Research Ideas and Outcomes 6: e55522
doi: 10.3897/rio.6.e55522

Project Report

# Technical capacities of digitisation centres within ICEDIG participating institutions

Naomi Cocks‡, Laurence Livermore‡, Vincent Stuart Smith‡, Matt Woodburn‡

‡ The Natural History Museum, London, United Kingdom

Corresponding author: Laurence Livermore (l.livermore@nhm.ac.uk)

Reviewable v1

Received: 17 Jun 2020 | Published: 03 Jul 2020

Citation: Cocks N, Livermore L, Smith VS, Woodburn M (2020) **Technical capacities of digitisation centres within ICEDIG participating institutions**. Research Ideas and Outcomes 6: e55522.
https://doi.org/10.3897/rio.6.e55522

## Abstract

DiSSCo, the Distributed System of Scientific Collections, is seeking to centralise certain infrastructure and activities relating to the digitisation of natural science collections. Deciding what activities to distribute, what to centralise, and what geographic level of aggregation (e.g. regional, national or pan European) is most appropriate for each task, was one of the challenges set out within the EC-funded ICEDIG project. In this paper we present the results of a survey of several European collections to establish current digitisation capacity, strengths and skills associated with existing digitisation infrastructure. Our results indicate that most of the institutions surveyed are engaged in large-scale digitisation of collections and that this is usually being undertaken by dedicated teams of digitisers within each institution. Some cross institutional collaboration is happening, but this is still the exception for a variety of funder and practical reasons. These results inform future work that establishes a set of principles to determine how digitisation infrastructure might be most efficiently organised across European organisations in order to maximise progress on the digitisation of the estimated 1.5 billion specimens held within European natural science collections.

## Keywords

Staff Skills, Digitisation Equipment, Natural History Specimens, Digitisation Prioritisation, Digitisation Protocols, Digitisation Methods

## 1. Introduction

This report summarises the technical capacities of digital centres within collection-holding institutions in the ICEDIG project. The longer term aim is to combine this data with some work previously completed on policy to produce underlying best practice policy for digitisation for the Distributed System of Scientific Collections (DiSSCo) Research Infrastructure (RI) for natural science collections.

The survey is intended to provide information that will help to demonstrate:

- Specific areas of excellence within each institution.
- Which collections have been prioritised for digitisation and why.
- The resources that are available for digitisation of various collections across ICEDIG partner institutions.

A previous report (MS43) provided a summary of common policy elements within ICEDIG partner institutions, creating a dashboard of results that is available on the ICEDIG website (ICEDIG Project 2020). It highlighted a number of needs:

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

These points will need to be considered later when combining work on the technical capacity of digitisation centres, with establishing best practice policy for digitisation.

This report aims to identify strengths and capacity for digitisation within collection-holding institutions in the ICEDIG Project. We debated which methods were most appropriate to gather this information, as survey fatigue is potentially an issue across the ICEDIG partners. However, this tends to be a quick and potentially simple way of gathering information from institutions across Europe. For this reason, we decided to create a survey whilst bearing in mind the potential disadvantages of this technique. For example, questions and answers given can be misinterpreted or survey answers may lack detail or clarity.

All links referenced in this report were archived using the Internet Archive's Wayback Machine save page service on 02-07-2020.

### 1.1 Project Context

This project report was written as a formal Milestone (MS44) as part of Task 7.2 of the ICE DIG Project. It was previously made available to project partners and submitted to the European Commision as a report on 31 March 2019. While the differences between these versions are minor the authors consider this the definitive version of the report.

The original data were collected at the beginning of 2019 and do not necessarily reflect the current capacity of any institute included in the report.

## 2. Data Collection

### 2.1 Methodology

Due to the low number of institutions taking part in this task we combined the survey technique of data collection with an interview technique. We sent out a survey to six institutions:

1. Botanic Garden Mesie (APM)
2. The Finnish Museum of Natural History (LUOMUS)
3. The Unviersity of Tartu (UTARTU)
4. Muséum national d'Histoire naturelle (MNHN)
5. Naturalis Biodiversity Centre (Naturalis)
6. The Natural History Museum, London (NHM)

We refer to each institute using the abbreviated name given in brackets.

In addition, there have been reports and surveys completed previously with some information on technical capacity. Where possible, we used this existing information to make a first attempt to complete relevant parts of the survey. The initial approach taken for this analysis was the extraction of digital components from institutional strategy documents. Information on institutional priorities and digitisation capacity was extremely limited from these sources. The next focus was searching for relevant information in other recent EC project documents such as the SYNTHESYS+ proposal (Smith et al. 2019). However, this information was not provided specifically for this survey, and so was not suitably structured to answer many of our questions. To fill in the remaining gaps we emailed the survey to each institution, and provided the option of having a meeting to either go through the survey together or answer any remaining questions.

### 2.2 Survey Design

Interpretation can be a challenging topic when discussing digitisation and what is meant by fully digitising an object. For the purpose of this report and to create clarity, we decided to use the DiSSCo survey definitions for digitisation.

The survey (Suppl. material 1) looks at three main areas within digitisation that would provide useful information around institutional technical capabilities; ‘Current Status’, ‘Workflow’ and ‘Resources’.

‘Current Status’ has been included as it will give us a quick overview of what stage institutions are at in digitising their collections. We have the current status data from the previous DiSSCo survey, however this will give us the opportunity to update this information if needed. ‘Workflow’ will allow us to see what methods each institution uses to digitise their collections and how they prioritise what to digitise. We can then compare and contrast what works well, what could be improved and any potential barriers. Lastly, ‘Resources’ will reveal what institutions have at their disposal for digitisation and how this relates to workflow. The survey consists of open-ended and tick box questions. All questions have space in the last column for free text to ensure all possible data is captured.

## 3. Results

Each of the seven collection-holding institutions amongst the ICEDIG partners completed the survey, provided via a link to the appropriate Google Sheet. While a meeting or phone call was preferred to support the process, this was not always possible due to a combination of time constraints and the logistics of gathering the relevant data. We recognize that potentially not all the information available within institutions was included due to members of staff being absent or information not being found within the allotted time for the survey to be completed. All institutions were extremely helpful, replying in a timely manner within the deadline and completing as much of the survey as possible.

### 3.1 Current Status

The ‘Current Status’ section of the survey was pre-populated as much as possible with data from the DiSSCo survey. Each institution was asked to update the current status of digitised collections, if new data was available. This section was divided into three sections (along with the percentages for total collections):

- Number of specimens catalogued (i.e. records exist on in-house collection management system).
- Number of specimens digitised (i.e. specimen information in collection management system with partly or fully transcribed labels).
- Number of specimens fully digitised (i.e. specimen information in collection management system, with fully transcribed labels and images).

As previously mentioned, we used the DiSSCo definitions for all of the above. This remains a challenging subject, as interpretation of the terms (‘catalogued’, ‘digitised’ and ‘fully digitised’) can vary drastically and this was raised by most institutions.

Current status is summarised in Table 1 below.

Table 1.

Current status of digitised collections circa March to April 2019.

* records exist in in-house collection management system.

† specimen information in collection management system with partly or fully transcribed labels.

‡ specimen information in collection management system with fully transcribed labels, and images.

| Question | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Number of specimens *catalogued**** | 1,700,000 | 41,000,000 | 1,730,000 | 1,740,000 | 10,060,426 | 1,200,000 | 4,259,118 |
| **Percentage of collections catalogued** | 42.5 | 100 | 20 | 10 | 15 | 37 | 5 |
| **Number of specimens *digitised*†** | 1,700,000 | 10,250,000 | 1,730,000 | 1,300,000 | 7,150,585 | 554,000 | 3,600,000 |
| **Percentage of collections digitised** | 42.5 | 25 | 20 | 10 | 10 | 45 | 5 |
| **Number of specimens *fully digitised*‡** | 1,400,000 | 6,150,000 | 711,000 | No total estimates available | 1,869,945 | No total estimates available | 1,200,000 |
| **Percentage of collections fully digitised** | 35 | 15 | 8 | N/A | 3 | N/A | 2 |

## 3.2 Workflow

In order to understand each institution's digitisation process, we examined workflows to better understand the different work streams and tasks involved in digitising collections. We wanted to see how each institution streamlines the process in order to facilitate the complexities of digitisation.

**Where does your digitisation take place**?

All responding institutions currently digitise in-house, although APM also employs external contractors to digitise specimens on-site, noting that this makes it easier to accurately track progress in the project and minimise the risk of damage in transporting specimens. Several institutions (Naturalis, MNHN, NHM) have in the past outsourced digitisation to contractors such as Naturalis for off-site digitisation of herbarium sheets, the former two at scale and the latter as a smaller pilot project.

**What collections are you currently able to digitise**?

Each institution has a variety of different collections which in turn creates a variety of expertise when it comes to digitisation. Table 2 shows the collections that each institution can currently digitise.

Table 2.

Institutional digitisation capabilities by collection/specimen type.

| Collection/specimen type | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Herbarium sheets** | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Microscope slides** | Yes | Yes | Yes | Yes | Yes | No | Yes |
| **Vertebrates (dry preserved)** | N/A | Yes | N/A | Yes | Yes | Yes | No |
| **Spirit (liquid preserved) material** | No | Yes | Yes | Yes | Yes | Yes | No |
| **Mineralogical** | N/A | Yes | N/A | No | Yes | Yes | No |
| **Palaeontological** | N/A | Yes | N/A | Yes | Yes | Yes | No |
| **Anthropological** | N/A | No | N/A | No | Yes | No | No |
| **Pinned insects** | N/A | Yes | N/A | Yes | Yes | Yes | Yes |
| **Other** | No | No | Yes | Yes | No | No | No |

All institutions can digitise herbarium sheets, showing considerable expertise in this area across ICEDIG partner institutions, but also emphasising the relative simplicity of the workflow. Similarly, six out of seven institutions have the facility to digitise microscope slides. MNHN selected all collection types listed in the survey, suggesting capacity to digitise in some form a wide variety of collections. RBGK also included additional collections that they are able to digitise that weren't explicitly separated in the list: 'mycology', 'seed', 'DNA and Tissue Bank' and 'in vitro'. LUMOUS added in the 'other' option the ability to digitise galls with stacking equipment.

**What specimen handling techniques are used**?

Two institutions use either automated or assisted positioning of specimens. APM use a template that guides technicians on where to position the specimen before imaging. LUOMUS uses automated conveyor belt techniques as well as manual. RBGK stated that it mainly uses manual handling techniques, however the microscope slide workflows are slightly more automated with the use of a Zeiss Axio Scan. MNHN, NHM and UTARTU all use manual handling techniques.

**Do you currently have any high throughput digitisation activities going on**?

Five out of seven institutions (APM, Naturalis, RBGK, LUOMUS and NHM) reported to currently be running high throughput digitisation activities. Naturalis are working on

digitising bagged butterfly collections, RBGK are currently digitising 220 herbarium specimens per day and the NHM are currently working on high throughput projects for pinned insects, herbarium sheets and microscope slides. MNHN have finished their latest high throughput project, and have completed digitising their herbarium sheets.

**What formal processes are used to prioritise digitisation? For example, is there a scoring system in place**?

In the past, some institutions surveyed had designed formal prioritisation frameworks for collections digitisation. However, in some cases (NHM, APM) those institutions have not persisted with using them for various reasons. Naturalis have adopted a collection evaluation system that was developed by the Smithsonian which provides structured data to underpin the prioritisation process. A variant of this system ('Join the Dots') has been rolled out at the NHM, and is beginning to be incorporated into the prioritisation process.

Institutions which do not currently use a formal scoring process still use various criteria on a less formal basis to help them prioritise which collections to digitise (Table 3). APM are currently working on African and Belgian collections, as their scientists are working on these collections. The Belgian collections were also chosen due to being native, and the African collection due to their colonial past. RBGK align with their science strategy and/or their collections strategy. MNHN prioritise type specimens. In UTARTU it is the curators who make decisions on what to digitise, however they are currently in the process of formalising this procedure. At the NHM projects are assessed on a case by case basis, looking at a set of criteria, with external funding being the most important factor.

Table 3.
Prioritisation criteria for digitisation projects.

| Prioritisation Criterion | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Technical viability** | Yes | Yes | No | Yes | No | No | Yes |
| **Affordability/resource commitment** | No | Yes | No | Yes | Yes | Yes | Yes |
| **Scalability and strategic importance** | Yes | Yes | No | Yes | Yes | Yes | Yes |
| **Curatorial benefit** | Yes | Yes | No | No | No | No | Yes |
| **Research benefit** | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Public engagement** | No | Yes | No | No | No | No | No |
| **Funding potential** | No | Yes | Yes | Yes | Yes | Yes | Yes |
| **Technical innovation** | No | Yes | No | No | Yes | No | Yes |
| **Other** | No | No | Yes | No | No | No | No |

Research benefits were a consideration for all seven institutions when prioritising collections to digitise, with funding also being a factor for six out of seven institutions. Public engagement was the least popular criterion with only Naturalis taking this into

consideration. This was followed by technical innovation (3 of 7 institutions) and curatorial benefit (4 of 7 institutions). RBGK added an additional criterion: the time required for specimen selection, due to how their collections are arranged. They noted that it is prohibitively slow to select material by collector or even country, so taxonomic grouping or large geographical regions are preferred.

**What information resources are available to support digitisation activities**?

Five out of seven institutions (APM, RBGK, Naturalis, LUOMUS and NHM) provide written guidelines for handling specimens, while only two out of seven (APM and Naturalis) have guidelines on mounting/rehousing specimens. In the latter case, LUOMUS noted that their guidelines were not written down, while NHM commented that mounting and rehousing is commonly carried out by curators rather than digitisers (and where not, digitisers receive specialist training prior to the project). MNHN do not have written guidelines, but note that for both activities any manipulation is carried out by specialized technicians.

Six out of seven institutions have other relevant guidelines for digitisation processes. APM have guidelines for transcription of label information and for in-house imaging. Naturalis have health and safety guidelines. RBGK have data entry and also imaging guidelines that are adapted for each project. LUOMUS have frequently asked questions, and guides to georeferencing and transcription. The NHM have guidelines on recording information and semi-automated ingestion of data into the collection management system.

**Do you have mobile digitisation stations**?

Only one institution, RBGK, reported having a fully mobile digitisation station. This includes an adjustable desk, Mac, camera with adjustable column and a one sided open box with LED lighting at the top. This mobile station can be moved close to wherever the collections are being digitised. Some institutions have digitisation equipment that could technically be moved if needed, however it would not be an easy process and so remains static.

**How do you track movement of specimens to ensure they are returned to the correct location**?

A variety of answers were given for this question. At APM, the external company they hire to digitise collections have their own tracking system that uses QR codes. In the APM herbarium storage facility, the cupboards are divided into pigeon holes (64 per cupboard). When a pile of specimens are taken out of a pigeon hole and placed onto a trolley, a sheet is added to the pile with a QR code. An identical sheet with the same QR code is left in the empty pigeon hole.

Naturalis have coded their storage locations. All digitised specimens have this code with their standard location in the registration system, so they are always returned to the correct location after digitisation.

At RBGK, when collecting individual specimens for digitisation, tags are placed in the cupboard where the specimen is stored in place of the specimen so that it can then be

placed back in the correct location. There are also collection Excel spreadsheets that are completed by the digitiser. Here the name, region and folder information are recorded. If entire cupboards are being moved, the boxes of specimens are numbered for order. When there are new digitisers working, more qualified staff check that the specimens have been put back in the correct location.

LUOMUS have a drawer number tracking system. Curators at UTARTU, similar to RBGK, place markers on specimens to show to where it should be returned. MNHN have a tracking system (details on this system was not provided). The NHM were the only institution not to have a tracking system however specimens taken from storage are returned on a regular basis and most collections are indexed based on their taxonomic name or have location labels.

**Which protocols for barcoding are used in specimen digitisation**?

Six out of seven institutions used barcodes in some way, with UTARTU being the exception. At APM every specimen (even if there is more than one) on each herbarium sheet gets a barcode, and similarly at Naturalis each digitised object is given a barcode.

At RBGK, most specimens are barcoded. Microscope slides are manually barcoded before digitisation, then the software used during the digitisation process automatically reads the barcode and creates filenames containing these barcodes for images produced. Economic botany and spirits are not barcoded but numbered. Not all new herbarium accessions are barcoded however all new fungarium accessions are databased and there is also software currently being developed to print out barcodes on fungarium labels.

LUOMUS use encoded CETAF stable identifiers. MNHN have protocols for barcoding botany, QR codes for entomology and marine invertebrates. NHM have protocols for barcodes including generating and purchasing labels, attaching them to specimens and reading/recording them into the CMS.

**Quality Assurance policy/standards in place**?

Four out of seven institutions reported that quality assurnace (QA) policy or standards were in place on some level. APM were the only institution that did not report any conditions for QA being implemented. Circumstances that affected the degree of QA included the particular demands of the project (Naturalis) and resources available (RBGK), for both images and transcribed data.

RBGK also added that QA guidelines are defined on a project by project basis, and tailored to the size of each project. Checks by digitisers and peer reviews are carried out with an aim to check about 5% of work completed for a project. However, constraints of project demands and funding can again limit the amount of QA activities.

NHM conducts QA on images at a basic level, mostly looking at file names along with some random checks. For transcribed data, controlled lists are generated where possible along with manual checks of the data.

**Image elements included when digitising**

Three out of seven institutions include all of the five elements (colour calibration chart, scale bar, labels, barcodes, institute name) when digitising specimens (Table 4). No institution named any additional elements to those listed. RBGK include their institution name on the scale bar so that those two criteria are always paired. UTARTU stated that curators do not always include a colour calibration chart and scale bar, so this is only sometimes included when digitising.

Table 4.

Image elements included when digitising.

| Image element | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Colour calibration chart** | Yes | No | Yes | Yes | Yes | Sometimes | No |
| **Scale bar** | Yes | No | Yes | Yes | Yes | Sometimes | Yes |
| **Labels** | No | Yes | Yes | Yes | Yes | No | Yes |
| **Barcode** | Yes | Yes | Yes | Yes | Yes | No | Yes |
| **Institution name** | Yes | Yes | Yes | Yes | Yes | No | Yes |
| **Other** | No | No | No | No | No | No | No |

**How do you track time and costs of digitisation workflows**?

Five out of seven institutions reported using some kind of method to track time and cost of collection digitisation. Naturalis have team leaders for each digitisation team, whose responsibility it is to monitor and report results and capacity. At RBGK, digitisers fill out progress reports and flexi timesheets which record each day's activities along with timestamps on data entry, where possible. RBGK are currently using Toggl to categorise activities to see how long each section of the workflow takes. At LUOMUS they track time and cost on mass digitisation lines using logbooks. NHM also manually enter time and cost in spreadsheets, and MNHN complete a monthly evaluation of production by unit which is used in their global annual report.

## 3.3 Resources

**How can digitised collections be accessed by researchers or the general public**?

All institutions have made their digitised collections easily accessible to researchers and the general public. The main insitutional or national online collection is listed first, followed by their insitutional contributions to other large aggregator sites:

APM

- Institutional Collections: www.botanicalcollections.be

- Data Aggregators:
    - Europeana: https://www.europeana.eu/en/search?page=1&qf=DATA_PROVIDER%3A%22Meise%20Botanic%20Garden%22
    - GBIF: https://www.gbif.org/publisher/a344ee9f-f1b7-4761-be2c-58ee6d741395
    - JSTOR Global Plants: https://plants.jstor.org/partner/BR

Naturalis

- Institutional Collections: http://bioportal.naturalis.nl
- Data Aggregators:
    - Europeana: https://www.europeana.eu/en/search?page=1&qf=DATA_PROVIDER%3A%22Naturalis%20Biodiversity%20Center%22
    - GBIF: https://www.gbif.org/publisher/396d5f30-dea9-11db-8ab4-b8a03c50a862
    - JSTOR Global Plants: https://plants.jstor.org/partner/NHN

RBGK

- Institutional Collections: https://www.kew.org/science/collections-and-resources/data-and-digital/collections-catalogues
- Data aggregator sites:
    - iDigBio: https://www.idigbio.org/portal/recordsets/710a8a54-783c-41aa-ad9a-05544cdb4c55
    - eflora Virtual Herbarium: http://reflora.jbrj.gov.br/reflora/herbarioVirtual/ConsultaPublicoHVUC/BemVindoConsultaPublicaHVConsultar.do?modoConsulta=LISTAGEM&quantidadeResultado=20&herbarioOrigem=K
    - Europeana: https://www.europeana.eu/en/search?page=1&qf=DATA_PROVIDER%3A%22Royal%20Botanic%20Gardens,%20Kew%22
    - GBIF: https://www.gbif.org/publisher/061b4f20-f241-11da-a328-b8a03c50a862
    - Genesys: https://www.genesys-pgr.org/partners/39331cc7-91d0-4af7-8bb1-46824864c1c8

- GGBN: http://www.ggbn.org/ggbn_portal/search/browse?&institution=RBGK,%20London
- JSTOR Global Plants: https://plants.jstor.org/partner/K
- Plants of the world online: http://www.plantsoftheworldonline.org/

LUOMUS

- Institutional (National) Collections*1 : https://laji.fi/en
- Data aggregator sites:
  - GBIF: https://www.gbif.org/publisher/e5585950-488e-11db-a1c2-b8a03c50a862 and https://www.gbif.org/publisher/04fd2e13-6881-4e5c-9dd1-8fdd9ab993c1
  - JSTOR: https://plants.jstor.org/partner/H

MNHN

- Institutional Collections : https://science.mnhn.fr/ and https://www.mnhn.fr/fr/collections/ensembles-collections
- Data aggregator sites:
  - Europeana: https://www.europeana.eu/en/search?page=1&qf=DATA_PROVIDER%3A%22Mus%C3%A9um%20national%20d%27Histoire%20naturelle%22
  - GBIF https://www.gbif.org/publisher/2cd829bb-b713-433d-99cf-64bef11e5b3e
  - JSTOR Global Plants: https://plants.jstor.org/partner/P

UTARTU

- Institutional Collections: UTARTU contribute to national aggregators listed below with an overview here: https://www.natmuseum.ut.ee/en/biodiversity-informatics
- Data aggregator sites:
  - eLurikkus: https://elurikkus.ee/en
  - GBIF: https://www.gbif.org/publisher/0870a77b-587c-4369-a8ed-bc3d347b8e1c
  - Geoscience Collections of Estonia: https://geocollections.info/
  - PlutoF: https://plutof.ut.ee/

NHM

- NHM collections via Data Portal: http://data.nhm.ac.uk/
- Data aggregator sites:
  - Europeana: https://www.europeana.eu/en/search?page=1&qf=DATA_PROVIDER%3A%22Natural%20History%20Museum%20Library,%20London%22&qf=DATA_PROVIDER%3A%22The%20Trustees%20of%20the%20Natural%20History%20Museum,%20London%22
  - GBIF: https://www.gbif.org/publisher/19456090-b49a-11d8-abeb-b8a03c50a862
  - GGBN: http://www.ggbn.org/ggbn_portal/search/browse?institution=NHM,%20London
  - JSTOR Global Plants: https://plants.jstor.org/partner/BM

**Do you have any planned or current crowdsourcing projects**?

Three out of seven institutions are currently running active crowdsourcing projects, while Naturalis, LUOMUS, UTARTU and NHM do not have any current or planned crowdsourcing projects.

APM run and support the DoeDat project, a multilingual crowdsourcing platform based on the code of DigiVol and the Atlas of Living Australia.

RBGK are involved in DigiVol and always aim to have a number of crowdsourcing projects on herbarium and fungarium transcription each year. They complete around 30,000 specimen records per year through crowdsourcing. Overall on DigiVol RBGK have now completed over 40,000 tasks in 38 expeditions with 502 volunteers.

MNHN have a crowdsourcing programme called Les herbonautes which has been operational since 2013. It was developed for the transcription of herbarium labels and is now also used for paleontology.

**Do you support Digitisation on Demand requests**?

Six out of seven institutions support Digitisation on Demand requests. At APM, if there is a request for a physical loan and the specimens have not yet been digitised, they are first digitised so that the researcher can select from them for the physical loan. At the RBGK they have a maximum of 10 images per request, and specimens requested are logged through a loan management system. LUOMUS supports Digitisation on Demand but currently has no formal procedures in place. At MNHN Digitisation on Demand requests are processed on a web demand management tool (MNHN collections - Requests) and validation is completed by the collection manager. NHM do not currently support Digitisation on Demand, but are actively working on it.

**Do you have one or more specialised digitisation teams and what number of dedicated digitisers do you currently have**?

Six out of seven institutions report having a specialist digitisation team. Naturalis have two team leaders who are dedicated digitizers, and NHM has a specialised digitisation team consisting of five personnel. UTARTU do not have a specialised team or dedicated digitisers.

APM have their technicians complete two hours a week (*ca.* 150 images) of digitisation. They also have 13 dedicated digitisers, including 10 technicians, and three volunteers who each spend roughly three hours imaging per week.

At RBGK they currently have three core staff in the digital collections team, however within this team effectively only one person spends 0.25 of their time imaging. They also have one dedicated person for scanning microscope slides. Team numbers are variable and depend on what projects are going on, funding and what interns and volunteers they have.

LUOMUS have a new team that started this February consisting of personnel from four teams and 13 people. Permanent personnel consist of seven people. Two of these employees are full-time (1.0 FTE) and four employees are part-time (0.5 FTE), with six additional full time employees.

MNHN have specialised digitisation teams consisting of a systematic collection management team, and specialised platforms for 3D and CT scans. Within this team there are 4 personal linked to 3D platforms and CT scans and 80 technicians in the collection management team.

**What are staff trained in**?

Staff working on digitising collections are a valuable resource. We wanted to briefly look at the main skill set staff working within digitisation have (see Table 5). The skills of staff will obviously be tailored to what collections institutions hold. The area most covered by training was photography with six out of seven institutions having staff trained in this area. This was followed by scanning with five out of seven institutions having staff trained in this area. Only two out of seven institutions had members of staff trained in machine learning. RBGK added another area not listed in the survey which was training on R and SQL queries along with how to test new products and tools introduced. For example, a new platform such as Transkribus.

Table 5.

Staff training in digitisation and related skills/expertise.

| Skill/expertise | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Photography** | Yes | Yes | Yes | Yes | Yes | No | Yes |
| **Scanning** | No | Yes | Yes | Yes | Yes | Yes | No |

| Skill/expertise | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Other visible light imaging** | Yes | Yes | No | No | No | No | Yes |
| **Optical character recognition** | No | No | Yes | No | No | Yes | Yes |
| **Machine learning** | No | No | No | Yes | No | No | Yes |
| **Software development** | Yes | No | No | Yes | No | No | Yes |
| **Other** | No | No | Yes | No | No | No | No |

### What equipment is available?

A vital part of assessing technical capacity for digitisation is examining what equipment institutions have access to. Table 6 shows a list of equipment available at each institution.

Table 6.
Digitisation equipment available at each institution.

| Equipment Type | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Macrophotography setup (camera, stand & lighting** | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Image stacking photographic equipment** | Yes | No | No | Yes | Yes | Yes | Yes |
| **Multiple-camera setup (linked cameras, stand & lighting)** | No | Yes | No | No | No | No | Yes |
| **Custom specimen digitisation hardware (i.e. cradles, rigs, holders)** | No | No | Yes | Yes | Yes | No | Yes |
| **Herbarium imaging setup (camera, stand & lighting)** | Yes | Yes | No | Yes | Yes | No | Yes |
| **Herbarium scanner** | No | Yes | Yes | Yes | Yes | Yes | No |
| **Flatbed scanner** | No | Yes | No | Yes | Yes | Yes | No |
| **Other scanner (please specify)** | No | Yes | No | No | No | No | Yes |
| **Infrared scanner** | No | No | No | No | No | No | No |
| **3D laser scanner** | No | No | No | Yes | Yes | No | Yes |
| **CT scanner** | No | No | No | No | Yes | No | Yes |
| **Micro-CT** | No | No | No | No | No | No | Yes |
| **X-ray** | No | No | Yes | No | Yes | No | Yes |
| **Scanning electron microscope** | Yes | No | No | No | Yes | No | Yes |
| **Automated slide scanner (please specify)** | Yes | No | Yes | No | Yes | No | Yes |
| **Semi-automated microscope (please specify)** | No | No | No | No | No | No | Yes |

| Equipment Type | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Manual microscope (please specify)** | No | No | No | Yes | Yes | Yes | Yes |
| **Specialist microscope (please specify)** | Yes | No | No | No | Yes | No | Yes |
| **Other (please specify)** | No | Yes | No | No | No | No | No |

MNHN and NHM reported the most diverse lists of digitisation equipment, which is likely to be a reflection in part of the scope and heterogeneity of those collections. All institutions have access to a macro photography set up, showing that a basic capacity for 2D imaging is now quite ubiquitous across institutions.

Naturalis also described a setup of three conveyor belts with innovative software that enables rapid digitization while still ensuring high quality. This allowed for more than 4 million herbarium sheets to be scanned over a 10 month period.

**What types of software do you currently use in the digitisation workflows**?

Within digitisation workflows, a variety of software solutions may be used for to support the various steps such as imaging and transcription (Table 7).

Table 7.

Software used as part of digitisation workflows.

| Software Provenance | APM | Naturalis | RBGK | LUOMUS | MNHN | UTARTU | NHM |
|---|---|---|---|---|---|---|---|
| **Commercial** | Yes | Yes | Yes | Yes | No | Yes | Yes |
| **Open source** | No | Yes | No | No | Yes | Yes | Yes |
| **Developed in-house** | Yes | Yes | Yes | Yes | No | No | Yes |

Six out of seven institutions reported using commercial software as part of their workflows, including Aetopia, BGBase, Capture One, EMu, Microsoft Access, Photoshop, Sketchfab and the conveyor belt control system Digitarium (Tegelberg et al. 2017).

Four out of seven institutions used open source software, and five out of seven develop software products in house. LUOMUS have developed their own collections management system (Heikkinen et al. 2019), APM have developed their own automated barcode recognition software, as have NHM with image segmentation functionality also included (Hudson et al. 2015). RBGK reported that they have developed all of their collections management systems in-house, but are now looking to combine these and will probably not take the bespoke development approach for this.

## 4. Discussion

This survey provides an overview of the digitisation capacities of the participating institutions, but potentially reveals some interesting trends in addition to confirming information that might be well-known anecdotally.

The results reveal that most of the institutions are actively engaged in digitisation at scale, and that there is a strong current tendency towards on-site digitisation, even when employing the services of an external contractor. This suggests that all centres have some suitable space on-site for mass collection digitisation, but further conversations would be needed to assess how much those activities could be scaled up in those locations. It was also notable that outsourcing, both currently and historically, centred almost entirely on the digitisation of herbarium sheets (Le Bras et al. 2017, Anonymous 2016). As one would expect, herbaria are more focused on a smaller number of collection workflows (herbarium sheets and slides), with expertise and equipment more relevant to 2D imaging of flat objects. Although a number of institutions reported the capability of digitising most of the different kinds of collections, any existing and previous high-throughput digitisation projects reported were related to herbarium sheets, microscope slides or pinned (or bagged) insects.

The survey also indicates that centralised, specialist digitisation teams have become the norm, rather than relying purely on ad hoc databasing and imaging by curatorial and research staff. In some cases, this was a fixed, core team, and in others the number of digitisers varied depending on current projects. Barcoding specimens has also become standard practice across the surveyed institutions, at least for centralised, large scale digitisation projects. Tracking locations by electronic methods appears to be lagging behind somewhat, but there are a number of good systems employed within various institutions to act as references for others.

There was in general a reported lack of formal, structured methodologies for prioritising digitisation across the surveyed institutions. Some had attempted these in the past but not managed to embed them into standard practice. However, on an informal basis, many of the same criteria are being used in practice to prioritise projects. This suggests that a common framework, if suitably flexible and pragmatic, could be developed to assist institutions and digitisation centres in this work. Similarly, it's a significant task to write the various documents and guidelines to support digitisation workflows, and the gaps displayed in this data confirm that initiatives to start sharing these resources more widely are well-founded. This extends to better documentation on the software used in digitisation workflows, including the underlying business decisions on whether to develop in-house software or use off-the-shelf or open source solutions. It would be useful for other organisations to have an up-to-date list of recommended software used in digitisation workflows and how it is being used.

The survey data also suggested that quality assurance (QA) policies and standards are commonly lacking, minimal or ad hoc. This is an element of digitisation workflows which

appears to be commonly under-resourced and lacking in expert input. Some of these issues are discussed in more detail by Phillips et al. (2019) and Groom et al. (2019).

As has been the case for many previous surveys and conversations related to collections digitisation, one of the challenges has been to provide clear definitions for terms such as ‘catalogued’, ‘digitised’, and ‘mass’ or ‘high-throughput’ digitisation. While we aimed to be consistent as much as possible, there is still some room for interpretation which may have been reflected in some of the responses. However, we expect that wider feedback on the survey results should help to resolve some of these inconsistencies.

## Acknowledgements

We would like to thank our colleagues at the following institutions for providing the underlying survey data for this report:

- Botanic Garden Mesie (APM)
- The Finnish Museum of Natural History (LUOMUS)
- The Unviersity of Tartu (UTARTU)
- Muséum national d'Histoire naturelle (MNHN)
- Naturalis Biodiversity Centre (Naturalis)

We would also like to thank Daniel Mietchen for his editorial comments prior to publication.

## Funding program

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

## Grant title

ICEDIG – “Innovation and consolidation for large scale digitisation of natural heritage”, Grant Agreement No. 777483

## Author contributions

**Authors**:

**Naomi Cocks**: Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualisation, Writing - Original Draft. **Laurence Livermore**: Validation, Writing - Reviewing and Editing. **Vincent Stuart Smith**: Conceptualization, Supervision, Writing - Reviewing and Editing. **Matt Woodburn**: Conceptualization, Methodology, Writing - Reviewing and Editing.

Contribution types are drawn from CRediT - Contributor Roles Taxonomy.

## References


- De Smedt S, Bogaerts A, Stoffelen P, Groom Q, Engledow HR, Sosef M, Van Wambeke P, Dessein S (2016) DOE! Mass digitisation of the BR Herbarium at Botanic Garden Meise. Missouri Botanical Garden Open Conference Systems, TDWG 2016 ANNUAL CONFERENCE. TDWG 2016 ANNUAL CONFERENCE, Santa Clara de San Carlos, Costa Rica. URL: https://mbgocs.mobot.org/index.php/tdwg/tdwg2016/paper/view/1146
- Groom Q, Dillen M, Hardy H, Phillips S, Willemse L, Wu Z (2019) Improved standardization of transcribed digital specimen data. Database 2019 https://doi.org/10.1093/database/baz129
- Heikkinen M, Riihikoski V, Kuusijärvi A, Talvitie D, Lahti T, Schulman L (2019) Kotka - A national multi-purpose collection management system. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37179
- Hudson L, Blagoderov V, Heaton A, Holtzhausen P, Livermore L, Price B, van der Walt S, Smith V (2015) Inselect: Automating the Digitization of Natural History Collections. PLOS ONE 10 (11). https://doi.org/10.1371/journal.pone.0143402
- ICEDIG Project (2020) Policy Analysis. https://icedig.eu/content/policy-analysis. Accessed on: 2020-6-03.
- Le Bras G, Pignal M, Jeanson M, Muller S, Aupic C, Carré B, Flament G, Gaudeul M, Gonçalves C, Invernón V, Jabbour F, Lerat E, Lowry P, Offroy B, Pimparé EP, Poncy O, Rouhan G, Haevermans T (2017) The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. Scientific Data 4 (1). https://doi.org/10.1038/sdata.2017.16
- Phillips S, Dillen M, Groom Q, Green L, Weech M, Wijkamp N (2019) Report on New Methods for Data Quality Assurance, Verification and Enrichment. Zenodo https://doi.org/10.5281/zenodo.3364509
- Smith V, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston E, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e46404
- Tegelberg R, Kahanpaa J, Karppinen J, Mononen T, Wu Z, Saarenmaa H (2017) Mass Digitization of Individual Pinned Insects Using Conveyor-Driven Imaging. 2017 IEEE 13th International Conference on e-Science (e-Science) https://doi.org/10.1109/escience.2017.85


## Supplementary material

**Suppl. material 1: Survey Template** doi

**Authors:** Naomi Cocks, Matt Woodburn
**Data type:** Excel spreadsheet
**Brief description:** An Excel version of the Google Sheets template used to collect the data for this report.
Download file (25.29 kb)

## Endnotes

[*1] LUOMUS manage the FinBIF site which aggregates national natural history data for collections and observation records.