002-8351-4028
Pedro Arsénio: https://orcid.org/0000-0003-3860-9789
Elspeth Haston: https://orcid.org/0000-0001-9144-2848
Laurence Livermore: https://orcid.org/0000-0002-7341-1842

Affiliation

Finnish Museum of Natural History (Luomus): Esko Piirainen, Zhengzhe Wu
Natural History Museum, London: Lisa French, Laurence Livermore
Meise Botanic Garden: Sofie De Smedt
Universidade de Lisboa: Pedro Arsénio, Rui Figueira
Royal Botanic Garden Edinburgh: Elspeth Haston

Contributors

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

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

Abstract
-

Content keywords

Project reference
DiSSCo Prepare (GA-871043)

WP number
WP3

Project output
Milestone report

Deliverable/milestone number
MS3.6

Dissemination level
Public

Rights

License
CC0 1.0 Universal (CC0 1.0)

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 – MS3.6 Best Practice Standardised Extract, Transform and Load (ETL) procedures

Esko Piirainen, Zhengzhe Wu, Lisa French, Sofie De Smedt, Rui Figueira, Pedro Arsénio, Elspeth Haston, Laurence Livermore

LUOMUS
FINNISH MUSEUM OF NATURAL HISTORY

Index

01 Introduction

DiSSCo (Distributed System of Scientific Collections, https://www.dissco.eu/) is a pan-European Research Infrastructure (RI) that among other things aims to create a digitisation infrastructure for natural science collections. An overview of its infrastructure is described in its conceptual design blueprint (Har2020). DiSSCo is a new world-class research infrastructure for natural science collections. It is estimated that in order to digitise the majority of important public natural history collections in the coming decades, up to 40 million specimens may need to be digitised each year. Digitised specimens, each up to hundreds of megabytes, will be created at different distributed digitisation facilities across Europe. The large amount of data generated at the digitisation stations will go through various Extract, Transform, Load (ETL) procedures before appearing in the Collection Management System (CMS) and/or data sharing/publication portals. ETL procedures are critical in the digitisation process, and it is therefore necessary to provide best practice on standardised ETL procedures to facilitate and optimise the digitisation process at DiSSCo institutions.

This project report was written as a formal Milestone (M3.6) of the DiSSCo Prepare Project (https://www.dissco.eu/dissco-prepare/). The following text is the formal description (Subtask 3.2.2) from the DiSSCo Prepare project's Description of the Action (workplan):

Subtask 3.2.2 Standardised Extract Transform and Load (ETL) procedures. *Handling metadata and images during digitisation involves many transformations, as information is modified and held in various temporary (staging) environments, before reaching the institutional collection management Systems (CMS) and being made accessible through public portals.*
This subtask will document best practices for these processes, where necessary including the computational workflows required to support data transformations.

The best practice in this work will help enhance natural science collection specimen digitisation capacity across DiSSCo partners and the DiSSCo national nodes. Together with other deliverables of WP3.2, this document will form a Community Digitisation Manual. This work was done by the following task partners in this task:

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

We first give an overview of the goal and the scope of this Best Practice Document (BPD). Secondly, digitisation workflows from partner institutions and other related work were reviewed to find the potential ETL procedures in each part of the workflow. Thirdly, we made a list of the best practice recommendations. This report concludes a discussion.

2. Overview of the Work

2.1 Scope

Extract, Transform, Load (ETL) is a higher concept that can mean moving any data from place A to B. The term ETL is most commonly used in the context of moving data from multiple databases into a single data warehouse for analytics. In the context of a natural history collection digitisation workflow, ETL processes can be considered to start from getting data from digitisation stations up to the point where the data is stored in the Collection Management System (CMS) and/or data sharing/publication portals. After the ETL process, there may be many further steps (often handled for example by the CMS) concerning data validation, cleaning, annotation, crowdsourcing, using AI-based methods on processing, mining, enriching the data, and finally sharing it to 3rd party platforms such as Global Biodiversity Information Facility (GBIF). Those steps are not covered by this Best Practice Document (BPD), except for AI-based methods which may be applied during data transformation during the ETL procedures. Long-term archiving may happen as a part of the initial ETL process or after: we have included it in this BPD as it is an important data transformation. Figure 1 shows the generalised data flow and the scope of the BPD on the ETL procedures in the digitisation process.

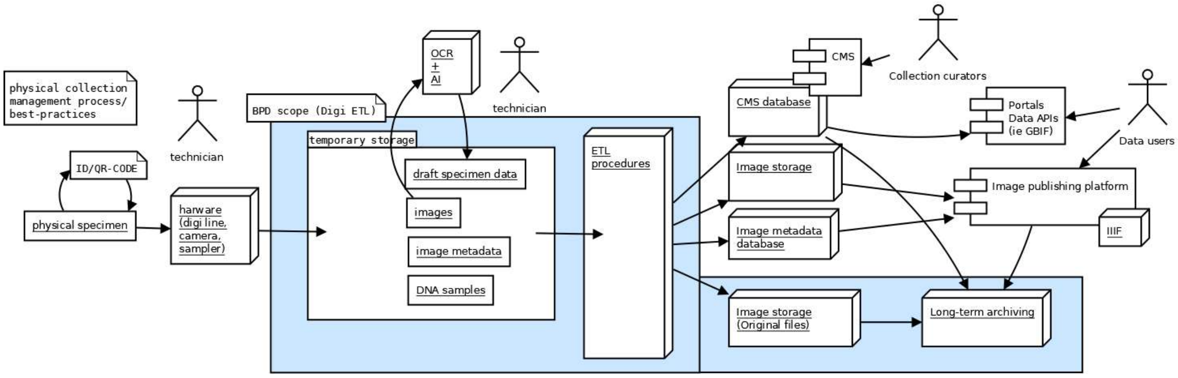


Figure 1: Generalised view on how information from physical specimens can reach data users as a result of digitisation. (OCR=Optical character recognition; AI=Artificial Intelligence; ETL=Extract-Transform-Load procedures; CMS=Collection Management System; API=Application Programming Interface; GBIF=Global Biodiversity Information Facility; IIIF=International Image Interoperability Framework; Long-term archiving=medium where data is stored "forever")

The level of specimen digitisation varies between different institutions and digitisation projects. There can be a variety due to the following properties:

- Collection (preservation) types (insects, herbarium sheets, mosses, microscope slides, fossils, rocks, ...)
- Collection sizes (from few specimens to millions of individual specimens)
- Digitisation media (textual data, images, CT scans/3D models, DNA barcodes, ...)
- Organisational maturity (from "disorganised" to higher levels of institutional organisation)
- Organisation size (many teams vs individual people)
- Technical advancement (from manual to semi-automated to almost fully automated and from human made labour to AI and robotics based methods)

Different digitisation levels demand different approaches. For example, for massive insect collections, a high level of automation is needed, but setting up such an infrastructure for a small rock collection would not be ideal. Thus, **it is not possible to propose a single standardised procedure**. Instead, this BPD lists a number of recommendations. Each party can use the recommendations to evaluate if they apply to their particular digitisation projects.

Furthermore, since institutions operate inside very different infrastructures, this BPD does not recommend any particular software, service providers or other concrete methods on how to implement the recommendations. **The recommendations should be considered as goals** that institution staring to set up/improve their digitation infrastructure should try to meet.

2.2 Audience

The main target audience of the Community Digitisation Manual has been agreed to be institutions that are at the beginning of building up their digitisation process. The recommendations in this BPD are categorised according to their level of advancement from very basic/must have recommendations to more advanced recommendations. To be able to implement the requirements, the organisation must have a certain level of technological capacity and resources (servers, infrastructure, etc.): for example a single individual can not successfully implement recommendations of this BPD. More details of the agreed target audience and ALA Digitisation maturity model is in DiSSCo Prepare Milestone 3.5

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, because organisations at this level would require detailed guidance and bespoke support.

2.3 BPD Template

To assess the best practice is this work, we use a template defined by Alwazae et al. (Alw2015) to formalise the goals and maturity of this BPD as in Table 1.

Table 1. Overview of this BPD based on the template defined by Alwazae et al. (Alw2015).

Summary	This BP helps implement Extract-Transform-Load (ETL) procedures for data from digitisation stations to their final publishing platforms and how the ETL procedures should fit in the overall DiSSCo infrastructure.
Goal	Applying the BP ensures data is not lost, improves efficiency and makes the data more interoperable in the digitisation process.
Means	Successfully applying this BP requires IT specialists with skills in (1) programming, (2) server administration, (3) image processing AND availability of cloud based servers and sizable storage capacities.
Cost	Technology costs: Institutions may have access to "free" services provided by academic research infrastructures nationally or internationally, or their parent university or other organisations may be able to provide the needed technology. However, someone ultimately has to pay for the needed resources. In terms of computation power, a minimal viable digitation ETL process does not incur great costs. Use of advanced methods like OCR or AI techniques require more computation resources. Storing and long-term archiving digital material (images, 3d models) and the cost of moving the

	material in and out of storage can be very costly, up to tens of thousands EUR/year. The costs are examined in more detail by DiSSCO Prepare project WP4.1.
Barriers	Obstacles or problems that may occur before, during, and after applying the BP should be documented here after demonstration of success. (None so far / 2022-04)
Barrier Management	Procedures to follow if certain obstacles or problems are encountered should be documented here after demonstration of success. (None so far / 2022-04)
Acceptability	This document will be reviewed by the leading experts participating in the DiSSCO Prepare project.
Usability	Feedback to which degree this BP is easy to use should be documented here after demonstration of success. (None so far / 2022-04)
Comprehensiveness	This document does not describe a comprehensive BP. Instead, various individual recommendations are listed. We have selected the most important recommendations for institutions that are at the beginning of building up their digitisation process.
Prescriptiveness:	This BP offers concrete proposals for solving the problems; however the underlying infrastructures vary so much that only examples of actual implementations can be provided.
Coherence	This BP does not form a coherent unit: certain parts only apply to certain types of digitisation efforts.
Consistency	This BP is consistent with existing knowledge and vocabulary used in digitisation of the natural science collection sector and knowledge domain as leading experts in the field have participated and reviewed the living document.
Demonstration of Success	See end of this document (None so far / 2022-04)

3. Review of Digitisation Workflows

To identify the ETL procedures in the digitisation processes, we did a review of digitisation workflows from available publications, reports, and project partners' documents. The extensive list of related literature is in Appendix IV and V. We extracted steps/procedures from those workflows and tried to list them into three categories regarding ETL procedures (before, within, and after it). Those lists can be seen from Appendix VI, VII, and VIII respectively.

The outcome lists of digitisation steps are not recommended workflows, or even functional workflows, rather it is a union of all potential workflows in digitisation of natural history collections. It is a tool used in mapping the landscape so that BP recommendations on ETL can be created from the different steps/processes/procedures.

3.1 Infrastructure

Infrastructure is the one of the most fundamental core parts in the digitisation processes. It is the hardware and services that the digitisation processes are built on and it requires careful planning. Therefore, good practice should be followed in setting up the digitisation infrastructure. Within the scope of the ETL processes, the infrastructure involves the local digitisation station, remote servers, CMS system, data backup system, etc. In general, there are the following components:

Local storage at the digitation site to which the digitation line/other hardware connects to - typically a local machine (not a server).

Staging area to which raw digitised material is transferred from the local machine for processing - NOT meant to store data for longer periods of time - server based.

Image archive to which large original TIFF (etc) files are stored - cloud or server based.

Publishing platform file storage (image server) to which ready material is transferred so that it is accessible from the web - cloud or server based.

CMS/data repository (relational or other database) to which specimen data, image metadata etc is stored - cloud or server based.

Backup storage to which resources from image archive, publishing platform are periodically backed up)- cloud or server based

3.2 Organisational Models

Different institutions have different organisation structures for digitisation. Some may have their own in-house IT development staff. Some may have outsourced digitisation to contractors. For example, at the Finnish Museum of Natural History (Luomus), the digitisation workflow involves a digitisation team and three IT staff, one digitisation software developer/manager, one server administrator, and a data manager (Luo1). There are digitisers and one IT manager at Herbarium João de Carvalho e Vasconcellos (LISI), Universidade de Lisboa,. For the Natural History Museum (NHM) London, digitisers and the data management team are involved in the digitisation workflows. Different models have their advantages and disadvantages, which should be well considered before the starting of digitisation activities based on institutions' own situations. We made BP recommendations in the next chapter.

3.3 Data Models

There are variants in the data models in different workflows. At Luomus (Luo1) data model, the specimen data does not have direct links to specimen’s media files; instead in their CMS, the associated media are queried in real time from Image-API. Furthermore the information is available in the CMS search engine (ElasticSearch) and FinBIF national data warehouse. At Herbarium João de Carvalho e Vasconcellos (LISI), Universidade de Lisboa, the links to media data are stored in the CMS Specify 6, which manages specimen data. Images are managed with Specify Web Asset Server. A Digital Asset Management (DAM) system is used at the NHM London to store digital assets, including images uploaded to Emu CMS (Sco2019). Via the DAM API, specimen images at a suitable downscaled web resolution are displayed at the NHM data portal.

3.4 Identified Workflow Procedures

3.4.1 Pre-ETL Workflows

The pre-ETL workflows are mostly on the digitisation stations, as shown in the Appendix 'Pre-ETL workflows'. Most of those workflows are related to the barcoding of the specimens, imaging, image quality control, image processing, metadata generation and upload. It also involves the data transmission from the digitisation station to the staging area, CMS, image publishing platform, and data backup. The objects in those workflows are identifiers, images, image metadata, and specimen data. There are different actions, such as manual, semi-automated, automated, on those workflows. Some of the actions require manual work as a must.

3.4.2 ETL Workflows

The Appendix 'ETL workflows' lists the examples of ETL workflows. ETL workflows are mostly in the staging area. They are doing image pooling, data quality control, file renaming, data export and publishing, image conversion, and data backup. The data will go to the CMS, image publishing platform, image archive, in different workflow steps. Most of the workflows here are done semi-automated or fully automated. Automating those workflows as much as possible will increase the efficiency of the digitisation process and minimise the potential risks of human mistakes.

3.4.3 Post-ETL Workflows

In the Appendix 'Post-ETL workflows', it lists all those post-ETL workflows we found after reviewing the related work. They are mostly run on CMS, Image archive, data backup storage, and long-term data archive. They are related to data backup, data linkage, data enrichment, and data long-term preservation. There are different actions, such as manual, semi-automated, automated, on those workflows. Some of the actions here still require manual work as a must.

4. Best Practices

This chapter lists the Best Practices (BP) recommendations we have been able to determine in the reviews of digitisation workflows done in Chapter 3. We used a customised template for the BP in this work to follow. This will unify the formats of all the BPs and make it easy to follow. An example of the template is in Table 2 and the explanation of the items can be found in Table 3.

Table 2. Template of 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

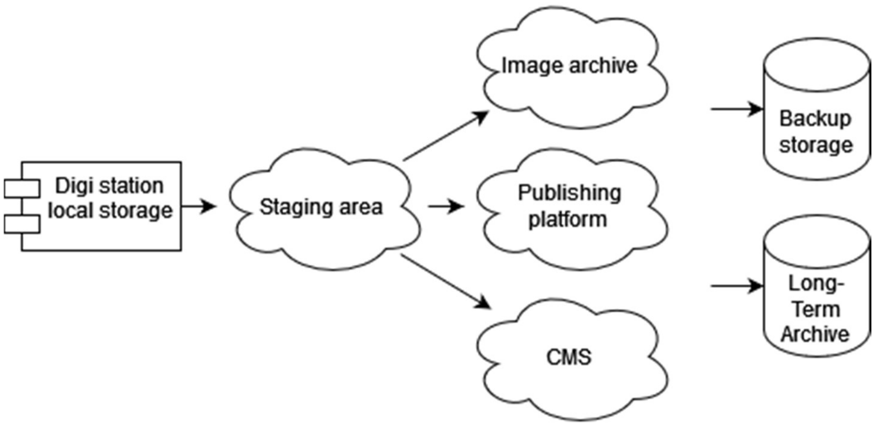
Discussion	Rationale behind the recommendation
Implementation example	One or few references/examples on how the recommendation has been implemented in practice if applicable
References	Link, Ref

Table 3. Explanations of the items in Table 2 of Template of Best Practices

Id	To make it easier to communicate about an individual recommendation
Level	Level: how demanding the recommendation is BASIC: A fundamental goal that everyone doing digitisation should try to fulfil ADVANCED: Next steps in automating and improving performance STATE-OF-ART: New upcoming techniques that should perhaps not be attempted to take into account at first
Use case	An use case which acts as a motive for the recommendation
Best practice recommendation	Procedure to follow/task to accomplish that fulfils the use case
Discussion	Discussion about the rationale of the recommendation Implementation example
Implementation example	One or few references/examples on how the recommendation has been implemented in practice if applicable
References	Links to external documentation or publication which is the source of the use case or recommendation and/or to implementation example

The recommendations in this chapter are given as examples. During the course of DiSSCo Prepare and DiSSCo projects, the BDP is expanded, reviewed and maintained. The final location of the recommendations will be a dynamic website (opposed to a static, frozen document). The website is located at https://dissco.github.io/. The previously linked page is updated by making pull requests to a Git repository: https://github.com/DiSSCo/dissco.github.io. (Read more on Chapter 5 on how the recommendations are maintained and reviewed in the future). The pages are 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.

4.1 Infrastructure Recommendations

Id	INFRA1
Level	BASIC (+STATE-OF-ART)
Use case	**As** digitation manager **I want** no significant data loss to occur and have a reliable system **so that** the digitisation process is not delayed
Best practice recommendation	Your digitation/ETL/publishing/CMS infrastructure should generally have the following components **Local storage** at the digitation site to which digitation line/other hardware connects to - typically a local machine (not a server) **Staging area** to which raw digitised material is transferred from the local machine for processing - NOT meant to store data for longer periods of time - server based **Image archive** to which large original RAW/TIFF (etc) files are stored - cloud or server based, possibly tape based. **Publishing platform** file storage (image server) to which ready material is transferred so that it is accessible from the web - cloud or server based **CMS/data repository** (relational or other database) to which specimen data, image metadata etc is stored - cloud or server based **Backup storage** to which resources from image archive, publishing platform are periodically backed up (see recommendation INFRA2) - cloud or server based STATE-OF-ART: **Long-term archive** to which all data is eventually replicated to be stored "forever" (see recommendation INFRA3)  Digi station local storage Staging area Image archive Publishing platform CMS Backup storage Long-Term Archive

Id	INFRA1
Discussion	**Local storage:** Data is not meant to stay for a long time at the local digitisation station. It should be moved daily or at least weekly forward. Loss of the data stored in these stations does not incur significant data loss. Setting up the environment again may take a long time and mean delays in the digitisation process. STATE-OF-ART: Docker (or other container environment or VirtualBox) based environment is recommended so that it is quickly set up on any new local computer. The idea of containerising the environment is that all required software is installed in the container, and the user just needs to run the container instead of starting with a long list of software to install and configure. This however may not always be possible because of software licences etc. **Staging area:** ETL procedures may require computing power which is best done on server / computing clusters rather than on the local machine; procedures are automated and software driven so ease of deploying new versions is a benefit. State-of-art environment would for example be a Kubernetes container cluster to which different ETL process steps are deployed as individual services/pods and co-operate to provide the ETL procedure. A test environment exists where software is tested before being put to production. **Image archives** should be cloud based to prevent data loss. Hard disk failures are common, which can be alleviated by running a RAID disk server. However, we do not recommend institutions to run their own disk servers or any other servers, as cloud based services are more cost efficient, professionally managed and data loss is almost impossible (except for human error - so backups are still needed). It is a good idea to separate the live-publishing server data storage (containing smaller JPGs etc) and the original raw data (TIFF etc). This allows for example to use a faster disk for publishing. Furthermore, as data in image archives is not needed often, it does not need to be accessible from the internet. It can be for example an object storage database instead of a conventional file system. In case of very large datasets, it may be the case that the image archive needs to be a tape based solution, and the images are fetched from tape and copied to another environment on need basis. **Publishing platform file storage:** Uptime and performance are important here as it is the prevention of data loss (which causes downtime). We recommend a cloud based service for those above-mentioned reasons. **CMS/data repository:** Data loss in your CMS database would be catastrophic. It needs to be professionally administered and backed up. Cloud based solutions are a must. Databases contain text and do not typically take much space. Regular backups should be done in professional manners.

Id	INFRA1
	Backup storage: Even if original data is located on cloud based servers, data loss can occur as a result of human error. It is problematic to find another large enough place to put your biggest data: finding a suitable place for the image archive can be difficult and for backup there would be a second location, as having data twice on the same service doesn't quite fit the need. If no other solution can be found, image archive and backup storage can reside in the same service, which at least helps in case of human-made accidental deletion. **Long-term archive (LTA):** it would be the third place your data resides. It doesn't always fulfil the function of backup storage, as data is stored to LTA in formats that are designed to be ever-lasting and may be somehow modified as a result. It might not be easy to recover data from LTA as getting lots of data out from LTA is not typically what they are designed for. LTA is almost impossible to implement by your own institution, so you should seek research infrastructures that can provide the service for you. We have marked LTA to be "STATE-OF-ART" (very demanding) using this BPD's three level scale. It is not something you should try to set up first.
Implementation example	Finnish Museum of Natural History (Luomus) **Local storage:** Helsinki University IT centre provides local workstations, administrates security, network, user accounts etc **Staging area:** Finnish IT Centre for Science (CSC) provides virtual servers (cPouta; OpenStack based) **Image archive:** CSC research data storage service (IDA) - for even larger 3d scans in the future CSC object storage (Allas) providing space in petabytes and not based on conventional file system **Publishing platform file storage:** CSC virtual server mounted disk (cPouta; OpenStack based) **CMS/data repository:** Helsinki University IT centre provided Oracle database (running on their OpenStack based virtual server environment) **Backup storage:** For publishing platform images: Helsinki University provided disk; for Image archive: none so far **Long-term archive:** Not yet implemented; will be at CSC provided national service (Digital Preservation Service (DPS))
References	Luo1, Luo2

Id	INFRA2
Level	BASIC
Use case	**As a** digitation manager **I want** no significant data loss to occur and have a reliable system **so that** the digitisation process is not delayed
Best practice recommendation	Use traditional hard disks instead of SSD disks on local storage of the digitisation stations.
Discussion	On digitation stations, the I/O access to the storage is usually quite high especially for the high-throughput mass digitisation. High volumes of data are frequently written to the disk from the imaging devices, read to transfer the data to the staging area, and then deleted. SSD disks have a limited number of reads they can do and are more expensive when compared to the traditional hard disks.
Implementation example	Finnish Museum of Natural History (Luomus) **Local storage:** Traditional hard disks are used at the digitisation stations of the mass digitisation systems.
References	Luo1

Id	INFRA3
Level	BASIC
Use case	**As a** digitation manager **I want** no significant data loss to occur and have a reliable system **so that** the digitisation process is not delayed
Best practice recommendation	Implement automated, periodical backup to cloud based backup storage.
Discussion	A second data storage for data backup is necessary to prevent potential human errors and system hardware failures. With the development of digitisation techniques, the size of individual images is getting bigger and the number of output is also increasing considerably. However, the staging area usually has limited storage. Transfer the original RAW/TIFF images and 3D scans to the backup storage and only keep compressed or low resolution versions of images if needed. There are regional, national, and also commercial services available. When choosing the service, you need to

Id	INFRA3
	consider several factors, such as the data privacy and the location of the storage.
Implementation example	Finnish Museum of Natural History (Luomus) Images and their metadata are backed up at CSC research data storage service (IDA). For larger 3D scan data, CSC object storage (Allas) is planned to use. Specimen data is stored in Oracle database which is backed up by Helsinki University IT centre.
References	Luo1, Luo2

Id	INFRA4
Level	STATE-OF-ART
Use case	**As a** digitation manager **I want** no significant data loss to occur and have a reliable system **so that** digitisation process is not delayed
Best practice recommendation	Implement the data archive for long-term data preservation.
Discussion	Long-term archives are used for preserving the data for a very long time. There are requirements on the data types and also the metadata formats. It is usually stored offline and may not be suitable for quick data recovery. LTA is almost impossible to implement by your own institution, so you should seek research infrastructures that can provide the service for you.
Implementation example	Finnish Museum of Natural History (Luomus) It is planned to use national service from CSC (Digital Preservation Service (DPS))
References	Luo1, Luo2

Ideas for further recommendations:

- INFRA4: Information security recommendations...
- INFRA5: Clean up of digitation stations

4.2 Organisational Recommendations

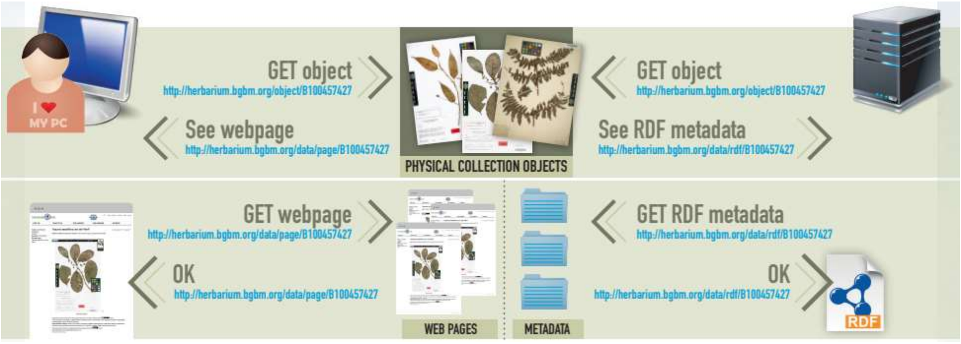
Id	O1, O2
Level	BASIC
Use case	**As a** museum director **I want** to use limited monetary resources efficiently **so that** I can provide best value to society.
Best practice recommendation	O1: Automate recurrent routine tasks as much as possible as part of the ETL process. O2: Employ/acquire one or few software developers instead of adding more digitisation staff to speed up digitisation.
Discussion	Software development is expensive, but spending development resources in automating tasks will eventually save money by reducing staff costs (or allow using those staff more efficiently).
Implementation example	Instead of having staff manually create thumbnails with an image editor, develop an image service that does the job; use existing image libraries available (such as ImageMagick).
References	All2019

Id	O3
Level	BASIC
Use case	**As a** digitisation manager **I want** to prioritise digitisation efforts based on scientific criteria instead of existing procedures **so that** I can provide that information which is most needed to research
Best practice recommendation	O3: Maintain sufficient in-house skills in IT (software development and server administration)
Discussion	More subjective than most recommendations. Digitisation is a continuously changing field where advances are continuously made. New kinds of digitisation projects start and end as digitisation moves to different types of collections. This means changes to existing processes are needed often. Buying services from an outsourced party may not be flexible enough. On the other hand, should an excellent partner exist, they could be able to keep up

Id	O3
	to technological advancement better than small institutions own IT staff. Ordering software and services from an outside partner requires IT skills and knowledge so you can order and explain to the tech people what exactly is the service you need. TODO: Weight PROs and cons of the approaches
Implementation example	
References	

4.3 Identifier Recommendations

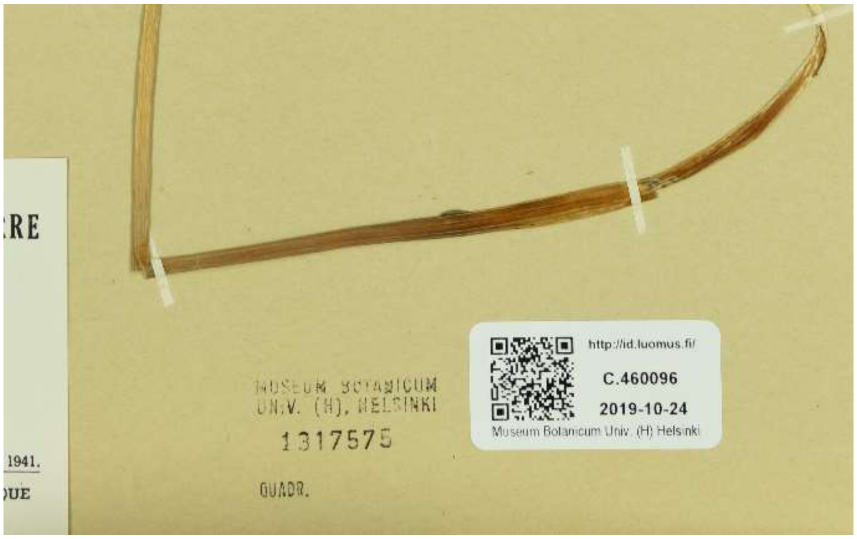
Id	ID1
Level	BASIC
Use case	**As a** digitisation manager **I want** to have the specimens with persistent identifiers **so that** I can make the digitised specimens retrievable through the whole data lifecycle
Best practice recommendation	ID1: CETAF Stable Identifiers and OpenDS identifier minting
Discussion	The digitised specimen data has to be accessible through the whole data lifecycle, not only in the ETL processes in the digitisation but also in the publishing and CMS portals. All specimens have to be assigned with Persistent Identifiers (PID). Different technologies are available for PID, such as LSIDs, DOIs, GBIF ‘Triple IDs’ and HTTP URIs. The CETAF identifier system is based on HTTP-URIs and Linked Data principles. The system has been successfully implemented in 14 member institutions, which are listed in the following implementation examples. When using the CETAF identifier system, the institution can use their domain name in the identifiers, such as http://id.luomus.fi/C.460096 from Luomus. In such cases, the domain names are critical and have to be maintained for long-term by the institution. There is the possibility to use centralised systems such as the services of PURL.org or DOI.org to avoid domain name changes in the URIs.

Id	ID1
	 I ❤ MY PC GET object http://herbarium.bgbm.org/object/B100457427 See webpage http://herbarium.bgbm.org/data/page/B100457427 PHYSICAL COLLECTION OBJECTS GET object http://herbarium.bgbm.org/object/B100457427 See RDF metadata http://herbarium.bgbm.org/data/rdf/B100457427 GET webpage http://herbarium.bgbm.org/data/page/B100457427 OK http://herbarium.bgbm.org/data/page/B100457427 WEB PAGES METADATA GET RDF metadata http://herbarium.bgbm.org/data/rdf/B100457427 OK http://herbarium.bgbm.org/data/rdf/B100457427 RDF
Implementation example	**Botanischer Garten und Botanisches Museum Berlin** Example: https://herbarium.bgbm.org/object/B100277113 Catalogue: https://ww2.bgbm.org/Herbarium/ Redirect to machine-readable representation: yes Finnish Museum of Natural History, Helsinki Example: https://id.luomus.fi/GL.749 Catalogue: no Redirect to machine-readable representation: no **Institute of Botany, Slovak Academy of Sciences, Bratislava** Example: https://ibot.sav.sk/herbarium/object/SAV0001234 Catalogue: https://ibot.sav.sk/herbarium Redirect to machine-readable representation: no (no redirection by passing rdf header, but rdf is accessible at https://ibot.sav.sk/herbarium/data/SAV0001234.rdf) **Museum für Naturkunde, Berlin** Example: https://coll.mfn-berlin.de/u/ZMB_Orth_BA000061S01 Catalogue: no Redirect to machine-readable representation: yes **Muséum national d'histoire naturelle, Paris** Example: https://coldb.mnhn.fr/catalognumber/mnhn/ec/ec32 Catalogue: https://science.mnhn.fr/all/search Redirect to machine-readable representation: yes **Naturalis Biodiversity Center, Leiden**

Id	ID1
	Example: https://data.biodiversitydata.nl/naturalis/specimen/RMNH.AVES.110103 Catalogue: https://bioportal.naturalis.nl/ Redirect to machine-readable representation: yes **Natural History Museum, London** Example: https://data.nhm.ac.uk/object/a9bdc16d-c9ba-4e32-9311-d5250af2b5ac Catalogue: https://data.nhm.ac.uk/ Redirect to machine-readable representation: yes **Natural History Museum - University of Oslo** Example: https://purl.org/nhmuio/id/41d9cbb4-4590-4265-8079-ca44d46d27c3 Catalogue: https://nhmo-birds.collectionexplorer.org/ Redirect to machine-readable representation: yes **Royal Botanic Garden Edinburgh** Example: data.rbge.org.uk/herb/E00421509 Catalogue: https://elmer.rbge.org.uk/bgbase/vherb/bgbasevherb.php Redirect to machine-readable representation: yes **Staatliches Museum für Naturkunde Stuttgart** Example: https://col.smns-bw.org/object/S10000227722006 Catalogue: https://www.smns-bw.org/db/datenbank.php Redirect to machine-readable representation: no **Staatliche Naturwissenschaftliche Sammlungen Bayerns** Example: https://id.snsb.info/snsb/collection/97112/153455/93009 Catalogue: https://www.snsb.info/dwb_biocase.html Redirect to machine-readable representation: yes **Zoologisches Forschungsmuseum Alexander Koenig, Bonn** Example: https://id.zfmk.de/collection_ZFMK/2003261 Catalogue: https://www.collections.zfmk.de/ Redirect to machine-readable representation: yes (https://herbal.rbge.info/?uri=https://id.zfmk.de/collection_ZFMK/2003261)

Id	ID1
	Botanic Garden Meise Example: https://www.botanicalcollections.be/specimen/BR0000008422330 Catalogue: https://www.botanicalcollections.be/#/en/home Redirect to machine-readable representation: yes **Royal Museum for Central Africa** Example: https://darwinweb.africamuseum.be/object/RMCA_Vert_2011.003.P.1885-1898 Catalogue: https://darwinweb.africamuseum.be/search_specimens Redirect to machine-readable representation: no
References	Gün2017, CETAF

Id	ID2
Level	BASIC
Use case	**As a** collection manager **I want** to find the specimen data **so that** I can curate specimens effectively
Best practice recommendation	ID2: Use a standard two-dimensional QR/matrix code embedding the persistent identifiers to barcode the specimen
Discussion	The Persistent Identifiers (PID) can be embedded into Two-Dimensional (2D) QR/matrix code. 2D code can embed more information and is easier to read when compared to the conventional one-dimensional codes. The generated code can be physically attached to the specimen or virtually added to the digitised specimen images. By scanning the code manually, the information of the specimen can be retrieved via the embedded PID. Computer applications can be used to detect and decode the code to automate the workflow in the digitisation process.
Implementation example	Finnish Museum of Natural History (Luomus) Qualified URI based PID of the specimen is present as text and QR code on the imaged specimen. For example, ID http://id.luomus.fi/C.460096 is shown as follows.

Id	ID2
	 :RE 1941.)UE MUSEUM BOTANICUM UNIV. (H), HELSINKI 1317575 GUADR. http://id.luomus.fi/ C.460096 2019-10-24 Museum Botanicum Univ. (H) Helsinki
References	Luo1

4.4 Image Transformation Recommendations

Id	TRANSF1, TRANSF2, TRANSF3
Level	BASIC
Use case	**As a** researcher, **I want** to see specimen images **so that** I can determine if it can be included to my research.
Best practice recommendation	TRANSF1: The high-resolution image without lossy compression needs to be acquired in the imaging process to provide a good base for transformation and processing in the later usage. TRANSF2: Different versions of images, like compressed JPEGs of different spatial resolution and small scale 3D models from CT scanning data, need to be extracted for different purposes such as OCR, online viewing and sharing. TRANSF3: For long-term data preservation, usually TIFF images are used and proprietary image formats like Nikon's NEF and Canon's CR2 raw images may not be supported.
Discussion	For conventional 2D imaging, high-resolution TIFF or RAW images have large file sizes up to hundreds of MB. TIFF images with lossless compression can reduce half of the uncompressed TIFF file size. Converting 12/14/16-bit

Id	TRANSF1, TRANSF2, TRANSF3
	images to 8-bit will reduce the file size but will also lose colour and tone information. The same image in the JPEG format is usually less than 10 MB file size. However, the lossy compression used in JPEG results in the loss of information in the image. And reducing the spatial resolution of the image will further reduce the file size, such as thumbnail images with only tens of KB file size. Imaging process is one of the key parts in the digitisation process. It not only provides images for viewing but also is a starting point for many ETL processes. High-resolution images have to be acquired without lossy compression (ie. all imaged raw data must be available and no data must be lost because of image compression) at the imaging process to preserve as much information as possible of the specimens. This will provide a good base for various later use cases. Usually TIFF images or RAW images like Nikon's NEF and Canon's CR2 formats are used. 3D imaging acquires large amounts of raw data from the imaging devices. For microCT scanners, data is up to hundreds of GB. Data transformation and processing have to be done to generate a 3D model with tens of GB file size. To further reduce the file size, a small scale version can be extracted from the 3D model. Therefore, depending on the use cases, the original acquired images will be transformed into different versions with different image formats, spatial and colour resolutions. Original images will be needed in tasks that require accurate information of the objects in the image, such as quality check, OCR, and segmentation in the digitisation process. JPEG images with the full spatial resolution can be used for online sharing. For online viewing, images have to be transformed into JPEG images with different spatial resolutions, like the small thumbnail images. Similarly, small scale 3D models are needed for online viewing. For long-term data preservation, original data has to be kept. Usually TIFF images are used and proprietary image formats like Nikon's NEF and Canon's CR2 raw images may not be supported.
Implementation example	Finnish Museum of Natural History (Luomus) High-resolution TIFF images without compression are acquired from mass digitisation systems. At the imaging stations, JPEG images with the same spatial resolution are converted from TIFF images and the TIFF images are lossless compressed. Different spatial sizes of JPEG images are generated for online viewing and sharing. All images are achieved for backup and will be stored in the planned long-term preservation.

Id	TRANSF1, TRANSF2, TRANSF3
References	Luo1, Luo2

4.5 Specimen Data Recommendations

Id	DD1, DD2
Level	ADVANCED
Use case	**As a** researcher **I want** to know if data is reliable/complete **so that** I can determine if it can be included to my research.
Best practice recommendation	DD1: When data is extracted from the digitalisation platform to CMS, make sure there is information available about a missing datafield: (1) if the field is marked empty/missing by the digitation operator or (2) if the field was not databased at all by the operator. DD2: If OCR is applied during the ETL process, the CMS should support marking the data field to be "automatically filled" and the ETL process should make sure to fill in this information.
Discussion	Data field value can be one of the following: -absent: information has not been documented at time of collection event and can not be later resolved -unknown: information is documented but is not yet databased -unknown:missing: the information could have been databased but is absent -unknown:indecipherable: the information appears to be present but failed to be captured -automatically filled: information has been databased using automated methods (OCR) but not yet cleaned/verified by a human -default: information is present and has no known problems -erroneous: information is present but contains errors/marked as unreliable by a human -unknown:withheld: information is databased but has been withheld by the provider (Note: not a factor for ETL processes; this is a data publishing problem)
Implementation	See DiSSCO Digitation Guide:

Id	DD1, DD2
example	https://dissco.github.io/ElectronicDataCapture/Transcription.html
References	Interoperability of Collection Management Systems, p5 recommendation #8 (Dil2019) Improved standardization of transcribed digital specimen data, table 2 (Gro2019)

4.6 Media Metadata Recommendations

Id	MM1, MM2
Level	BASIC
Use case	As digitisation manager I want to keep all media metadata so that I can use that information in different use cases.
Best practice recommendation	MM1: Keep media metadata, like EXIF tags embedded in the image and imaging information stored in standalone txt files, as much as possible in the digitisation process MM2: Mark sensitive fields in the media metadata and make them invisible if necessary in required cases.
Discussion	Media metadata can provide useful information to the digitisation process. Cameras usually generate related imaging information, such as image resolution, orientation, date and time, location, in EXIF tags embedded in the images during the image capture. EXIF information can be read and modified by computer software, which can be used to automate the digitisation process to reduce the manual labour work and the potential human mistakes. There may be fields in the media metadata that are sensitive in some special cases. Those potential fields should be marked and invisible to the users.
Implementation example	Finnish Museum of Natural History (Luomus) EXIF information of the images from mass digitisation is extracted from the images and saved as standalone text files. The data and time of the processing and transformation on the image, like rotation and quality check, are recorded in a text file associated with the images. For endangered species, the location information in the occurrence images was removed and the accurate location information was not shown to the public.

Id	MM1, MM2
References	Luo1

4.7 Quality Control Recommendations

Id	QC1
Level	BASIC
Use case	As digitisation manager I want to have the quality control in the digitisation process so that I can provide high quality data
Best practice recommendation	Establish quality control procedures in all the stages of the digitisation process.
Discussion	Quality control is one of the essential parts in the digitisation process. It will ensure the digitised data is of a high quality level for different usages. In each step of the workflow of the digitisation process, quality checks have to be performed in time to find out the errors and mistakes and alerts for checking and re-digitisation. That will prevent the expansion of the errors to the following steps of the digitisation process and minimise the efforts of corrective actions. Regarding ETL procedures in the digitisation process, quality control mainly involves two parts, image check and specimen data check. There is still a large amount of manual work involved in the check due to the corresponding work in the workflow that has not been automated. Automating the digitisation workflow and related quality control will reduce human mistakes and improve the work efficiency in the digitisation process.
Implementation example	Meise Botanic Garden (MBG) MBG implemented a digitisation workflow to digitise the herbarium sheet specimens. This workflow is based on modular designs containing tasks of in-house and outsourced digitisation, processing, preservation and publishing. In each of the tasks, there are quality concerns that quality control has to be performed as shown in the following table from Hid2020.

Id	QC1

Task	**Sub-tasks**	**Quality Concerns**
1 Pre-digitisation curation	Selection of specimens to digitise. Retrieval from storage. Identification of specimens (barcoding). Conservation/restoration of specimens selected for digitisation. Specifying safeguards for handling specimens. Marking specimens that are already digitised. Extraction exceptions for internal imaging (e.g. capsuled specimens or specimens that needed to be imaged twice due to added booklets). Creation of metadata record / adding cover barcodes for external transcription of the labels. Transfer to digitisation station.	Specimens are selected and prioritised for digitisation by collection curators. Some sheets may be damaged or fragile or specimens may need to be remounted to display relevant features.
2 Imaging	Station(s) Setup Digitisation equipment selection, acquisition and set up. Equipment testing/calibration. Training of digitisation technicians.	Equipment should be calibrated to minimise image postprocessing after digitisation.
	Digitisation Mounting for imaging Digitisation of a specimen, creation of a master file (TIFF). Unmounting and return of specimen. Data capture, based on the image when outsourced.	Identification, digitisation and [meta] data capture, so that images are correctly linked to the corresponding specimen records.

Id	QC1

3 Image processing	Retrieval of master files (TIFF) from temporary storage. Creation of derivatives for publishing and distribution (JPEG2000 and JPG); Verification of naming and linking of files (based on barcode ID). Verification of file formats.	Verification of master image resolution format. Verification that derivatives adhere to quality standards.
4 Imaging (alternate)	Imaging (2) and image processing (3) are integrated. The Task receives specimens and produces full sets of images (TIFF, JPEG2000 and JPG).	Same as those for 2 and 3 above.
5 Image processing (alternate)	Verification of image sets (correspondence of master and derivatives). Verification of naming and linking of files (based on barcode ID).	The task is simpler. However, the load increases considerably, from 5,000 to 25,000 weekly specimen image sets to process (400% increase).
6 Store images	Transfer of master and derivative files to archive servers and image servers. Create and preserve links to storage.	Verify that master and derivative files are not corrupted in transfer to storage.
7 Archive images	Deposit master files on external archives for long term preservation.	Verify master is not corrupted in transfer and images are recoverable.
8 Data transcription	Extraction of data from images, populating/complementing specimen record. Final verification/correction of specimen data.	Verify readability of image data for transcription. Verification against reference image and recorded data before publishing.
9 Data transcription (alternate)	Extraction of data from images, populating/complementing specimen record.	Verify readability of image data for transcription.

<table>
<tr><td><b>Id</b></td><td>QC1</td></tr>
<tr><td></td><td>
<table>
<tr><td><b>10 Data transcription validation</b></td><td>Final verification/correction of specimen data.</td><td>Verification against reference image and recorded data before publishing.</td></tr>
<tr><td><b>11 Publish digital specimen</b></td><td>Creation of digital specimen, verifying links to images, data, physical specimen and collection management system data.
Publishing of digital specimen.</td><td>Data, metadata, persistent identifiers and links are used to build stable long-lasting specimens which adhere to FAIR data principles.</td></tr>
</table>
</td></tr>
<tr><td><b>References</b></td><td>Har2020, Hid2020, Hid2020b</td></tr>
</table>

Id	QC2
Level	BASIC (+ADVANCED and STATE-OF-ART)
Use case	**As a** digitisation manager **I want** to have the quality control in the digitisation process **so that** I can provide high quality data
Best practice recommendation	QC2: Establish quality control procedures for images.
Discussion	Specimen imaging data are one of the key outputs from the digitisation process. It is critical to keep the image quality at a high level. The quality control for the images involves the image acquisition, processing, and storing processes. In the image acquisition process of mass digitisation, the images are captured and usually transferred to the imaging station immediately. The captured images need to fulfil the following quality control checks: - format validation - file integrity check - image size, resolution and metadata verification - image colour check - image sharpness check The above quality control measures can be done automatically by the computer applications on the imaging station in real-time. This will find the error images in

Id	QC2
	time for the re-imaging process of the specimens. After passing the above quality controls, the original images are ready for the imaging processing tasks in the workflow, like image renaming by decoding barcodes in the image and image transformation to other formats. In the image processing process, the quality control measures can be done at the imaging station or on a remote server depending on the workflow as - image file name format verification - image derivatives (such as JPEG and PNG images) check (similar QC measures that were done in the above image acquisition part) - image duplication check . The above quality control measures can be done automatically by the computer applications. Often they can be performed offline, since the imaging tasks are based on the original images and do not need access to the physical specimens. After image acquisition and processing, different versions of the image are transferred and stored at different storage areas, such as the staging area, image archive, and long-term data preservation. Also the storage of the imaging station and buffer server have to be cleared periodically after the successful image transfer to other data storages by checking - file integrity Some of the above complex QC tasks done by computer applications belong to the ADVANCED level. Moreover, for different types of specimens, there are different objects in the image, such as specimen, labels, colour chart, scale bar, and barcode, as summarised in the following table from Har2020. It is necessary to make sure that those objects are shown in the image correctly. This work is usually done manually before the digitisation or during the barcoding process. With the development of computer vision and AI techniques, the computer program can achieve relatively high accuracies to detect those objects. However, it requires a large training dataset and computing resources to train the mode and perform the task. This belongs to the STATE-OF-ART level.

Id	QC2

Overview of image elements. Legend: R = Required; C = Conditions apply; NR = Not required; O = Optional (Har2020)

Imaging Workflow	Speci-men	Background	Colour Chart	Scale Bar	Labels	Barcode	Institution Name	Other Elements	Conditions
Microscopy Slides	C	C	NR	NR	C	R	NR	O	Specimens can be hard to capture without special equipment due to size. Background is generally white to facilitate viewing of slide elements. Labels can be placed on both sides of the slide, requiring additional images per slide. Special type labels are important for classifying specimens.
Skins and Vertebrate Material	R	C	C	C	R	R	O	O	Background must maximise the identification of the specimen, avoiding glossy or reflective materials that can hinder border detection. Some specimens may not require a colour chart as colour is

Id	QC2

									not a main feature (e.g., bone samples). The placement of the scale needs to consider the depth and angle at which images are acquired.
Liquid preserved specimens	C	C	O	O	C	C	O	C	The containers can contain more than one specimen and require additional handling. Sometimes the specimens are removed and imaged outside the container, but this takes longer time. Background must be neutral, especially for see-through containers. Barcodes can refer to one container with multiple specimens. Labels can be hard to image due to the shape and placement in or on the container. Paper records

Id	QC2

									which describe the specimens in a container will need to be digitised as well.
Pinned Insects	R	C	R	R	R	R	O	O	Background must maximise the identification of the specimen, avoiding glossy or reflective materials that can hinder border detection.
Herbarium Sheets	R	NR	R	R	C	R	R	O	Labels can be hard to image due to the overlapping with other labels and:or specimen parts. Additionally, some labels are placed at the back of the sheet, requiring additional imaging.
3D Specimen Models	R	C	C	NR	C	C	O	O	Background must maximise the identification of the specimen, avoiding glossy or reflective materials that can hinder border detection. Colour

Id	QC2

									charts can help if the model includes colour or texture. Labels may be captured separately, but some cases such as pinned insects, 3D scanning may help in rapid imaging while minimising specimen handling. Barcodes can be part of the label set

Implementation example

Meise Botanic Garden (MBG)

At MBG, the following quality control measures are applied to the images in the image acquisition and processing subtasks, and image storing subtasks in the digitisation workflows of herbarium sheet specimens.

Table Image acquisition and processing subtasks. (Hid2020)

num	**sub-task**	**type**	**dataset**	**state**		
				start	**success**	**fail**
1	Check file name	AT, QA	TIFF set		names_ok	names_error
2	Check tiff file size, image dimensions and resolution	AT, QA	TIFF set	names_ok	fssr_ok	fssr_error
3	Generate JPEG 2000 derivatives	AT, IH	TIFF set	fssr_ok	jp2_gen	jp2_gen_err
			JP2 set		jp2_gen	jp2_gen_err

Id	QC2

4	Generate jpeg derivatives	AT, IH	TIFF set	jp2_gen	jpg_gen	jpg_gen_err
			JPG set		jpg_gen	jpg_gen_err
5	Check metadata file structure	AT, QC	TIFF set	jpg_gen	md5_ok	md5_error
6	Check duplicates	AT, QA	TIFF set	md5_ok	unique	duplicate
7	Check structure and file size	AT, QA	TIFF set	unique	fss_ok	fss_error
			JP2 set	jp2_gen	fss_ok	fss_error
8	Visual qc tiff files	MT, QC	TIFF set	fss_ok	vqc_ok	vqc_error
9	Check filename	AT, QA, IH	JPG set	jpg_gen	jpgn_ok	jpgn_error
Sub-task Type: AT automated task, **MT** manual task, **QA** quality assurance task, **QC** quality control task, **IH** sub-task performed in-house only.						

Table Image Storing sub-tasks. (Hid2020)

num	sub-task	type	dataset	state		
				start	**success**	**fail**
1	Remove duplicates and bad crops (Table 11)	MT, QA	TIFF set	vqc_ok	dup_rmv	
			JP2 set	fss_ok	dup_rmv	
			JPG set	jpgn_ok	dup_rmv	
2	Copy files to archive	AT	JP2 set	dup_rmv	stg_ok	stg_error
			JPG set	dup_rmv	stg_ok	stg_error
3	Generate image viewers	AT	JP2 set	stg_ok	vwrg_ok	

<table>
<tr><th><b>Id</b></th><th>QC2</th></tr>
<tr><td></td><td>
<table>
<tr><td>4</td><td>Copy files to ftp server</td><td>AT</td><td>TIFF set</td><td>stg_ok</td><td>svrc_ok</td><td>svrc_error</td></tr>
<tr><td>5</td><td>Copy files to external archive</td><td>AT</td><td>TIFF set</td><td>svrc_ok</td><td>arc_ok</td><td>arc_error</td></tr>
<tr><td rowspan="2">6</td><td rowspan="2">Check jp2 and jpg sets</td><td rowspan="2">AT, QA</td><td>JP2 set</td><td>vwrg_ok</td><td>stgv_ok</td><td>stgv_err</td></tr>
<tr><td>JPG set</td><td>stg_ok</td><td>stgv_ok</td><td>stgv_err</td></tr>
<tr><td>7</td><td>Clear buffer server</td><td>AT, QA</td><td>TIFF set</td><td>arc_ok</td><td>bufc_ok</td><td>bufc_err</td></tr>
<tr><td rowspan="2">8</td><td rowspan="2">Clear buffer server</td><td rowspan="2">AT</td><td>JP2 set</td><td>stgv_ok</td><td>bufc_ok</td><td>bufc_err</td></tr>
<tr><td>JPG set</td><td>stgv_ok</td><td>bufc_ok</td><td>bufc_err</td></tr>
<tr><td colspan="7"><b>Sub-task Type: AT</b> automated task, <b>MT</b> manual task, <b>QA</b> quality assurance task.</td></tr>
</table>
</td></tr>
<tr><td>References</td><td>Har2020, Hid2020, Hid2020b</td></tr>
</table>

Id	QC3
Level	BASIC (+ADVANCED)
Use case	**As a** digitisation manager **I want** to have the quality control in the digitisation process **so that** I can provide high quality data
Best practice recommendation	QC3: Establish quality control procedures for specimen data.
Discussion	Specimen data is the most important core part in the digitisation process along with specimen imaging data. In mass digitisation, usually preliminary specimen data with minimum information level are extracted from the specimen during the barcoding process to speed up the digitisation. More extensive transcription can be done later with the specimen image at a collection management system or dedicated transcription portals. To improve the quality of the specimen data in the digitisation process, the quality control measures must be applied in the digitisation process, such as simple data format validation of

Id	QC3
	- date, time, and higher locality names . By further utilising the list of controlled vocabularies/terms from the authorised sources to validate - scientific names - localities - peoples' name . Automated geo-referencing processes can be used to improve the data quality.
Implementation example	Finnish Museum of Natural History (Luomus) At Luomus, the preliminary specimen information is recorded at the barcoding step in the mass herbarium digitisation process with a web-based system. In the system, the list of controlled vocabularies from the authorised sources is used to validate the scientific name, country, and municipalities of the specimen. The special cases of the localities are alerted in the system with highlights to the user. The formats of year, and the links between country and municipalities are validated instantly after the input of the fields. At the mass digitisation of the pinned insect, automated geo-referencing is used to achieve high specimen data quality.
References	Luo1, Har2020

5. Maintaining the PBD

In DiSSCo Prepare project WP3.2 (capacity enhancement) we have created a documentation website (https://dissco.github.io/) for digitisation guides, including ETL best practices described in this document. The work on the best practices is not however over. They will be expanded and more BP recommendations will be created. For the BP recommendations to be valid, there must be a review process. In this chapter we outline how the BP recommendations will be maintained in the future.

The documentation website is based on GIT version control. The GIT repository that contains the documentation is located on the GitHub platform (https://github.com/DiSSCo/dissco.github.io). The GitHub site is not open for anyone to make alterations, but anyone can make a "pull request" to change the contents of the documentation. GitHub provides a method for reviewing and accepting these pull requests. Once accepted, the changes will automatically become visible on the final documentation website.

Before pull requests are accepted, changes to the recommendations or the new recommendation must be reviewed by experts, so that the acceptability of the BDP remains high.

The process will be as follows:

1. A need for a change or a new idea for a BP recommendation is thought of by anyone working in the field of digitisation - most likely a person involved in technical implementation (later called contributor).
2. The contributor will "clone" the Git repository to his/her own computer and make modifications to the recommendations.
3. Once satisfied, the contributor will "push" the changes to a new "branch" to the GitHub repository
4. The contributor will then ask to create a "pull request"
5. Owners of the repository will be notified about the pull request.
6. Owners can ask for a team to review the changes. GitHub platform provides powerful tools that can be used to discuss the pull request (for example an individual line of text), to ask for changes, accepting the made changes and so on. (Read more: https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/about-pull-requests)
7. Reviewers will comment and finally approve the changes to BP recommendations.
8. Once reviewers are satisfied, the pull request is accepted by Owner of the repository and "merged" to the main branch of the documentation.
9. The changes will become visible in the documentation website.

It should be noted that Git version control maintains the full history of the documentation. The history is not browsable on the documentation website, but anyone can browse the "commit history" of the GitHub repository.

As a further development point, the BPD located on the documentation site should contain explanations of this process, these instructions and further formatting instructions on how to use the GitHub Pages with Jekyll, and Just the Docs Jekyll theme.

6. Discussion

In this work, we made a list of recommendations of the best practices for the ETL (Extract, Transform and Load) procedures in the digitisation of natural history collections. Firstly, we made an overview of the goal of the work that we would like to achieve and the scope of the work to cover in the digitisation process. Secondly, we gathered the related work on the digitisation workflows from task partner institutions, reports from relevant projects (such as ICEDIG), and scientific publications. They are listed in the APPENDIX: 'Literature with digitisation workflows'. We intensively reviewed those workflows to identify the potential ETL procedures in each step of the workflow and categorised them into three groups: pre-ETL workflows, ETL workflows, and post-ETL workflows. The extensive list is in the related APPENDIX chapters. Based on those workflows, we made 18 best practices recommendations in 7 categories as

1. Infrastructure recommendations

2. Organisational recommendations

3. Identifier recommendations

4. Image transformation recommendations

5. Specimen data recommendations

6. Media metadata recommendations

7. Quality control recommendations

Each best practice recommendation has a unified format based on a customised template. It contains items of id, level, use case, best practice recommendation, discussion, implementation example, and references. With the unified standardised format, the recommendations are easy to follow. There are three levels of best practice recommendation, basic, advanced, and the state-of-art, indicating how demanding the recommendation is. The discussion part provides more information about the recommendation. The implementation example presents examples of how the recommendation is implemented in practice.

Those recommendations are mainly targeted to institutions that are at the initial phase of building up their digitisation processes at the ALA Digitisation Maturity Level 1 and 2. Institutions at Level 0 are out of scope of this work, because they would require detailed guidance and individualised support.

Moreover, recommendations in this work do not include any particular software, service providers, or other concrete approaches to implement, because institutions have their own organisational model and digitisation related infrastructures. Moreover, different digitisation levels work with different digitisation methods. One single standardised procedure may not work for all digitisation projects. Institutions can evaluate the recommendations in their particular digitisation projects and use the recommendations as goals to meet in establishing the new digitisation infrastructure or improving the existing digitisation activities.

REFERENCES

All2019: Allan L et al. (2019) Digitisation using Automated File Renaming and Processing. Microscopes Slides. (TODO PUBLISHED?)

Alw2015: Alwazae M., Perjons E, & Johannesson P (2015) Applying a Template for Best Practice Documentation. Procedia Computer Science 72 (2015) 252 – 260. https://doi.org/10.1016/j.procs.2015.12.138

Dil2019: Dillen M, Groom Q, & Hardisty A. (2019). Interoperability of Collection Management Systems. Zenodo. https://doi.org/10.5281/zenodo.3361598

Dri2014: 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. https://doi.org/10.3897/phytokeys.38.7168

Gro2019: Groom Q et al. (2019) Improved standardization of transcribed digital specimen data. Database, Volume 2019, 2019, baz129. https://doi.org/10.1093/database/baz129

Gün2017: Güntsch et al. (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects, Database, Volume 2017, 2017, bax003, https://doi.org/10.1093/database/bax003

Has2012a: Haston E, Cubey R, Pullan M, Atkins H, Harris D (2012) Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. ZooKeys 209: 93-102. https://doi.org/10.3897/zookeys.209.3121

Has2012b: Haston E, Cubey R, & Harris D J (2012) Data concepts and their relevance for data capture in large scale digitisation of biological collections. IJHAC, Volume 6, Issue 1-2. https://doi.org/10.3366/ijhac.2012.0042

Har2020: 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

Hid2020: Nieva de la Hidalga A, Rosin PL, Sun X, Bogaerts A, De Meeter N, De Smedt S, Strack van Schijndel M, Van Wambeke P, Groom Q (2020) Designing an Herbarium Digitisation Workflow with Built-In Image Quality Management. Biodiversity Data Journal 8: e47051. https://doi.org/10.3897/BDJ.8.e47051

Hid2020b: Nieva de la Hidalga A, van Walsun M, Rosin, P, Sun X, Wijers A (2019) Quality Management Methodologies for Digitisation Operations, ICEDIG Project Report. https://doi.org/10.5281/zenodo.3469521

Sco2019: Scott B, Baker, E, Woodburn M, Vincent S, Hardy H, Smith V S (2019) The Natural History Museum Data Portal, Database, Volume 2019, 2019, baz038, https://doi.org/10.1093/database/baz038

CETAF: CETAF n.d., Best Practises of CETAF Stable Identifiers (CSI), Retrieved from https://cetaf.org/resources/best-practices/cetaf-stable-identifiers-csi-2/

APPENDIX I: REVISION HISTORY

Date	Author(s)	Description
2022-03-xx	Esko Piirainen, Zhengzhe Wu, Lisa French, Laurence Livermore	Initial version of the milestone deliverable.
2022-04-01	Sofie De Smedt	Added many useful comments and additions from the point of view of their institution's technical digitization infrastructure.

APPENDIX: REVIEW HISTORY

This will actually be maintained in the Git based Digitisation Guide manual, this milestone is not reviewed.

Review date	Reviewed version	Reviewer	Notes

APPENDIX II: IMPLEMENTATION DEMONSTRATIONS

Template for reporting BP was applied in an organisation:

Implementing organisation	Name of org, contact person, contact info
Implementation time	Start date, end date
Implementation cost	How many person work months implementation took
Experiences and feedback	
Measurements	See appendix: Measurement Report measurable improvements in performance

The BP has not been currently demonstrated in practice.

APPENDIX III: MEASUREMENT

PLACEHOLDER: Git Indicators for measuring the quality and performance of the BP guide will be added to the GitHub Digitisation Guide manual.

APPENDIX IV: Workflows / documentation provided for this WP

TODO REPLACE LINKS TO POINT TO DISSCO Knowledge base [LF: KB not currently allowing upload of new files (08/04/2022) – will be added after milestone review]

Link	Organisation	Desc	Ref
PDF	Luomus	Workflow for insect-line mass digitisation process Workflow for non-mass digitisation processes	Luo1
HTML	Luomus	Plans on how CT scan/3d model workflow will happen	Luo2
PDF PDF v2 Doc v2	LISI Inst de Agronomia - Univ de Lisboa	LISI Herbarium Digitization Workflow	LIS1

Link	Organisation	Desc	Ref
PDF Doc	RBGE Royal Botanic Garden Edinburgh	RBGE Digitisation Workflows	RBGE1
PDF Doc	RBGE Royal Botanic Garden Edinburgh	RBGE ETL Processes	RBGE2
Article	RBGE Royal Botanic Garden Edinburgh	Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach	Has2012a
Article	RBGE Royal Botanic Garden Edinburgh	Data concepts and their relevance for data capture in large scale digitisation of biological collections	Has2012b
Article	RBGE Royal Botanic Garden Edinburgh	The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels	Dri2014
PDF Doc	NHM, London	Summary of other doc + specimen data to CMS	NHM1
PDF Doc	NHM, London	Slide Digitisation - End of day checklist	NHM2
PDF PDF	NHM, London	eMesozoic workflow diagram	NHM3
PDF Doc	NHM, London	ALICE Workflow	NHM5
PDF Doc	NHM, London	Microscope slides digitisation - article	All2019
PDF Doc	NHM, London	Bee types digitisation workflow	NHM8
Article	NHM	The Natural History Museum Data Portal	Sco2019

Link	Organisation	Desc	Ref
Article PDF	Meise Botanic Garden	Designing an Herbarium Digitisation Workflow with Built-In Image Quality Management	Hid2020
HTML/txt	Museum für Naturkunde Berlin	MfN workflow ETL summary	MfN2

APPENDIX V: Literatures with digitisation workflows

Link	Organisation	Name	Ref
Article	ICEDIG	Interoperability of Collection Management Systems	Dil2019
Article	ICEDIG	Quality Management Methodologies for Digitisation Operations	
Article	ICEDIG	Mass-imaging of microscopic and other slides	
Article	ICEDIG	Best practice guidelines for imaging of herbarium specimens	
Article	ICEDIG	State of the art and perspectives on mass imaging of pinned insects	
Article	ICEDIG	State of the art and perspectives on mass imaging of liquid samples	
Article	ICEDIG	State of the art and perspectives on mass imaging of skins and other vertebrate material	
Article	ICEDIG	Methods for Automated Text Digitisation	
Article	ICEDIG	Conceptual design blueprint for the DiSSCo digitization infrastructure	Har2020
Article	NCSU	Results and insights from the NCSU Insect Museum GigaPan project	

Link	Organisation	Name	Ref
Article	NHM	No specimen left behind: industrial scale digitization of natural history collections	
Article	INHS	InvertNet: a new paradigm for digital access to invertebrate collections	
Article	Swiss Aca of Sci	Handbook on natural history collections management – A collaborative Swiss perspective	
Article		Improved standardization of transcribed digital specimen data	Gro2019
Article	Uni Coimbra	A Strategy to digitise natural history collections with limited resources	
Article		Back to the future: A refined single-user photostation for massively scaling herbarium digitization	
Article	NHM	Georeferencing the Natural History Museum's Chinese type collection: of plateaus, pagodas and plants	

APPENDIX VI: Pre-ETL workflows

Infra	Step	Action type	Type	Ref	Notes
Digi station	Fully qualified URI Identifier of the specimen (globally unique persistent identifier) is present as QR-Code on the imaged specimen	Manual (repeated for each specimen)	Identifier	Luo1	
Digi station	Barcode is created / scanned	Manual (repeated for each specimen)	Identifier	LIS1	Most likely internal catalogue number (not fully qualified URI identifier) based on rest of the workflow
Digi station	Apply barcodes / scan barcodes	Manual (repeated for each specimen)	Identifier	RBGE 1	Unclear if fully qualified URI identifier or internal catalogue number
Digi station	Before capturing image, specimen data is entered Camera operator enters their details into an online form that queries the imaging database to check if there are records for the specimen barcodes already.	Manual + Semi-automated (repeated for each specimen)	Specimen data	RBGE 1 RBGE 2	
Digi	Images are taken in	Manual	Image + Identifier	RBGE	

Infra	Step	Action type	Type	Ref	Notes
station	RAW format using CaptureOne software. The barcode on the specimen is scanned by the camera operator and used as the filename for the image. A mask for the crop is applied manually by the camera operator	(repeated for each specimen)		2	
Digi station	The operator selects image(s) and these are processed to TIF format by CaptureOne software. As part of this conversion process the image is cropped to the mask applied by the camera operator and sharpening is applied to the TIF.	Manual (repeated for each specimen)	Image (Transformations)	RBGE 2	
Digi station	After imaging the dorsal and lateral views, the images have to be rendered (Helicon Focus) and renamed (BardecodeFiler). Then we need to generate a filelist for the dorsal images (command prompt) in order to associate each UID with the correct PTN. After this is done, we remove the PTN from the name of the image (Bulk Rename Utility)	Semi-automated (daily / overnight)	Image (Transformations)	NHM8	

Infra	Step	Action type	Type	Ref	Notes
	and we crop the images to remove the IRN tags and the dead space (Lightroom) 8. Leave to run overnight				
Digi station	Something called "Syrup" is done after image capture Rename image file with concatenation of scanned specimen barcode and drawer barcode, plus incremental suffix for more than one image of item	Automated (presumed) (on-the-fly)	Image + Identifier	NHM3	Level of automation?
Digi station	Before capturing image, "System quality control" is done	Automated (on-the-fly?)	Specimen data? (Quality control)	RBGE 1	What is controlled?
Digi station → CMS	System creates a record for each new barcode and populates record with data	Automated (on-the-fly?)	Specimen data	RBGE 1	
Digi station → CMS?	The metadata are managed in a MySQL image data management database, and in the image file exif data. The metadata for the original image files are held in one table, and	Automated ?	Image metadata	RBGE 2	Is the mySQL a temporary image metadata repository or also the final one? See doc for exact metadata

Infra	Step	Action type	Type	Ref	Notes
	comprise information copied from the image exif along with additional metadata derived from a folder structure developed for capturing metadata not included in the exif data. A watched folder for the image files sits within a hierarchical folder structure with each folder name holding the relevant metadata for the image file.				fields
Digi station	The camera operator checks to see that there is a pair of images (a RAw & TIF) for each barcode. If either file is missing the images can not be processed, as the image processing service is expecting both.	Manual (repeated for each specimen)	Image (Quality control)	RBGE 2	
Digi station	Manual checks are done: Look through the list of file names in the final folder - Common errors to check: - duplicate unique identifier (UID) barcodes (i.e. a slide that hasn't been barcoded when imaged and therefore	Manual (end of day)	Image (Quality control)	NHM2	

Infra	Step	Action type	Type	Ref	Notes
	has been renamed with the UID of the previous slide). - “Missing” numbers from the UID sequence (i.e. a slide was barcoded but then not imaged before returning it to the drawer) Check that there isn’t any “unexpected” _additional images (i.e. images of the front side of a slide; duplicates of envelopes				
Digi station	We perform the quality checks; Check that all images look alright	Manual	Image (Quality control)	NHM8	
Digi station → Staging area	Image is captured and transferred to dropbox Both the RAW and TIFF files are saved onto a network share drive. The folder structure for this includes the camera the operator is using and the operator's username. This is a temporary storage location.	Manual (repeated for each specimen)	Image	RBGE 1 RBGE 2	
Digi station → Staging area	Files manually moved to different folders copy the date folder (with the “images”	Manual (end of day)	Image	NHM2	(Destination seems to be a network drive)

Infra	Step	Action type	Type	Ref	Notes
	folder within) in final to: emu-import-Slides_SLR_X (\\dfs-ctdb)				
Digi station → Image archive	Copy the date folder (with the “images” folder within) in final to: Emu-import-dcp_digitisation (This is our back-up area)	Manual (end of day)	Image (Backup)	NHM2	
Digi station → ?	Copy the date folder (with the “images” folder within) in final to: DCP-1 - EXTERNAL HARD DRIVE (NOTE: It’s going to be tricky for everyone to save to the hard drive if you’re all leaving at the same time, so you can do this step the next day)	Manual (end of day)	Image (Backup)	NHM2	Reason for the external hard drive?
Digi station	Metadata such as digitiser name/operator is generated and stored at the digitisation station as text file	Automated (on-the-fly)	Image metadata	Luo1	
Digi station	Metadata of all images are generated using XnView and a .ipt-template	Semi automated (once a day)	Image metadata	LIS1	x1 - difference between this and x2 is not clear
Digi	Metadata of all images is generated	Semi	Image metadata	LIS1	x2 - difference between this

Infra	Step	Action type	Type	Ref	Notes
station	using Limbs digitization software	automated (once a day)			and x1 is not clear See doc for exact metadata fields
Digi station → Backup storage?	Copy images+metadata in current day folder to external drive	Manual (once a day)	Image, Image metadata	LIS1	Is the external drive for backup purposes?
Digi station → Staging area	Copy images+metadata to staging area using FileZilla program	Manual (once a day)	Image, Image metadata	LIS1	
Digi station → Staging area	Images loaded onto EMu server	?	Image	NHM3	Needs more info
Digi station	Post-Processing can include color corrections and rendering of scale bars	Manual	Image	MfN2	
Digi station → Staging area	Manual workflow for 2D imaging on demand: DNG and PNG files are stored in structured file system Manual upload to DAM system after quality check	Manual	Image (+Quality control)	MfN2	
Digi station	In case of multi-focus imaging: image acquisition and rendering of multi-focus images are	Manual	Image	MfN2	

Infra	Step	Action type	Type	Ref	Notes
	separated steps.				
Digi station	Backups are not done at digitisation station	--	Infa (Backup)	Luo1	
Digi station	Specimen data is entered to Excel spreadsheet	Manual (repeated for each specimen)	Specimen data	Luo1	
Digi station	Object related metadata: - Mostly metadata are acquired with Excel spreadsheets, which are designed for enabling bulk uploads into the CMS	Manual (repeated for each specimen)	Specimen data?	MfN2	"Metadata" == data? Not image metadata?
Digi station → CMS → Image publishing platform	Images are captured and uploaded straight to CMS using Web UI; thumbnails etc are generated by image API; metadata is created and stored; images are moved to image publishing platform	Semi automated (repeated for each image)	Image, Image metadata	Luo1	The "ETL" parts are done automated and instantaneously without a specific ETL part in the workflow
Digi station	Raw scans are done using CT Scanner	Manual (repeated for each specimen)	3d/CT scan	Luo2	
Digi station	3d model is generated from raw CT scans	Manual (repeated for each	3d/CT scan	Luo2	

Infra	Step	Action type	Type	Ref	Notes
		specimen)			
Digi station	A smaller scale 3d model is discretized from the model	Manual (repeated for each scan)	3d/CT scan	Luo2	
Digi station → CMS, publishing platform	Small scale 3d model is uploaded straight to CMS using Web UI; thumbnails etc are generated by image API; metadata is created and stored; images and 3d scans are moved to image publishing platform	Semi automated (repeated for each model)	3d/CT scan	Luo2	The "ETL" parts are done automated and instantaneously without a specific ETL part in the workflow
Digi station	Mostly CT images from scientific projects. Processing is mostly done by requesters and/or student helpers. Raw and processed files are stored in the file system and managed by the lab technicians. Upload routines for long-term-archiving and publication are not established yet	Manual (repeated for each specimen)	3d/CT scan	MfN2	
Digi station	Multiple specimens Give one barcode per specimen. Make sure it is clear which barcode corresponds to each specimen (written on the barcode,	Manual	Identifier, Image (multi specimen)	NHM8	

Infra	Step	Action type	Type	Ref	Notes
	examples: male/female, a/b/c, type etc.) Image as many times as the specimens, each time with only one UID visible (the other ones reversed)				
Digi station → Staging area	Automated workflows for data acquisition with mobile devices (vertebrate collections and assessments): W e use the app ODK Collect. Data is uploaded to a central ODK server.	Automated		MfN2	

APPENDIX VII: ETL workflows

Infra	Step	Action type	Type	Ref	Notes
Staging area	System polls dropboxes; starts to execute if new files found	Automated (running background task)	Image	RBGE 1	
Staging area	Quality control is done Checks include: Filename - the file name is checked for format and length. It should be the letter E followed by 8 numbers. Any additional images for a particular barcode should be suffixed using _. If a filename does not pass, this is returned to an Errors folder which is manually	Automated + Manual	Image (Quality control)	RBGE 1 RBGE 2	

Infra	Step	Action type	Type	Ref	Notes
	checked by a Digitisation Officer. Filesize - the size of the file is checked, if it falls outside of the set parameters the file is returned to an Errors folder which is manually checked by a Digitisation Officer. Image pair - whilst a manual check has been performed by the camera operator these can still be missed. If one of the files is missing it is returned to an Errors folder which is manually checked by a Digitisation Officer.				
Digi station → staging area	Images and metadata are fetched in real-time or in batches to staging area	Automated (on-the-fly OR daily)	Image, Image metadata	Luo1	
Staging area?	This script takes individual images with metadata encoded in the filename and creates a specimen record with appropriate attachments to the taxonomy and location modules. Metadata encoded in format: "UIDBarcode_LocationIRN_TaxonIRN.jpg"	Semi-automated ?	Identifier, Image, Image metadata	All2019	
Staging area	Specimen identifier URI is detected and extracted from specimen image and image is named to match the ID and image metadata is updated to contain the specimen ID	Automated (running background task)	Identifier, Image, Image metadata	Luo1	
Staging area?	Systems perform all processing steps and deliver two image files (Tiff/Raw and Png). All technical	Automated (when?)	Image (Transf) Image	MfN2	

Infra	Step	Action type	Type	Ref	Notes
	and administrative Metadata related to the images are delivered with a json sidecar file (XML in METS format for library and archival material)		metadata		
Staging area? → CMS	Object-related metadata are acquired in different ways. In one case they are delivered together with the images in the json sidecar file and parsed by the database management team (this process is not yet fully established). In most cases object related metadata are acquired in Excel spreadsheets and imported to the respective CMS	Automated Semi-Automated (when?)	Specimen data or Image metadata?	MfN2	
Staging area → CMS	Images are attached to records in Specify (CMS) based on catalog numbers in file names	Semi-automated (how often?)	Image	LIS1	
Staging area? → Image publishing platform	Automated import of image files and related metadata to the digital asset management system	Automated (when?)	Image	MfN2	
Staging area → Image publishing platform	Specify script creates copies of original large image files and creates thumbnails (PNG) to Specify Attachment Server and renames based on UUID; original filename and location is kept in attachment metadata	Semi-automated (how often?)	Image, Specimen data (Transformations)	LIS1	Originals are tiff files, about 5574x7370 px,8-bit sRGB Thumbnails are png files, about 93x123 px, 8-bit

Infra	Step	Action type	Type	Ref	Notes
					sRGB
Image publishing platform	Original TIFF files are converted to JPEGs running a Python script that uses the ImageMagick library. TIFF images are kept.	Semi-automated (how often?)	Image (Transformations)	LIS1	
CMS	Run SQL UPDATE to modify attached TIF files links to point to generated JPEG links instead	Manual (how often?)	Specimen data	LIS1	
Staging area	EMu eMesozoic Batch Operation script	?	?	NHM 3	Needs more info
Staging area	Each TIFF is processed through OCR software; the OCR output is recorded as unstructured text to CMS as separate record (not to primary specimen data) A copy is made of the TIF file which is submitted to an OCR pipeline.	Automated	Specimen data (OCR)	RBGE 1, RBGE 2	(to what level automated?)
Staging area → Digi station? → CMS	Batches of records with shared collectors or geography were then transcribed by digitisation staff, using a record set of the data records in the CMS along with an identical set of images presented in the same order in image-viewing software	Semi-automated? (what intervals?)	Specimen data (OCR)	RBGE 1	
Staging area → Image publishing platform	Automated workflows for data acquisition with mobile devices (vertebrate collections and assessments): Data is uploaded to a central ODK server. The process for integration into the media repository and the CMS is also automated	Automated? (what intervals?)	Specimen data Image Image metadata?	MfN2	

Infra	Step	Action type	Type	Ref	Notes
→ CMS					
Staging area	Original sized JPG and smaller thumbnails are generated	Automated (running background task)	Image (Transformations)	Luo1	
Staging area	Creation of JPG and zoomify files A high resolution JPG is produced. This is stored in an online accessible repository and can be downloaded from our online catalogue. A tiled image is created. This is stored in an online accessible repository and can be viewed on the online catalogue.	Automated (running background task)	Image (Transformations)	RBGE 1 RBGE 2	
Staging area?	Image rotation and cropping using XnConvert 12) XnConvert watches the hot folder “renamed”. 13) The renamed image file is copied then rotate 180o (step 1) and cropped to specified coordinates to remove the temporary location and taxon IRN label from the final image (step 2; Figure 5). 16) The cropped image is then automatically to the folder “cropped”. 17) At the end of each day the renamed and processed image files are manually transferred from “cropped” to an “images” folder within a date folder “YYYY_MM_DD” within a “final” folder. 18) The image files in the folders “original_processed” and “renamed” are manually deleted	Semi-automated? Automated? (when?)	Image (Transformations)	All2019	

Infra	Step	Action type	Type	Ref	Notes
	daily				
Staging area	The metadata for the transformed files produced by the ETL processes are also managed in the MySQL image data management database. Each will have a record of the original file from which it was derived, along with ...	Automated ?	Image metadata	RBGE 2	See doc for exact metadata fields
Staging area → Image publishing platform	JPG and zoomify files are moved to Image streaming online service	Automated (running background task)	Image	RBGE 1	
Staging area → Image publishing platform → CMS	Images loaded into EMu multimedia Load images from source folder into EMu Multimedia. For each unique specimen barcode number in the image file name, spawn a new eMesozoic barcode stub record and attach the image(s) to it. At least: location id attached via location barcode; media id via specimen barcode (???)	Automated (presumed) (on-the-fly?)	Image + Image metadata??	NHM 3	Needs more info on level of automation and details
Staging area → CMS	Script takes individual images and attaches them to an existing record by matching the UID (NHMUK barcode) in the filename with an existing record in EMu. Metadata encoded in format: "UIDBarcode_suffix.jpg"	Semi-automated?	Data linking	All2019	
Staging	"Sapphire script" - copy image	Automated	Image + Specimen	NHM	Needs more info

Infra	Step	Action type	Type	Ref	Notes
area	and location to specimen record Search for the specimen number entered (applying search filters). On matching, copy the attached Location and multimedia to the destination specimen record. New specimen records arise from editing the barcode stub	(presumed) (on-the-fly?)	Data	3	on level of automation
Staging area → Image archive	Original TIFF images are moved to image archive and deleted from staging area	Semi-automated (couple times a week)	Image	Luo1	Done using command line tools but could (should!) be automated in the future
Staging area → Image archive	Archive raw and TIFF files	Automated (what intervals?)	Image	RBGE 1	
Staging area → Image publishing platform	Generated JPG images including thumbnails are moved to image publishing platform	Semi-automated (couple times a week)	Image	Luo1	Done using command line tools but could (should!) be automated in the future
Staging area → Image publishi	URL of published images and other image metadata is stored to image metadata database	Semi-automated (couple	Image metadata	Luo1	Done running a Python

Infra	Step	Action type	Type	Ref	Notes
ng platform		times a week)			script
Staging area	Backups are not done at staging area	--	Infra (Backup)	Luo1	
Staging area → Image publishing platform	The automated pipeline moves images for publication and download.	Automated	Image, Image metadata(?)	RBGE 1	

APPENDIX VIII: Post-ETL workflows

Infra	Step	Action type	Type	Ref	Notes
Staging area → Backup storage	"At a later stage" images will be copied to INCD cloud service for backup archiving	?	Image	LIS1	Possibly semi-automated ?
Image archive → Long-Term Archive	Images are moved from image archive to long-term archive	TBD	Image	Luo1	Future feature
CMS	CMS starts to show specimen images once images are in	Automated	Data linking	Luo1	

Infra	Step	Action type	Type	Ref	Notes
	publishing platform and the URLs of the images are in image metadata service				
CMS	A SOLR index is used to link the image files to the data records for display on our online catalogue	Automated	Data linking	RBGE2	
Staging area	The camera operator's perform a second check, once all of the images should have been processed to ensure that this has been successful. This is using the same online form as they used prior to processing the images. If any barcodes are showing as unprocessed then the camera operator can resubmit them for processing again, or pass the issue onto a Digitisation Officer to see if they can identify the reason for this failing. Once the camera operator is satisfied that all of the images have been successfully process	Manual	Image	RBGE2	

Infra	Step	Action type	Type	Ref	Notes
	ed they are deleted from the temporary storage location.				
CMS	Specimen data is upload to CMS using Excel spreadsheet	Manual	Specimen data	Luo1	
CMS	Object related metadata: - Mostly metadata are acquired with Excel spreadsheets, which are designed for enabling bulk uploads into the CMS	Manual	Specimen data?	MfN2	"Metadata" == data? Not image metadata?
CMS	Georeferencing, validations etc are done by CMS	Automated	Specimen data (Quality control)	Luo1	
→ Backup storage	Images, 3d models are automatically backed up to different cloud server environment Databases (specimen, image metadata) are backed up nightly to tape	Automated	Infra (Backup)	Luo1, Luo2	
Image archive	The archive folders are included in regular nightly backups. These are written to tape and taken offline, this is a manual process.	Manual	Infra (Backup)	RBGE2	

Infra	Step	Action type	Type	Ref	Notes
	They are also manually copied onto external hard drives to provide a backup and an easily accessible version of the data once it has been taken offline.				
CMS	OCR raw data was used as an aid to enhancing minimally database records see Dri2014	TODO	Specimen data (OCR)	RBGE2 Dri2014	
CMS	OCR raw data is picked up as part of the SOLR index and is displayed on our online catalogue as part of the specimens record.	Automated	Specimen data (OCR)	RBGE2	
Digi station	Manual clean-up: Once all the images have been saved and backed up you can empty the “Processing” folders:Empty the following folders	Manual (end of day)	Infra (Clean up)	NHM2	
Digi station	Scripts, developed in-house for the 2015 pilot, are also currently used for a series of processes known as Flows: (1) bulk transfer of image files from the imaging PC to the	Semi-automated	Infra (Clean up)	All2019	

Infra	Step	Action type	Type	Ref	Notes
	data managers, and (2) after ingest into EMu the deletion of the original image files on the imaging PC i.e. clear-down process (Flows; Workflow 3)				
?	Lu to import "drawer locations" spreadsheet (Locations module) Lu to import "specimen locations" spreadsheet (Catalogue module) Lu to import condition spreadsheet (Condition module) Lu to import treatment/storage spreadsheet (Processes module) Curators to resolve flagged merges	? (Every three months)	?	NHM7	
--	At present completely decentralised, following the (niche-)standards of the respective community	--	Specimen data - Analytical	MfN2	
--	The RBGE uses the CETAF stable	--	Specimen data - Analytical/chemical/molecu	RBGE2	

Infra	Step	Action type	Type	Ref	Notes
	identifiers to track material coming from the Herbarium and the Living collection via a molecular collection management system called EDNA. This is an in-house developed system that is under review.		lar data		