Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:
Deliverable **D3.3 / R / PU**
Work package **WP 3 /** Lead**: PIC**
Delivery date M17

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

D3.3 State of the art and perspectives on mass imaging of skins and other vertebrate material

Myriam van Walsum, Steven van der Mije, Agnes Wijers, Luc Willemse

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

Abstract

The topic of this report is applications of conventional photography and automated imaging for three-dimensional dry collections. No automated imaging solutions were found for these types of specimens which is due to variability of material and complexity of handling these specimens. Due to the variability of object types, shapes and research interest, mass imaging for scientific purposes is not realistic. Imaging labels for data entry and databasing purposes should be achievable. We recommend a process based approach of record creation with minimal data, combined with label imaging. This can be done very rapidly with minimally trained workers. The label images can then be used subsequently, during a separate phase, for data entry. Especially when data entry is partial or includes interpretation, the images provide potential for future use. It also has to be considered if including the specimen in the image adds enough value to justify the extra effort. We also provide decision trees to help design a demand based approach.

Key recommendations

- ➔ Pure research driven imaging can never be achieved in a mass process. Mass can only be achieved for databasing, either from the object or from photos.
- ➔ There is no miraculous mass digitisation solution for three-dimensional objects; the best that can be achieved is to gain efficiency through large scale projects or long term programs.
- ➔ Imaging the labels for subsequent (partial) data entry is the most efficient approach for many cases: division of less and more skilled work, less handling = lower risk, benefits of full data capture etc.
- ➔ When imaging of labels is planned, it has to be considered if including some specimens in the images is efficient.
- ➔ At minimum, enable tracing the image to the specimen, by the inclusion of these image elements: institute code and collection number. Desirable: scale bar.
- ➔ Speed and quality both depend on the simplicity of the workflow. The more choices and adjustments are needed, the greater the risk for errors and poor data.
- ➔ This also applies to imaging: make use of a professional photographer to establish the proper settings for the collection and design the workflow in such a way that there are minimal adjustments needed by the operators. For three-dimensional collections depth of field is especially important. This can be maximised through various variables.
- ➔ While curators may receive training to digitise and to make use of digitised data, the greatest user group are the researchers. To achieve the full potential of digitisation it is necessary to ensure that they have the skills to work with the digital data.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Contents

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Introduction

This task investigates applications of conventional photography and automated imaging of both specimens and/or specimen labels on a large scale. The following points were part of the task description and will be discussed in this report[1]:

1. What is the definition and demand of mass digitisation?
2. What are the current and near-term solutions?
   2.1. Case studies
3. What are the effects of scaling up imaging capacity to millions of specimens a year?
4. What is the efficiency gain of combining imaging and digitisation with curatorial processes like repacking, barcoding, cleaning and sampling?
5. What is needed to minimise risks and maintain high throughput while keeping costs low?
6. What is the information required by researchers?
7. Which existing automated imaging methods fit these collections compared to the information required by 1 and 5?
8. Is imaging beyond 2D useful?
9. Adept recommendations with regard to
   9.1. criteria,
   9.2. specifications,
   9.3. technical solutions,
   9.4. workflows,
   9.5. outsourcing,
   9.6. health & safety.

Current state of vertebrate digitisation

Compared to the quantity of specimens in entomological collections and the relative amenability for mass digitisation of herbarium sheets, vertebrate collections are usually smaller and associated with complicating factors during digitisation. To the general public vertebrate collections are often more visible and of higher interest. The scope of this report includes skins, disarticulated skeletal material, mounted skins and skeletons, and other material such as nests and eggs - as long as it is dry. However, there are other collections that share the three-dimensional, dry and unpinned properties, such as dry molluscs and other invertebrates, palaeontological and mineralogical objects.

Availability of images and data is a necessity for research and the verification of results. It has been noted for instance that the importance of voucher specimens is not fully understood by some researchers. The lack of citations of voucher data in GenBank was pointed out by Beaman and Cellinese (2012). Simply referring to a taxonomic identification does not eliminate the possibility of a misidentification, which could have serious

[1] The numbers are used in the headings to refer to the sections where the topic is discussed.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

consequences to phylogenetic trees. Digitisation enables the researcher to reference the sampled individual and imaging adds various benefits on top of (often limited) data entry. This report investigates these benefits and approaches to imaging for dry three-dimensional natural history collections.

Lack of planning and work description. Or: "Cottage industry"

Kalms (2012) stressed that digitisation is not an optional extra, but an integral part of natural history collection management. Digitisation should not be done haphazardly, but with a strategy and defined process, however it is not necessarily practical or required to digitise and especially image each object in the collection. The approach taken by Australian Museum and South Australian Museum is "Just Do It": begin with a contained and simple project to build experience with digitisation and all the aspects that are involved, to build on for future projects. Defining manageable portions of the collection to digitise and establishing digitisation protocols help to make digitisation efforts workable and durable. The GBIF task force advises a tiered strategy: first rapid and less expensive steps to make collections accessible, then moving on to more expensive and time-consuming steps to capture more detailed data and images (Krishtalka et al 2016, p5). The task force even goes so far as to call the current digitisation efforts "cottage industry". Blagoderov et al (2012) argue that mass digitisation can only be achieved when the limiting steps, such as databasing, are minimised and industrialised. They also point out that the common solution to prioritise based on user demand (see Berents et al 2010) leads to a variation in digitisation status within a collection. This will lead to more effort required to fill in the gaps at a later date to bring the whole collection to a certain minimum digitisation level.

The survey undertaken by Vollmar et al (2010, p109) concluded that costs of digitisation is the greatest impediment, as it reflects on funding staff and technology. This results in digitisation being done opportunistically (during regular collection maintenance work) and without strategy. This then, leads to non-standardisation and decrease of usability for potential users.

A survey undertaken for Synthesys3 provides some insight into digitisation capacity for various collection and storage types at a number of European institutes (Synthesys3 2017). Only three out of 18 responding institutes (excluding those with only botanical collections) indicate not having any capacity to digitise vertebrate material. One specified to have 70% digitised, of which 16% imaged. Another said to have 70% of their vertebrate collections digitised, while it remains unclear whether this includes imaging. The other respondents indicated that less than 10% had been digitised. Other dry collections like palaeontology have seen comparably little digitisation, especially with regards to images. 64% of respondents to the survey undertaken by Vollmar et al (2010, p103) indicate that some imaging was being done, usually taking <5 to 10 minutes per specimen (60% of 64% of total respondents).

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The final report of the rapid digitisation pilot project funded by the Atlas of Living Australia at the Australian Museum and the South Australian Museum claims "image digitising is the new databasing" with the primary benefit of creating a readily accessible digital voucher of each specimen and its labels for verification and reference (Australian Museum, 2011). Many digitisation projects since then have not included imaging of (all) specimens, which shows that the new databasing has not caught on everywhere.

The GBIF task force on accelerating discovery did find through their survey that 86% of respondents indicate to have started databasing. Challenges to establishing digitisation protocols and executing them are discussed in Kalms 2012 (p26) and the GBIF report with survey on impediments to digitisation. The impediments boil down to funding, effort required, personal expertise required and the feeling that errors in the data preclude digitisation and publishing. By establishing simple, efficient and affordable protocols for databasing and imaging , most of these impediments can be overcome.

Looking at the data available on the GBIF portal for vertebrate collections, it is clear that not many occurrences have associated images. For example, only 14.785 out of 1.072.647 Chiroptera preserved specimens have an image in GBIF. Looking closer, most of these are from Naturalis' mass digitisation project or catalogue scans from NHM. In the case of Passeriformes, 154.469 records out of 4.622.740 preserved specimens have an associated image. Table 1 shows the distribution of imaged preserved specimens among the collections with high image rates. A few interesting points can be made from this. Firstly, only 2,9% of Chordata preserved specimens in GBIF have an associated image. Second, of the top 16 institutes with Chordata images, only 4 are European. The top imaging institute is NHM, but most images are in fact scans of catalogues, which can't be filtered out in GBIF. NHM's own data portal[2] lists, in the phylum Chordata, 104.823 entries with images of registers and 16.053 entries with images of specimens (15.133 from palaeontology collections). Third, some collections have imaged very specific collections, such as fish groups, reptiles and amphibians. Not all institutes have uploaded their data into GBIF which means that this small analysis is not representative of the current state of the world's natural history institutes' vertebrate imaging efforts but it does give a good indication of the current status.

The variety of image layouts for vertebrate specimens available through GBIF, shows that there are few standards: inclusion and visibility of labels, scale bar and colour checkers. Examples are present where the image is a composite of multiple photos with the label cropped off. Other examples includes TIFF stacks from CT scans or X Ray images.

Detailed descriptions of finished and current vertebrate digitisation projects which include an imaging component will be discussed in section Current and near-term solutions [#2.1].

[2] Source: https://data.nhm.ac.uk/dataset/56e711e6-c847-4f99-915a-6894bb5c5dea/resource/05ff2255-c38a-40c9-b657-4ccb55ab2feb?q=&field=associatedMediaCount&view_id=203a0ae5-6a14-480a-a407-27eeb9373858&value=&filters=_has_image%3Atrue%7Cphylum%3Achordata (Date accessed 29/04/2019)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

	Chiroptera	Passeriformes	Chordata
Total preserved specimens	**1.072.647**	**4.622.740**	**26.562.126**
Total preserved specimens with image	**14.785**	**154.469**	**772.346**
Natural History Museum UK*	3.325	44.832	267.065
Natural History Museum UK (corrected for specimen images with numbers from their portal, including palaeontology)[2]	-	-	16.053
California Academy of Sciences**	0	0	224.689
Naturalis Biodiversity Center	10.394	75.644	97.136
National Museum of Natural History, Smithsonian Institution	167	1.429	38.167
Museum National d'Histoire naturelle	188	3.759	22.957
Museum of Vertebrate Zoology, University of California, Berkeley	79	8.192	15.369
Yale University Peabody Museum	7	3.720	13.886
Museum of Comparative Zoology, Harvard University	141	325	13.286
Field Museum of Natural History	181	595	11.421
Museums Victoria	33	5.971	11.006
The South African Institute for Aquatic Biodiversity	0	0	8.069
Estonian Museum of Natural History	33	1.325	6.070
Staatliche Naturwissenschaftliche Sammlungen Bayerns***	0	0	5.970
Australian Museum	20	238	4.583
Chicago Academy of Sciences	114	1.093	3.442
Berkeley Natural History Museums	23	17	3.260

Table 1. Counts of preserved specimens with image in GBIF (only collections above 3.000 Chordata specimens). Two taxonomic groups have been used as example (Chiroptera (bats) and Passeriformes ("songbirds" or perching birds)) and the whole Chordata group, which includes more than only Vertebrata[3]. In descending order for Chordata results.
*Mostly scans of catalogues
**Mostly Reptilia and Amphibia
***Mostly Actinopterygii (ray-finned fish)

Collection managers working on one group often are also involved with one or more of the other groups: be it mammalogy, ornithology, herpetology, ichthyology and to a lesser extent invertebrate zoology and (in)vertebrate palaeontology (survey results from Vollmar

[3] Source: https://www.gbif.org/occurrence/gallery?basis_of_record=PRESERVED_SPECIMEN&media_type=StillImage&taxon_key=44 (Date accessed 29/04/2019)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

et al 2010). Digitisation for these groups is marked by the greater time consumption of specimen handling and preparation.

Ornithology is characterised by great variety of parts of a single specimen scattered across the collection: besides the usual skeleton, skin and liquid preserved parts, there may also be eggs and nests, as well as sources such as photographs and recordings of bird sounds prior to collection. These sources also need specific database fields to accommodate the data. Ichthyological and herpetological collections consist for a large part of liquid preserved specimens, which are discussed in ICEDIG deliverable 3.4 (Van Walsum et al 2019). Vollmar et al describe that precision of location recording is especially important for specimens collected in rivers, as rivers may be thousands of kilometers long. Invertebrate collections often contain lots consisting of hundreds to thousands of organisms, which makes them difficult to track during loans. (In)vertebrate palaeontology collections are marked by the importance of locality data, almost on the same level as taxonomic data. This data needs to includes geologic information such as unit, age, series, formation, beds/members/zones. This requires a specific database structure to accommodate this data. Further, for palaeontological specimens it may be difficult to pinpoint preserved part, they can contain multiple specimens of varying taxa and are more likely to have multiple interpretations associated with them. Recording their label data is especially important.

The role of imaging in mass digitisation [#1]

The question that arises with the design of any digitisation project with an imaging component, is exactly what to image. First the decision whether or not to image has to be made, as data entry directly from the object is in many cases faster than imaging. As postulated by Berents et al (2010), the internal driver of curation is not sufficient for the great effort required for digitisation so that researchers and other external users need to drive digitisation priorities. This is also valid for deciding which specimens to image and which details need to be included, however researchers often require highly detailed images for taxonomic research which precludes mass digitisation.

For collections such as herbarium sheets and microscope slides, which are essentially 2D, automated imaging has been developed. Their two-dimensional nature simplifies the question of which views to image, how to position and stabilise the object, lighting and shading, and depth of field. Also, their variation in shape is constrained, as well as variation in size. This has direct impact on ease of handling, especially by non-experts. The threshold of cost-efficiency of imaging has been lowered through conveyor belts and multi-image stitching.

For most collections of three-dimensional objects this has not been achieved despite efforts in the past (for instance see section Current and near-term solutions [#2.1]). The above mentioned complicating factors are not resolved and some are intrinsic to their three-dimensional nature. The variability of the material due to preserved parts and the difference in taxonomic features have various impacts on the imaging workflow. Either it

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

takes a lot of time to adjust the workflow to capture the traits specific to this group, or a portion of the images will not be sufficient for much more than label data.

If imaging for vertebrate collections is (relatively) slow and complex, then there have to be good reasons to include it as part of digitisation. There are two approaches to imaging: one focuses on capturing the label to record the data, the other on the specimen for curation, presentation and research purposes. The two are not mutually exclusive, but because it does affect the imaging strategy, it is important to keep the goal clear to all parties involved. Where relevant, the differences between the two approaches will be made explicit in this report.

It is also not necessary to image every specimen. Overview photos of drawers and boxes containing multiple specimens can be useful for inventory purposes and may potentially capture labels. Some taxonomic groups are not well documented, in which case imaging one specimen per species can be valuable. For example, this approach was adopted by the Natural History Museum Rotterdam (The Netherlands) for the Mollusc collection, which resulted in an increase of attention not quite in proportion with its size. This shows that digitisation with imaging is a great tool for increased use and how small collections can be increasingly made visible. Taxonomy resources such as WoRMS[4] and Encyclopedia of Life[5] highly benefit from having one or more photos available to illustrate a taxon. An approach adopted by many institutes is to only image a selection of material during a larger digitisation project. For example: types, deteriorating, handwritten and poorly legible labels, high profile specimens. However, there are also reasons to image every label and every specimen, which are elaborated in Criteria [#9.1].

Effects of scaling up: minimising risks while maintaining high throughput and low costs [#3&5]

The anticipated positive effects of mass digitisation and imaging are a great jump in accessibility of collections. Following from that collections will be better protected and there will be more research and funding. The GBIF survey (Krishtalka et al 2016, p15) lists the top 8 perceived benefits of digitisation: “increased use of collections, increased exposure, better knowledge of holdings, better management of data, digital data preservation, enhanced data quality, new skills for staff and better management of physical specimens. Thirty percent reported new communities using the data, 35% saw increased publicity and reduced physical handling of the collection and 40% cited increased use of their collection data in research and publications, as well as increased public awareness of the importance of collections.” This increase in use, requests and inquiries was already noted by Vollmar et al (2010, p100), with the potential effect of increasing the workload for collection management. In any digitisation plan, this future effect needs to be kept in mind.

[4] World Register of Marine Species http://www.marinespecies.org
[5] https://eol.org

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

However, there are other aspects that need to be considered, even before embarking on mass digitisation efforts.

The stress on the IT infrastructure will need attention, such as temporary and permanent storage space, network load during data transfer and computing power for batch image processing and quality control scripts. Several finished and active digitisation projects have found that their greatest bottleneck was IT infrastructure. When a project applies the strategy of minimal initial data entry with label imaging for the purpose of data entry there will likely be a great quantity of image files processed and stored on the network, so this should be anticipated. If data entry is done more elaborately with less imaging, then number and efficiency of registrars is most likely the limiting factor. See for elaboration ICEDIG WP6[6].

Due to the upscaling of efforts, generally with non-experts, and emphasis on speed, there is a greater error risk. This applies both to data entry and specimen handling (for instance separation of specimen and label which is a great concern of curators). Errors in data entry should be checked through random checks and logical filters that can be applied to the data, such as impossible or highly unlikely combinations of collector & year, collector & region, species & region. Using thesauri for taxonomy, collectors and localities can help reduce data errors, but do require more effort at the start. For image quality control see ICEDIG deliverable D3.1 (Nieva de la Hidalga et al 2019a).

Further, during the design of any digitisation effort adequate support by curators needs to be taken into account (Australian Museum 2011, p26). Logistics can be a serious bottleneck for the digitisation effort, and is often underestimated with respect to demands on collection managers. Additionally, any specimen flagged for maintenance or as “puzzle” can pull curators away from their own priorities.

Efficiency gain of combining with other tasks [#4]

All handling actions have some risk for errors or damage to the specimen/label, so minimising the handling of specimens is crucial. For this reason, it may be better for risk management to image the objects and labels and perform data entry from those images, so that specimens don’t have to be handled again. Combining tasks such as recoding, barcoding, cleaning, repacking and tissue extraction with digitisation efforts can lead to an efficiency gain in certain situations. Especially when a collection is being rehoused and all specimens must be handled anyway, it is opportune to run the specimens through a digitisation project as well.

On the other hand, more complex tasks as part of a larger process may produce more errors than when these are done separately by specially trained workers. This includes collection maintenance, taxonomic update checks and reorganisation of the collection, processes which require specialist skill sets.

[6] Deliverables will be available at https://icedig.eu/content/deliverables

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Depending on the scale of the project and complexity of tasks, it can be advisable to divide the steps into clearly defined chunks, handled by dedicated staff, even if it is only for that shift. Finished projects learn that the better the pre-work by curators, the more smoothly the digitisation project can run.

Information required by researchers [#6]

The 21st century is marked by big data: connecting complex and large data sets. Data needs to be searchable through good descriptors to be fully made advantage of, while also acknowledging the long history of biodiversity collections, meaning that standardisation and pollution of data is present. During new imaging efforts, it should be tried to plan ahead so that this new dataset can be adequately searchable. Other new data sources, such as GIS-based spatial analyses, genomic research, isotope analysis, (micro)CT and 3D scans with geometric morphometric analyses, are adding to the complexity of biodiversity research and should be tied together (Cook and Light, in press). As extinction of populations and species proceeds, natural history collections are bound to provide the primary source for new research.

When there are specific research questions or purposes that the images are meant to address, this likely implicates that digitisation on a mass scale is not achievable. This could be specific views or zooming in on specific features, such as teeth and palate.

Image standards such as defined by GBIF[7][8] and TDWG should be used and made explicit in the metadata, where possible. Additionally, it can be useful to record what exactly the image is of: catalogue, specimen, label, microscope slide. Potentially even a narrowed down description of the specimen: skeleton, skin, TIFF stack, 3D render etc. This doesn't necessarily match up with preservation type. See for example this record[9] from the NHM where in Dynamic properties, amongst other information, details on image category are recorded: *"imagecategory": ["Register;Specimen"]* to indicate that the record is associated with images of both register and specimen. Currently GBIF and others don't provide for a way to filter this, but hopefully this can be added in the future, while NHM's own portal[10] does allow the use of this metadata for filtering results.

Atlas of Living Australia recommends including the unique identifier in the image[11]. In this case, when the image is separated from the rest of the specimen data, it can be traced back to its institution and specimen. For baseline purposes, ALA also recommends including the current species identification. By including the unique identifier and institute code (if not part of the UID) visibly in the image, the image can be traced back to the specimen if it becomes isolated, e.g. on the internet. A scale bar is advisable too. Some workflows assign

[7] Metadata fields from GBIF's Darwin Core Extension on Simple Multimedia: http://rs.gbif.org/extension/gbif/1.0/multimedia.xml

[8] GBIF Audubon Media Description: http://rs.gbif.org/extension/ac/audubon.xml

[9] https://www.gbif.org/occurrence/1056382890

[10] http://data.nhm.ac.uk/object/7c4baded-ce86-43f1-81ef-265bdf030f0a

[11] https://www.ala.org.au/who-we-are/digitisation-guidance/hints-and-tips/ (accessed 2019-03-19)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

unique identifiers after imaging and subsequent automated record generation (Blagoderov et al 2012, p142), this would mean that the label with unique identifier would not be visible in the image.

The Mollusca curators at Naturalis remark with some surprise that the images taken during the mass digitisation project have even been used for publication, while that was never an intended use and they would have been very happy to take better pictures if the author had contacted them. Further, since publication of data and images online in 2014 there have been less than ten request for better images because the published ones don't suffice.

In answer to which minimal MIDS level is required to make most digitisation efforts worthwhile, all interviewed curators from Naturalis responded that MIDS-2 is needed to answer the requests that they receive. Any additional information above MIDS-2 is too specific to research questions, taxonomic group and/or preparation type, to make general recommendations.

The survey by the GBIF task force on accelerating discovery received these answers for needed metadata[12]: taxonomic and spatial data (both >50%), type specimen information (45%), percent collection digitised, notable publications, percent georeferenced, place name coverage, name of collection manager, collection size and name of curator (Krishtalka et al 2016, p15). Making the data discoverable after digitisation through adhering to community endorsed metadata standards and connecting to data aggregators such as GBIF.

Usefulness of imaging beyond 2D [#8]

Due to the 3D nature of most vertebrate objects, 2D imaging is not always sufficient to capture essential features of these objects. Especially measurement recording for these objects from 2D images is problematic, because scale bars are only valid for the plane in which they sit due to factors such as perspective and parallax (Ariño & Galicia 2005, p91+108+117). Geometric morphometric analyses of 3D objects from 2D imagery has been done, but it is much preferable to do this on 3D scans. Additionally, it has been well established that measurements on 3D scans are much easier and more reliable and repeatable, as well as much less damaging to the specimen.

This means that 2D+ and 3D imaging is relevant, but due to capture speed and level of detail required for specific uses, the uptake of these techniques has been very limited. 3D imaging in the form of (micro)CT has been extensively used for very specific objects for research purposes. Other types of 3D imaging such as laser scanning, photogrammetry and structured light scanning, have been used for research purposes or public outreach in small projects. The captured data is often not stored for long term archiving and easy retrieval through the database. The characteristics of vertebrate material need to be considered when selecting an appropriate 3D imaging technique. Feather and fur are problematic,

[12] Multiple answers were allowed.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

especially when glossy and iridescent. Soft materials such as skins can be problematic for imaging techniques that require the object to be moved to capture the underside of the initial position too, which in fact means most 3D imaging methods. Complex structures such as crania and sacra can only be fully captured through techniques that penetrate the surface, such as CT scanning, otherwise the various apertures and occluding structures will lead to missing data. Thin structures such as bird bones may be problematic for laser scanners as they let through the laser, causing noise and distortion. Glossy surfaces on bones and other elements may cause noise in laser scanning, structured light scanning and photogrammetry. Laser scanning can also be problematic on very greasy bones, as the grease scatters the laser. Glossiness, greasiness and blackness can be alleviated by dusting the object with talcum powder, optionally sprayed on in an alcohol solution, but it has to be considered first whether this can be removed from the object afterwards (brushing, compressed air or rinsing) or poses other contamination.

The benefit of 3D scans can be to study the shape of the object without colour. In some cases the colour information is irrelevant (as for many bones and fossils) or actually obscures the shape. Not all 3D imaging techniques can capture colour, or allow controlled colour capture. Because photogrammetry is based on regular photography, it is best suitable for high quality colour information in 3D imaging as it can be kept to the high standards for archiving and specimen photography. Another technique is Reflective Transformation Imaging (RTI) which captures geometry and material properties by photographing the object multiple times from the same position, but with changing light directions. This allows visualising geometry without colour, iridescent materials such as bird and butterfly patterns as well as small surface details.

See ICEDIG deliverable D3.7 (Nieva de la Hidalga et al 2019b) for further information on 3D techniques and considerations for integration in a digitisation project.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Current and near-term solutions [#2.1]

Only a few vertebrate collections have been imaged during formal digitisation projects, let alone in mass workflows. Complexity and other priorities are at the root of this. Smaller scale efforts have been organised to digitise vertebrate specimens. Much of the information in this section comes from interviews and email correspondence undertaken for this report.

Dr. S.A. McLeod, Collections Manager of Vertebrate Paleontology at Natural History Museum of Los Angeles (NHMLA), describes that they don't have a formal imaging project but that imaging has been done for smaller numbers of specimens. The collection holds about 170.000 catalogued specimens and twice as many uncatalogued specimens which are sorted by locality which allows retrieval. The policy of NHMLA is to photograph everything that goes out on loan, to serve as a description of the specimen and as a snapshot of its condition. A good description of specimens would be desirable but is very time-consuming, so images can fill this need. A large part of his collection would require a microscope to image properly, like individual fossil teeth. Some parts of the collection are on the other extreme of the size range. In some cases, the fossil specimens need to be coated to deal with distracting reflection or colouring. Of the images that have been taken so far, not all are made available online because of the effort needed to upload everything correctly. For some very specific use cases drawer imaging has been used, but in general there is little utility. Due to the great number of specimens and the significant proportion that has not been catalogued to specimen level, barcodes are not in use.

However, the Vertebrate Paleontology department of NHMLA is running a 3D imaging (photogrammetry) project for types in the collection[13]. This project started when two experts offered their volunteer services. As part of this project specimen rehousing also was done. The photogrammetry setup includes a semi-automated turntable, lights with diffuser boxes and a DSLR camera mounted on a tripod with processing done in Agisoft Photoscan. Because the models haven't been available (on Sketchfab) for very long, it isn't possible to say anything about increased collection visibility and use. If possible, the aim is to make the quality high enough for research purposes. Archiving includes both input (photos) and output (finished 3D models) on an internal drive in a way that it can be retrieved, but there is no formal protocol for storage. This poses risks for durability of the data.

In Rotterdam, The Netherlands, Het Natuurhistorisch Museum Rotterdam (NMR) is located which houses around 390.000 collection units. The largest part of the collections is invertebrate while approximately 5% is vertebrate material including fossils. The whole collection is managed by one collection manager and one registrator, and supported by expert volunteers (called honorary conservators) who each manage one of five

[13] See https://sketchfab.com/blogs/community/3d-digitization-of-the-nhmla-vertebrate-fossil-type-collection/ for a description. (Accessed 05/04/2019)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

subcollections. Digital registration was adopted during the late '90s for which an online registration solution was developed (now at version 3.0). Imaging started in 2014 and raised issues with associating database record with image files. Currently 18.000 images are available online. They have seen an increase in feedback from taxonomists and loan requests. Also, it has put the collection on the world map so that researchers travelling to the large collections in London or Leiden will now also add Rotterdam to their itinerary.

The collection manager for molluscs initiated a project to photograph that collection. For mollusc species identification two views need to be imaged requiring exact positioning otherwise distortion interferes with identification. A photographer is involved to ensure that lighting and depth of field are correct. This way, NMR says, photographing is useful, but mostly applicable to types.

The mammal collection at the Natural History Museum in London is of such a size that it is dispersed over 9 floors, 2 areas in the basement and off-site storage. It has on average 5-6 visitors each day, which takes up a lot of time of curatorial staff. Label degradation and dissociation is a major concern, so that effort is taken to store the labels in archival bags.

NHM is currently not working on a workflow for mass digitisation/imaging for vertebrate specimens.

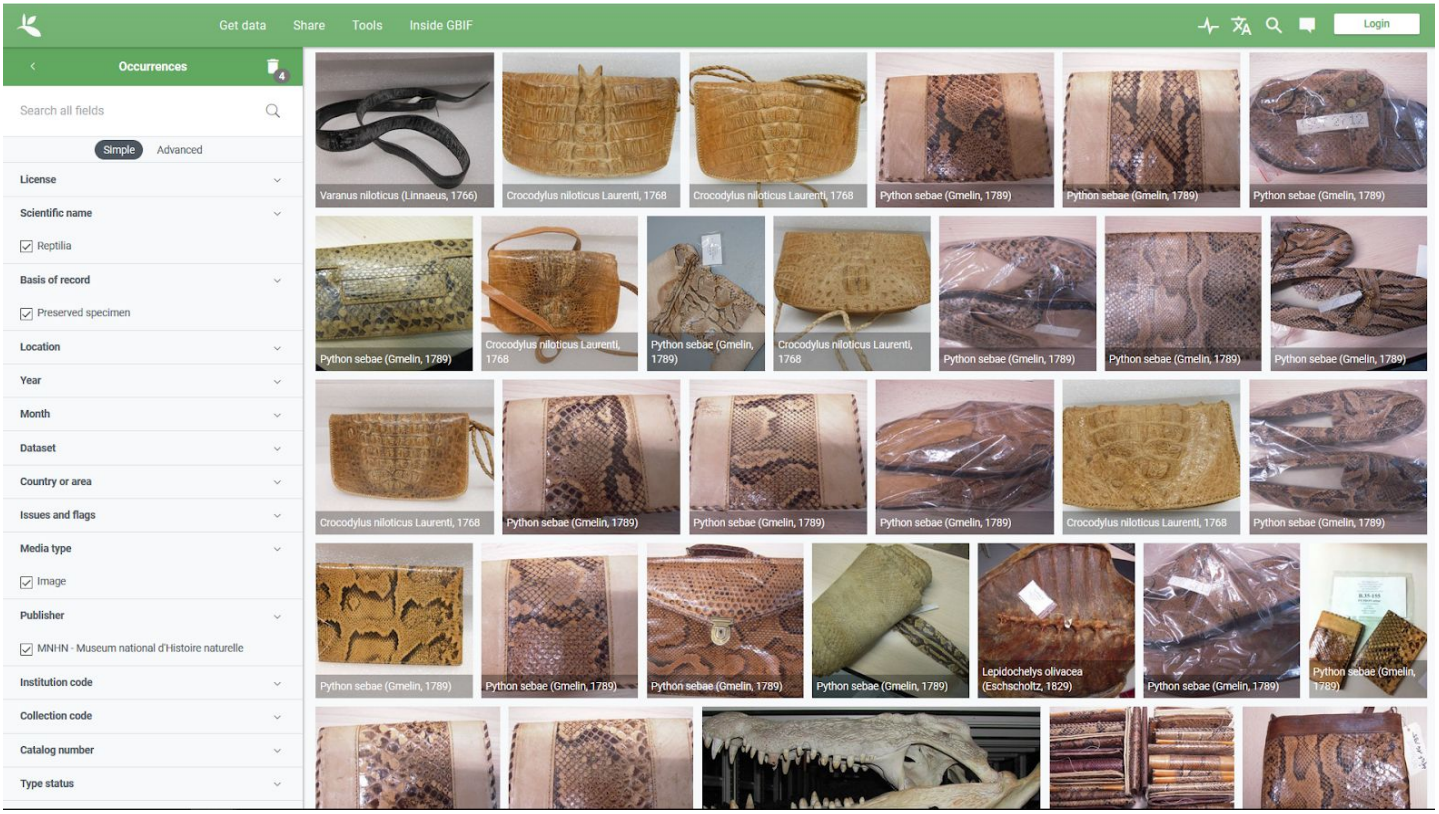
The Muséum national d'histoire naturelle (MNHN) in Paris has more than 465.000 vertebrate specimens on GBIF, including nearly 23.000 occurrences with images. The MNHN began its mass digitisation programs with herbarium sheets. This first homogeneous set of 6.000.000 digitised specimens was produced between 2008 and 2012. It was followed by 900.000 pages and sheets of registers and catalogues digitised between 2014 and 2017.

In 2014, a 5-year national program (e-ReColNat) was launched, still targeting herbaria plus types and published specimens from zoology collections (in volumes). Although priority has been given in recent years to digitisation and computerisation of types, digitisation plans are decided at the level of large collections in order to best meet the management and organisation of physical specimens.

Thus in entomology, for the types, the labels were detached from the specimen and photographed separately. For Hymenoptera in addition to a general photograph of the specimen, detailed photographs illustrate the main characters of the specimen. For Lepidoptera, photographs of the top, bottom and labels are grouped on the same image. More recently for this group, in addition to the types, the boxes are photographed and the species identified within the boxes, but still not the individual specimens beside the types or specific specimens on demand.

In vertebrates, for the naturalised bird collection, the emphasis is on labels, while in mammals only one general photograph of the specimen illustrates it. In reptiles, the processed skins from CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora) catches, which gives very special images, illustrated here.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483



With the end of the e-ReColNat program and its grant, the institution's policy is scaled down again and will focus on the computerisation and digitisation of specimens entering as acquisitions and leaving on loan. In addition to the daily work of the collection management teams, there are various projects to centralise and consolidate data on specimen collections. Thus, for 3D imaging, there is no mass digitisation program for specimens in large quantities. Specimens are digitised by different techniques: laser surface, photogrammetry, PTM, or CT scan. These scans, produced by internal and external research, are now intended to be integrated with the documentary data of objects in collections and to be made available to the community.

Case study: Naturalis[14]

The large scale digitisation effort that Naturalis undertook between 2011 and 2015 was divided into production lines based on handling specifics of the specimens, not based on taxonomic groupings. This means that there was not one production line for vertebrates. The production line dubbed "(e)Vertebrata" digitised the main part of vertebrate specimens in this project. However, it also covered dry invertebrate material, and vertebrate material was digitised in the production lines for specimens preserved in alcohol and for microscope slides. Other production lines were established for digitisation of other dry material (molluscs and geology) which adds information to this case study. The overarching project

[14] Sources used: Naturalis 2013, 2015, 2018. Heerlien et al 2015.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

aimed to digitise objects, link these through the database to physical storage and to integrate two major collections after the merger of the old collections of Rijksmuseum voor Natuurlijke Historie (RMNH, then Naturalis in Leiden) and Zoölogisch Museum Amsterdam (ZMA, Amsterdam). Another goal was to establish a permanent digital infrastructure for the museum, including registration system and workflows.

The (e)Vertebrata production line aimed to digitise 325.000 objects of 13 subcollections, including mammalian taxa, avian taxa (including eggs and nests) and dry invertebrate taxa. At the end of the project the target and budget were not reached (both 80%) only because over 70.000 birds had been included in the target and budget but had been digitised prior to this project. All boxes were also digitised as storage units, each object was tagged with a data matrix label and any multi-object specimens (skin, bones, other material) were linked through virtual umbrella records. Digitisation included data entry, labelling and imaging. Imaging was not done for each object while in some cases multiple images were captured.

The focus of imaging concerned data sources, i.e. if additional data was recorded on the box of the object or the reverse side of a label this was imaged too. The specimen was generally included in the images but of secondary importance. Various preservation types of one individual were linked through a virtual umbrella record, but spotting this often depended on small marks on the labels and the operator's skill and memory. For this reason, one species was entirely handled by a single operator. Eighteen database fields were included in the registration, and 8-9 additional fields only for certain cases such as types.

For imaging both DSLR photography and flatbed scanners were used. In general, skulls were scanned, while skins and mounted specimens were photographed (see table 1). In practice, sometimes the guidelines were deviated from (decided by curator per group, based on e.g. size) to benefit from the greater depth of field and lighting options of photography, while flatbed scanning was generally less time consuming. Mounted specimens were photographed to allow the whole specimen to be imaged, while extra photos were taken to capture label detail as needed. Even blank sides of unmounted labels were imaged, because graphically reproducing the blank side of the label informs the user that there is indeed no data, excluding the possibility that data was mistakenly neglected in the graphic reproduction process. Specimens that were too large for the copy stand were individually photographed by the in-house photographer. Whole skeletons with detached skulls, partial skeletons, or skulls with loose bones were imaged to capture both a palate and lateral view of the skull with both sides of the labels. Postcranial skeletal bones were included where practical. Because skins are of lesser use for research and due to time pressure, it was chosen to image only a single representative of the species (generally the first in the box). For the other skins of that species only the labels were imaged. In practice, for certain groups almost all skins were imaged (e.g. Chiroptera). Lateral views of skulls, balanced on the zygomatic arch, could be difficult to obtain, so the operator could use small

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

glass panes to stabilise and position the object while keeping it as inconspicuous as possible. These glass panes were also used to flatten any curling labels.

Flatbed scanning was done at 600dpi with manual selection of scan area for each scan and saved at the highest quality JPEG available. A scale bar was included in the scans. Photography was done with a Nikon D3000s DLSR at ISO 250, saved in JPEG Fine quality using Optimal Quality compression, and at largest image size (4288 x 2848). Most settings were kept fixed so that the operator only adjusted camera height to achieve the closest crop around the subject, exposure compensation for contrast and brightness, and the focus spot (on the label) for autofocus. The only scale information comes from the 1cm grid on the copy stand surface. Colour calibration was not done and the lighting setup was adjusted by the operator to achieve correct contrast and brightness. While colour correctness is considered important, calibration workflow was unknown or too time consuming. The resulting JPEG file sizes range from 500 KB – 6MB, but average around 2-4 MB each. In practice, one box could contain objects destined for both imaging options.

File naming was done through scanning the new data matrix labels in the image with an external barcode reader to match the registration number with a manually added suffix for the concerning view of the object or whether the label showed data on taxonomic identification or registration number. By reading the data matrix code from the image, this functioned as a check on sharpness and quality of the image.

Table 1 Graphic reproduction options with file suffixes.

	Scan or photo?	**Required Views**	**Filename Suffix**	**Comments**
Unmounted				
Skulls *with loose bones **(partial) skeletons *with **detached** skull*	S *S **S or P	palate lateral	RMNH.MAM.12345.a**_pal** RMNH.MAM.12345.**_lat**	Each view is paired with one side of the object's labels. Include postcranial bones when practical.
Skeletons (whole intact)	P	whole skeleton	RMNH.MAM.12345.a**_gen** RMNH.MAM.12345.**_reg** RMNH.MAM.12345.**_lab**	Depending on the object's size, a second photo may be required to capture label detail or the reverse of the labels.
Skins (study and flat): one representative per species	P	dorsal ventral	RMNH.MAM.12345.b**_dor** RMNH.MAM.12345.**_ven**	Photography only. Pair each view with one side of the object's labels.
Skins (study and flat): labels of remaining skins of same species	S	both sides of object labels	RMNH.MAM.12345.b**_gen** RMNH.MAM.12345.**_reg** RMNH.MAM.12345.**_lab**	If the label has *both* genus and registration number on the same side, choose the _gen suffix. If the label has neither genus or registration number, use the generic suffix ***_lab*** for 'label'.
Mounted				
Skulls, Skeletons, and Skins	P	whole object with labels	RMNH.MAM.12345.a**_gen** RMNH.MAM.12345.b**_reg** RMNH.MAM.12345.**_lab**	Depending on the object's size, a second photo may be required to capture label detail or the second side of the labels. If the label has *both* genus and registration number on the same side, choose the _gen suffix. If the label has neither genus or registration number, use the generic suffix ***_lab*** for 'label'.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The mollusc production line digitised 120% of the target objects (613.000 objects) after spending 100% of the targeted budget. Each object was tagged with a new data matrix label, taxonomic updates were executed and all material collected before 1940 or with handwritten labels was imaged. As a result, approximately 90% of the dry mollusc collection was digitised during this project. The geological production line targeted to digitise 200.000 Cenozoic molluscs, molluscs from a geological sample, palaeontological invertebrates, geological samples from the Netherlands, and meteorites and tektites. Again, every sample was tagged with a data matrix label, each box was also registered and tied to location and samples, imaging was performed for illegible labels and type specimens.

3D digitisation pilot

In 2014, one of the digistreets that was designed was for a pilot of 3D digitisation. Twenty objects were captured through photogrammetry with the help and advice from the ict department, photographers and an external company. Ten models were published on 3d.naturalis.nl, using a threeJS viewer. While a lot was learnt from the design and the process, it was decided not to continue the pilot due to too high investment compared to scientific and collection value due to quality (Van den Oever and Gofferjé 2012, Naturalis 2015, Heerlien et al 2015).

Results from mass digitisation project

During the Naturalis digitisation project the choice was made to keep the workflow simple. Many specimens were imaged. Some collections have basic data for all or a large part of the collection, and parts of it have even richer data available. The data and images are made available through an online portal but also through APIs that can output JSON or XML amongst others and has been uploaded to GBIF. Each curator uses the images at least weekly, if not daily, in their tasks. Quite soon after the project ended, the collection was closed to both researchers and curators due to a new museum and storage spaces being built. Only during a few times each month, the collections can be entered by curators to collect and return specimens. Due to the availability of data and images, the curators have been able to continue with loan requests and data entry. Many loan requests have been handled or made more specific due to data and images being available, especially during the inaccessibility of the collections. Because the data is available and queryable online, this has resulted in extra requests and increased use of the collections.

During planning and during the execution certain choices were made about prioritisation of collections and what data to capture. This also means that at the end of the various production lines, due to time pressure, it was sometimes necessary to record data for fewer fields or to image a narrower selection. The curators are now on the one hand very glad that images are available, but would also prefer to have their whole collection databased. In hindsight, they still can't give a clear answer whether they would have preferred more specimens to be imaged, or to have more specimens registered.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Lessons learned and changes since this project

After the conclusion of the mass digitisation project, Naturalis has continued with digitisation, albeit on a smaller scale. Especially any undigitised loans going out or returning are entered into the database. Imaging is also done for specific objects, in most cases using a DSLR on a copy stand: high profile specimens and labels older than 1940 or containing handwritten text. Through this effort Naturalis aims to fill in the remaining undigitised collections.

Steven van der Mije: "*The division of ICEDIG imaging subtasks illustrates scope issues. The current design (herbarium sheets, pinned insects, vertebrate material & skins, microscope slides and liquid preserved specimens) represents an odd assemblage of specimens, as vertebrate collections consist of a great variety of conservation methods resulting in an eclectic mix of objects. We can find almost every preservation technique in vertebrate collections from slides to spirit collections, flat skins to full mounts, loose bones to connected skeletons. The division of the subtasks in task 3.1 doesn't fully reflect this diversity and on the other hand there is serious overlap between the various subtasks: most vertebrate and insect collections also have substantial liquid collections, slides and other forms of specimen preservation. On the other hand we can find some of the same types of material in invertebrate and geological collections with the same challenges for digitisation.*

From 2011-2015 Naturalis digitised a large part its holdings in a program (FES / FCD) financed by the national government. In this program we set up production lines (digistreet) for registration and digitisation based on the handling specifics of the specimen, not on taxonomic division. So we didn't have a vertebrate digistreet, instead we had one for liquid specimes, for sheets, or for pinned specimens etc. Organising your digitisation effort this way enables you to invest in the most efficient way to maximise the numbers you can digitise.

In this period Naturalis digitised almost 9 million specimens on an object level, for a large part with an image. The remaining part of the collection (estimated > 30 million) was inventoried on a container level. This means we registered every box, every jar and every drawer, and we have a rough idea what species are inside and how many specimens."

Naturalis planned their mass digitisation project to be able to handle similarly preserved parts of the collection in a similar fashion. While real results were achieved, it also resulted in some issues due to the large scale of the project. While preservation type may be very similar, handling requirements or label data specifics differed between subcollections. If done in smaller batches, these differences may have been addressed better. On the other hand, they are aware that this will lead in a more fragmented approach. Another lesson learned is that professional and technical photography support during the design of the imaging workflow is essential for the quality of the imaging. Other improvements can be achieved through greater standardisation of positioning of specimen and labels. A higher level of automation could have been achieved by batch renaming of images and batch entry of repetitive data for a set of specimens or on drawer level.

A number of practicalities were improved after the FES project. E.g. instead of using glass planes to stabilise the specimens, a container of fine sand is used. One of the

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

bottlenecks was the collection registration system (CRS) that was still under development. It was attempted to create one data entry form for all types of collections, which made this step very cumbersome. The system was also occasionally troubled with connectivity issues etc. The curators now suggest that data entry into a custom Excel with subsequent data import into the CRS is a more practical and quicker approach. Another approach to speed up data entry they suggest when the situation is suitable is to use the catalogues for creating records and data entry, and then checking that it matches the presence of specimens in storage. While the focus of imaging during the mass project was on label data, that has now been shifted to equal emphasis on the specimen.

In the mineralogy and petrology collection, the curator has started taking pictures of objects as needed with a smartphone. Care is taken to photograph the object and its label together; in case of samples of large dimensions a close up photo of the label is captured as well. Users requesting objects often don't remember essential details of catalogue number, location, size and descriptive name. By taking photographing objects visitors are interested in, they can be more easily retrieved and its approximate dimensions estimated. At the end of the day, the photos are uploaded to the database and connected to the record. This type of imaging is not mass or performed in a standardised setting, but will in the long term help with the digital accessibility of the collection, especially for objects that are too large or heavy to be ever processed in a mass digitisation project.

Adapting existing imaging solutions to dry three-dimensional objects [#7]

Most dry vertebrate imaging projects have been executed with a camera mounted on a copy stand (Taylor 2005). This equipment is commonly present in collections and is flexible in orientation and subject size (eg. ICEDIG MS44 (2019a)). Very small objects may be imaged with a macro lens or microscope, and larger objects handheld or with a camera tripod. The benefit of a vertically mounted copy stand is the use of a spirit level for perfectly levelling the camera, although some cameras now have a virtual level.

Fine sand, covered with velvet, can be used for balancing irregular shapes for correct positioning (Taylor 2005, p148).

Ariño & Galicia (2005) describe the use of multiple mirrors, movable stages for the specimen and movable arms for the camera to achieve multiple views of the specimen. To achieve the same result, it may be simpler to use multiple cameras, e.g. for a dorsal and lateral view. By controlling the cameras from a computer they can be triggered simultaneously (or in rapid succession).

Collections that are uniformly mounted or laid out, that allow sufficient data capture of specimen morphology or labels could potentially be digitised using a SatScan type scanner (Blagoderov et al 2012). There is automation in place to detect specimen boundaries to generate images per specimen from the stitched whole (Hudson et al 2015). It is however

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

likely that few vertebrate or other dry three-dimensional specimen collections will meet the requirements to set up such a workflow.

Adapting the herbarium conveyor belt solutions

Picturae's conveyor belt set up for the project at Naturalis. The black structure contains the camera and lights, showing the distance of camera to subject. Greater distance can only be achieved in spaces that have high ceilings. (Image copyright Picturae)

Both Digitarium (since 2017 Bioshare Digitisation, part of Sertifer Consulting Ltd) and Picturae have developed conveyor belt solutions for mass herbarium sheet digitisation. With some modifications it can be adapted for certain types of objects. Limiting factors are weight and size of the object, but mostly the depth of field that can be achieved. The current Picturae setup (full frame camera, 90/100/120mm lens, f/11, distance usually at 1.5m) can achieve a maximum depth of field of 9 to 15cm. This would be sufficient for fossil slabs and small objects, including many vertebrate and invertebrate specimens and even certain drawers of these objects.

Depth of field can be even further extended by increasing the distance (up to 2m). Depth of field and pixel per inch (PPI) need to be balanced for the collection in question. Aperture can also be adjusted to increase depth of field, but at a certain point the diffraction from the smaller aperture is greater than the gain from the depth of field. At which aperture this happens varies per lens, but often this is f/16. Decreasing the aperture requires more light to achieve the same exposure, either by brighter lights or longer shutter speeds. The current herbarium conveyor belt can be fitted with improved lights which can compensate up to f/22 and even f/32.

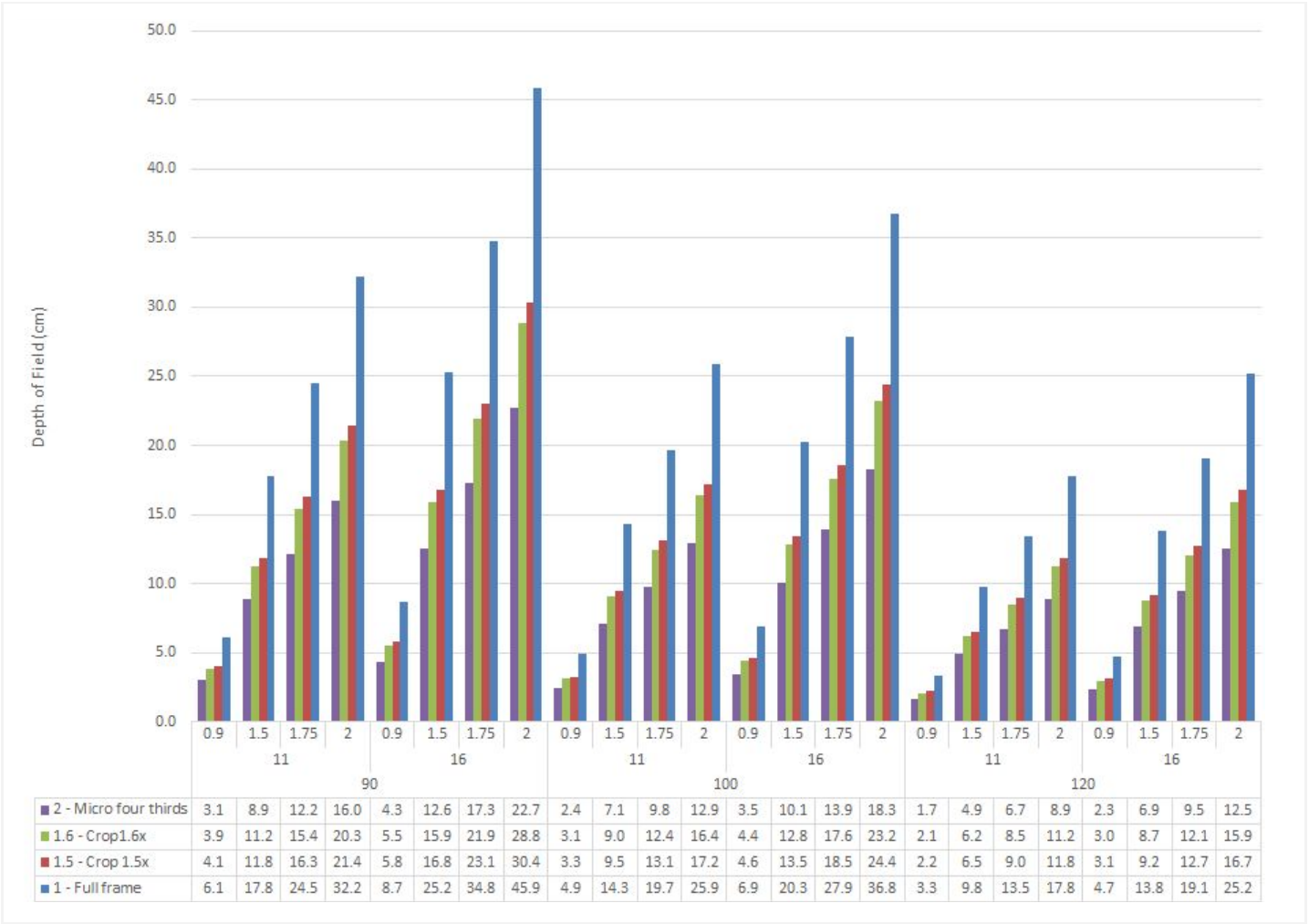
Adjusting both distance between camera and subject to 2m and aperture to f/16, the full frame setup can achieve a depth of field of 45cm, which is sufficient for a large part of the

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

type of three-dimensional objects that could fit on the conveyor belt system in terms of length, width and weight. When only the aperture is adjusted to f/16, a depth of field of 25cm can be achieved with a full frame camera and a 90mm lens, see also chart below.

Note that when distance to subject, aperture and lens focal length remain fixed; depth of field is deeper for larger sensor cameras. However, to obtain the same coverage of the subject on the photo it is needed to get closer to the subject with a larger sensor compared to a smaller sensor, or a longer focal length lens is used. This results in a smaller depth of field on a large sensor compared with a smaller sensor. This can be compensated by using a high resolution full frame camera (such as the PhaseOne cameras (80, 100 or 150 megapixel) often used by Picturae), so that the photos can be cropped to a closer framing while maintaining sufficient pixel coverage on the subject. Another consideration is that by narrower apertures come with increased diffraction, meaning that the sharpness at the edges of the photo decreases.

The herbarium conveyor belt solutions are further described in the report on herbarium sheet imaging (ICEDIG D3.6 (2019b)).



	90								100								120							
	11				16				11				16				11				16			
	0.9	1.5	1.75	2	0.9	1.5	1.75	2	0.9	1.5	1.75	2	0.9	1.5	1.75	2	0.9	1.5	1.75	2	0.9	1.5	1.75	2
2 - Micro four thirds	3.1	8.9	12.2	16.0	4.3	12.6	17.3	22.7	2.4	7.1	9.8	12.9	3.5	10.1	13.9	18.3	1.7	4.9	6.7	8.9	2.3	6.9	9.5	12.5
1.6 - Crop1.6x	3.9	11.2	15.4	20.3	5.5	15.9	21.9	28.8	3.1	9.0	12.4	16.4	4.4	12.8	17.6	23.2	2.1	6.2	8.5	11.2	3.0	8.7	12.1	15.9
1.5 - Crop 1.5x	4.1	11.8	16.3	21.4	5.8	16.8	23.1	30.4	3.3	9.5	13.1	17.2	4.6	13.5	18.5	24.4	2.2	6.5	9.0	11.8	3.1	9.2	12.7	16.7
1 - Full frame	6.1	17.8	24.5	32.2	8.7	25.2	34.8	45.9	4.9	14.3	19.7	25.9	6.9	20.3	27.9	36.8	3.3	9.8	13.5	17.8	4.7	13.8	19.1	25.2

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Depth of field (cm) for various combinations of sensor size, lens focal length (90, 100 and 120mm) and distance to subject (0.9, 1.25, 1.5, 1.75 and 2m)[15].

Adapting Picturae's sandwich set

One solution that was designed by Picturae for double sided 2D material is called the sandwich set, aimed at paper archives, photos and prints (unbound). This set may be adapted for use in natural history collections. The set consists of two horizontal glass panes, one fixed inside the frame and on which the material is placed. The second glass pane is attached by a hinge at the far end and is brought down by the operator with a handle (to prevent smudging the glass) to flatten the material and to set off a magnetic trigger connected to both cameras. A camera is mounted facing down at some distance above, and another is mounted facing up at some distance below. Due to the required distances the operator stands at this set. Lights are positioned to illuminate both faces of the material. The top, sides and back are surrounded by a closed black frame. The whole set can photograph according to the strictest Metamorfoze and FADGI standards (see Nieva de la Hidalga et al 2019a).

As is, the set can be used very well for labels and loose notes, when only photographs of those are needed. This can be especially interesting for labels that are curled or have dog ears, or have writing on both sides.

During their mass digitisation project, Naturalis was especially interested in the labels and for some collections always included a photograph of the reverse side. Even if there was no writing present, the photograph still serves to prove that no data was missed. Naturalis also used, like other institutes, a loose glass pane to flatten overly curved labels. This takes a disproportionate amount of time, especially considering fingerprints and smudges.

With some redesign this double photography set can be adapted to three-dimensional specimens. This can be interesting when two sides of an object are to be imaged. The time needed for two photographs would be the same as for one and the main time savings would be in not having to reposition the object, which is also safer for the object. The distance from camera to subject is 90 to 125cm; for the relevant depth of field see the chart above. The maximum depth of field with a full frame camera is achieved with F/16, 90mm lens, and 1.25m distance is 17cm, for 0.9m distance it is 8.6cm.

Of course, the top glass pane can't be used with three-dimensional objects. A two part design with one side for the object without top pane and one side for the labels with top pane is an option too. Another consideration is the stable and proper positioning of the specimen, as a sandbox or anything else opaque can't be used. Stabilisation between glass panes could be an option. Naturalis used a single glass pane and found that this was cumbersome, but two may provide more stability and speed.

[15] Formulas used for calculation are from: http://paulorenato.com/index.php/photography/151-excel-depth-of-field-calculator (accessed 23-05-2019)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Digitisation of acetate peels and other transparencies

One of the various types of material found in paleontology collections are acetate peels, which are an alternative to thin sections[16]. The peel is flexible and partially translucent, which can vary in size up to 25cm or more. Traditionally, they are studied under a microscope. During a quick test for the Field Museum of 50 acetate peels, Picturae found that these can be imaged in extreme resolution through use of their setups for scanning photo negatives and other transparencies, and image stitching. The transparency imaging set uses a glass plate to support the material, is lighted with strong high quality strobes from underneath while the camera is mounted above the subject. This set is designed for digitisation conform Metamorfoze and FADGI standards. Using a specially designed shutter remote, high throughput can be achieved. A PhaseOne 100 megapixel camera with Rodenstock 105mm objective was used, achieving a maximum magnification of two to one and resolution of 10000ppi. As this was a quick test, nothing can be said about throughput rates, especially because stitching at this point was done manually. In a real project settings an automated workflow for stitching would be developed.

Digitisation of written sources like catalogues and field notebooks

Besides imaging of specimens with their labels, imaging of catalogues, field notebooks, loose labels and other flat data sources can be very useful. This type of imaging is similar to the extensive imaging done for archives, conform Metamorfoze and FADGI standards. This means that there are many solutions available for this type of material.

[16] https://strata.uga.edu/cincy/fauna/bryozoanStudy/acetatePeels.html

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Recommendations [#9]

According to the experience in most projects it is easier to find funding for equipment than for hiring workers, because it is seen as an efficient long-term investment (Australian Museum 2011, p24). This is one of the reasons that automation is desirable.

Vollmar et al 2010 (p97) note that funding and availability of (especially, well-trained) staff are closely tied as impeding factors, and also acknowledges that funding limits technology options which can impede the efficiency of digitisation projects (p99, 101). They do not note the reverse: that technological advances can be used to bring cost down and to alleviate pressure on staff. Technological advances can increase cost-efficiency and simplify the workflow. This report focuses on the ways technology may be used to work for digitisation goals, instead of being an impediment through lack of funding.

They also note that cost and pressure on collection managers can be reduced by efficient use of non-experts for the first level of data entry. Respondents to their survey also noted that cultural changes at collection and management level are needed to improve their estimation of value of digitisation. Now, roughly a decade later, we should see some changes in this perceived value due to staff turnover and experience with digital collections. This should also affect the options for applying for funding of digitisation projects. The knowledge and skill required to perceive the possibilities of digital data and digitising methods will by now have progressed from optional training need to basic job requirement, at least in the larger institutes.

Criteria [#9.1]

Knowing the above mentioned issues with digitisation sharp criteria to determine what to do and what not are necessary. Every aspect of digitisation should be considered: imaging, full data capture, staff availability and expertise.

Reasons to image every specimen and every label[17]

Process and curatorial benefits:

- Data entry from image, especially when data entry during the initial project is minimal. By extension: OCR (Optical Character Recognition) and crowdsourcing.
- Data validation. Kalms (2012, p27) puts it like this: “Poor data quality is often used as reason to avoid digitisation; it should not be. While digitisation will highlight data quality issues, it also provides an opportunity to remedy these issues. [...] Use digitisation strategically to overcome data quality issues, eg if determinations are unclear, image specimens, publish the images and encourage external experts to comment on the nominated species name.” The GBIF task force on accelerating

[17] Sources: Australian Museum Digitisation Project: final report (2011, p3).

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

discovery also underwrites the use of digitisation as the most efficient method of identifying and overcoming data errors (Krishtalka et al 2016, p25).

- Limit handling. Every handling event comes with an error risk. Especially in combination with data entry by volunteers or temporary employees. Data entry from image limits the handling to a minimum during digitisation and afterwards, as observed by Nelson et al. (2012, p30).
- When repacking the collection, there may be loss of information which is not perceived until much afterwards which may be forestalled by imaging.

Improved and sustainable use of information:

- Collection mystery puzzles may be solved through pictures of specimens and labels.
- Increase of use and findability of collections.
- Recording of current conservation status, for monitoring collection condition.
- Virtually reconstruct historical collections, such as those from specific collectors or from areas with no historical natural history tradition and former colonies (virtual repatriation).
- Making the collection not on display accessible to the public. In many cases, these collections are paid for and maintained with public money.
- Improved collection audit and security (Blagoderov et al 2012).
- Public engagement and enhance public awareness of value of national/regional collections (Blagoderov et al 2012).

Availability for potential future developments and use:

- Potential for species identification from image, though limited.
- Potential for morphometric analysis and phenological population studies.
- Virtual specimen in case of loss or damage and during loans.
- Future technology advances. AI (artificial intelligence), OCR and HTR (Handwritten Text Recognition) are techniques that are already in use and are very promising for future development. AI for species recognition.
- Providing training datasets for HTR and OCR.
- During a mass digitisation project a major part of the collection is processed. The next opportunity to do so may present itself in another 150 years.

Even if data entry is done directly from the object, imaging the labels provides extra information. Often there is more information on the label than will be transcribed. Having the image allows this information to be accessible, even if it is not searchable in the database. Images of the labels allow a check of the data without the effort of locating the specimen in storage, both by curators and (external) researchers. When needed, the label images can be used for inhouse data entry of the extra data (at a later date when funding is available), crowdsourced efforts for transcription or OCR and HTR.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

When imaging is done for the labels it may be efficient to also include the specimen. The hardware and capture time are usually higher, but may pay off in the future. Experience from institutes such as Naturalis which have done imaging for certain parts of the collection learns that loan requests can be more specific after browsing the available specimens by image, or can even be replaced completely through the digital image. It also enables basic error detection, as well as more in-depth research uses. However, due to speed considerations, imaging as part of a digitisation project is generally not meant to facilitate research specific questions.

Specifications and technical solutions [#9.2 and #9.3]

Technical considerations are described in various sources, e.g. Ariño & Galicia (2005) and see ICEDIG deliverable D3.1 on quality control (Nieva de la Hidalga et al 2019a). One aspect which is especially important for the three-dimensional objects treated in this report is depth of field.

Of all the preservation types in collections considered in ICEDIG, this report pertains to the most complex objects to position. Microscope slides and herbarium sheets are flat, liquid preserved specimens can remain in their flat bottomed containers (depending on requirements of the image), and pinned insects will most likely be imaged on their pins. Smaller vertebrate and other dry three-dimensional specimens can be imaged in their containers (such as small bones or shells in open boxes), if there is no need for specific views of the specimen and an overview image is sufficient. However, if there is a need for specific views then positioning skulls, skins, eggs, shells in a specific manner and a stable position will slow down the imaging speed. This is also a factor in amenability for automation.

Various workflow descriptions of digitisation projects include manual adjustments of the camera height on the copy stand and/or adjustments of a zoom lens to frame the object closely (e.g. Australian Museum 2011, p9 and Naturalis 2013). While it is likely that the size of objects doesn't fluctuate between extremes continuously due to size continuity of digitising per taxonomic group, this is still a slowing factor. For example, the framing footprint of going from a sample of one egg to five eggs may need adjustment, as well as going from one skull to a whole skeleton. An automated solution to make the height and zoom adjustments would eliminate this manual step. This can be envisioned to automatically detect the footprint of the sample in the imaging area.

Another solution to the height and zoom adjustments for maximum framing lies in dedicated framing footprint set ups for the different sizes and shapes of objects. For a large digitisation project, it may be efficient to set up multiple digitisation stations based on sample dimensions. A design of specific frame footprint sizes can be made, for example matching A6, A4 and A2 paper sizes. By including an easy switch from front to downfacing, or multiple cameras for multiple views, these sets can accommodate a range of objects. This would cover smaller herbarium projects, liquid containers, dry three-dimensional material,

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

and more. The uptake of a new batch would then be determined amongst others by the required footprint and available stations.

For example, in the Mollusc digistreet at Naturalis a camera with macro lens was used on a height adjustable copy stand. Less than 1 percent of specimens was too large for this setup while over 40% required an even smaller footprint to be adequately imaged. About 60% could be imaged well with this setup, but height and zoom adjustments were occasionally required.

Photography with multiple cameras may be interesting for specific types of material. Experience from the Geology digistreet has taught that for minerals it can be very difficult for non-experts to position the specimen in such a way that the relevant side is photographed, e.g. when there is a small mineral on a larger one. If not all photos are relevant and storage is an issue, then the final selection can be made by experts or curator.

Workflows [#9.4]

Blagoderov et al (2012 p. 134-135) define a number of criteria for digitisation on industrial scale:

- Automation where possible (only excluding physical handling of specimens).
- Focus on total digitisation of a collection ("wall to wall"), as deciding which specimens (or drawers) to digitise and collecting them is inefficient and results in a fragmentation of digitisation status.
- Divide the labour intensive steps into smaller discrete steps (modularisation). This allows more efficient use of staff of various skill levels.
- Standardise collection of metadata and reduce to essential minimum.
- The last point is minimised by Blagoderov et al to image + unique identifier + storage location. It may be appropriate to adhere to the lower MIDS levels (minimum information for digital specimens, Saarenmaa et al 2019). The important point is that not all data needs to be recorded at once, especially when an image of the labels is available so that the specimen doesn't need to be accessed again for more extensive data entry (Beaman & Cellinese 2012).

For an example of a workflow in practice see also Case study: Naturalis [#2a].

Outsourcing or in-house, on-site or off-site [#9.5]

ICEDIG MS44 (2019) asked the participating collection holding institutes about out-sourced and off-site digitisation. Almost exclusively, this was done only for herbarium collections. When out-sourcing was done, it was also usually in-house meaning that institutes usually have sufficient space for their digitisation programs and prefer close contact with the external party.

Many digitisation projects work with temporary employees. Experience has taught that temporary employees with higher education levels may be available, such as new graduates,

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

but that they may get disinterested with the work more quickly than others and that the employee turnover can be undesirably high. Disinterest combined with low prospects at the institutes may also cause higher error rates. Especially for data entry tasks, it is possibly more efficient to attract those with good text processing skills. A few sufficiently experienced team leads are essential, but often are occupied mostly with project management tasks and should not be counted on to contribute to production numbers. When all non-team leads are hired through employment agencies then the circumstances are comparable to outsourcing. One benefit of using these types of digitisation workers, as pointed out by someone involved with the mass digitisation project at Naturalis, is that these people start without any prior preconceptions so that they are not impeded by apprehension of what might go wrong. To keep the error rate as low as possible it is important to let essential tasks, such as updating taxonomy, be executed by appropriately skilled workers. Another benefit is that outsourced workers are dedicated to the project, while internal staff may still have other tasks that distract them from the digitisation work and volunteers can be less reliable. While the greater knowledge and experience of internal staff may be a benefit, they can potentially get distracted from digitisation by non-essential activities such as redetermination and maintenance.

Regarding data entry by non-experts, Vollmar et al (2010, p. 99) note that tagging uncertainties for later review by experts allows for speed and data quality. Verbatim data entry goes even further, because almost any data is open to interpretation: e.g. synonyms, multiple identifications, date format, historical place names and changes in borders. Verbatim data entry reduces the difficulties to unclear handwriting and print, and mapping of data to database fields. Subsequent changes in e.g. date format can be done in bulk by an expert as a separate step. These choices enable outsourcing and off-site digitisation to a greater degree than when training and expert supervision are required for interpretation during data entry.

Error rate is an important consideration for outsourcing or digitisation by temporary non-expert staff, which can be used to judge quality per batch or per worker with various consequences. First it is necessary to define errors. When these workers are judged by unexpected situations that are not properly described in the workflow, then from their perspective it can't be considered as an error. Errors that they are actually judged on should be those when the workflow is not followed correctly. During their first large scale digitisation project, Meise Botanic Garden (APM) reports, the workflow was initially updated to deal with an increasing number of exceptions. This meant that the workflow was increasingly more complicated for the digitisers to follow, which resulted in higher number of perceived errors. These errors however shouldn't be regarded as errors, but seen as omissions in the description of the workflow. Realising this, the decision was made to keep the workflow lean with a minimal amount of rules and exceptions, where it was possible to perform quality control on the data on larger scale (e.g. through querying the data). This is something to consider during the design of new digitisation workflows: keep it as lean as possible and have an expert do quality control and correct exceptions.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

A consideration for off-site digitisation are regulations on dangerous substances (e.g. alcohol) and potential ecological vectors: not all material can be transported. On the other hand, contamination with pests and damage from environment changes pose risks to the collections. Freezing and quarantine add to the costs.

Health and Safety [#9.6]

Due to the historical depth of the collections at various institutes and the various preservation techniques applied, there is not a single definitive guideline on health and safety for all natural history collections. Potential chemicals include asbestos, arsenic, flammable liquids, formaldehyde, naphthalene. Institutes should be aware of which hazards are (potentially) present in their collections and how to deal with those. During mass digitisation projects, these should be adhered to. Because many digitisation projects involve volunteers or temporary employees, the health and safety guidelines should be clear and enforced during their onboarding. Continuous vigilance is required for any health and safety factors during the project as previously unseen parts of the collection might produce new hazards.

This also includes working conditions complying to health and safety regulations, because operators often sit at desks and computer screens for a significant portion of the time, combined with potentially suboptimal posture to accommodate handling of specimens.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Conclusion and discussion

As demonstrated above for the type of collections that are discussed in this report mass digitisation is not very feasible. This does not mean that digitisation of large numbers of objects is impossible. This can be approached from either the process or the demand. Both are discussed here. Depending on the situation, one section may be more relevant than the other. At the very least, these discussions will help think about how an institute's imaging programs can be designed.

Process based approach for fast but basic digitisation

Extending from the recommendations, we recommended an approach to image every label combined with quick minimal data entry from the images, considering the expertise needed to interpret label data, risks associated with multiple handling moments, the limited range of most data capture and the extended potential of images. We acknowledge that there may be cases where imaging is so complex or where data entry can be done so fast that this approach is not efficient.

The GBIF task force on accelerating discovery advised a tiered strategy with first rapid and least expensive steps, with a second phase for more detailed data capture and imaging (Krishtalka et al 2016, p5). Here we propose that, instead, imaging is the fast initial step that requires minimal expertise. This will capture all label data, so that during a second phase the images can be used to extract the data. The final report of the rapid digitisation pilot project funded by the Atlas of Living Australia at the Australian Museum and the South Australian Museum described imaging as the new databasing (Australian Museum 2011). The advantages of this approach are:

- **Data entry from image as separate phase:** This approach allows extended data entry to be done as a separate step, depending on availability of funds, even a long time in the future. Through minimal data entry and the images most information is available. Extended data entry will still be needed to achieve sufficient accessibility of data.
- **Risks of handling:** Each handling action poses risk to the specimen. When both imaging and data entry from the object are performed as part of a digitisation project this doubles the amount of handling. Even more when the specimens have to be returned to (temporary) storage in between. After digitisation, images provide additional information that (especially partial) data entry can't provide, so that many questions can be answered by the images without having to retrieve and handle the specimen.
- **Availability of expertise:** Workflows for minimum data entry and imaging can be kept simpler than workflows that encompass extended data entry, This means that the initial phase can be done by workers with minimal expertise of the collection. Extended data entry can then be done by the smaller pool of workers with the

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

required expertise, e.g. in reading handwriting, updating synonyms and georeferencing.

- **Incomplete data entry:** Many digitisation programs do not aim all data available for all specimens. Without full data entry, it may still be necessary to set up another digitisation project to capture the remaining data.
- **Extended potential:** The images have a greater potential.
    - **HTR & OCR:** While at least the label data is accessible in the image, in the mid-term future it may well be possible to use OCR and HTR to access the label data. At the very least the images can be used to train OCR and HTR, which is necessarily the first step.
    - **Curation tool:** By including the specimen in the images curatorial goals can be achieved, such as winnowing of loan request and snapshot of current condition.
    - **Research:** Also, a range of research purposes may be served. If not identification, then determining bad identifications so that they specimen can be selected for further study.
    - **Future developments:** Future developments and tools such as AI for species recognition may present themselves.

The goal of ICEDIG and DiSSCo is to mobilise biodiversity information, which requires digitisation of enormous amounts of specimens. With a process based approach, this can best be done with initial imaging of specimens and subsequent partial data entry, and planning for future use with the potential of the images.

After visits to various collection holding institutes to observe digitisation workflow, Nelson et al. (2012) describe three digitisation patterns which fit different types of process based approaches to digitisation:

> "The *data to occasional or optional image to distribution* pattern fits those institutions in which few or no specimens are imaged. Data capture follows curation and may include decisions about which specimens to submit for imaging. Rarely, imaging of exemplars is simultaneous with data entry of those exemplars.
>
> The *parallel data/image to distribution* pattern includes both data and image capture but treats them as independent and simultaneous rather than as sequential steps. This pattern is likely the most labor intensive of the three, especially when it requires specimen handling at two stages of the workflow, with attendant need for multiple trips to storage locations and increased opportunities for specimen damage. This pattern is made more efficient when data capture proceeds from bulk data sources (ledgers, cards), which requires specimen handling only during image acquisition.
>
> The *image to data to distribution* pattern fits institutions that image all specimens (e.g. most herbaria) and captures data from these images. It

> reduces specimen handling and with it the likelihood of specimen damage, increases efficiency by eliminating the need for return trips to storage locations, and offers the capacity to incorporate Optical Character Recognition and similar technologies within the data capture workflow."
> (From Nelson et al. 2012, p39.)

In short, they also found that data entry from the images was the most efficient way of (large scale) digitisation.

Demand based approach for mass imaging

Digitisation always should be considering the end-user of the information produced. However at the time of the project it is not always clear what this stakeholder needs. We therefore have developed decision trees to help with the design of the practical side of digitisation. They are not meant as a set of hard and fast rules and should be useable for various types of collections, not only dry vertebrate collections. It is not meant to assist with deciding when, what and why to digitise; it does address the question of when to image, what to image and how this relates to data entry.

First the decisions that can be made with this tool are described, a number of example projects are run through the trees to explain its use in appendix 1. A major consideration is when to apply a workflow for mass imaging, which is described as well.

These decision trees assist with the following decisions:

1. Will there be imaging?
2. Will data entry be done from the image or directly from the object?
3. Will the specimen be included in the image, or will imaging only be done of labels and textual information sources?

Defining mass digitisation and homogeneity as requirements for mass digitisation

A few decisions need to be made beforehand. The first is a consideration if the project and material at hand can be covered under a mass digitisation workflow. Based on the quantity of objects to be digitised and the hours required, as well as available staff and equipment a digitisation request it may not yield an efficiency gain when executed as a mass project. For example, many requests for research data are too small to warrant a massive digitisation effort but can be adequately handled following the workflow for incidental digitisation.

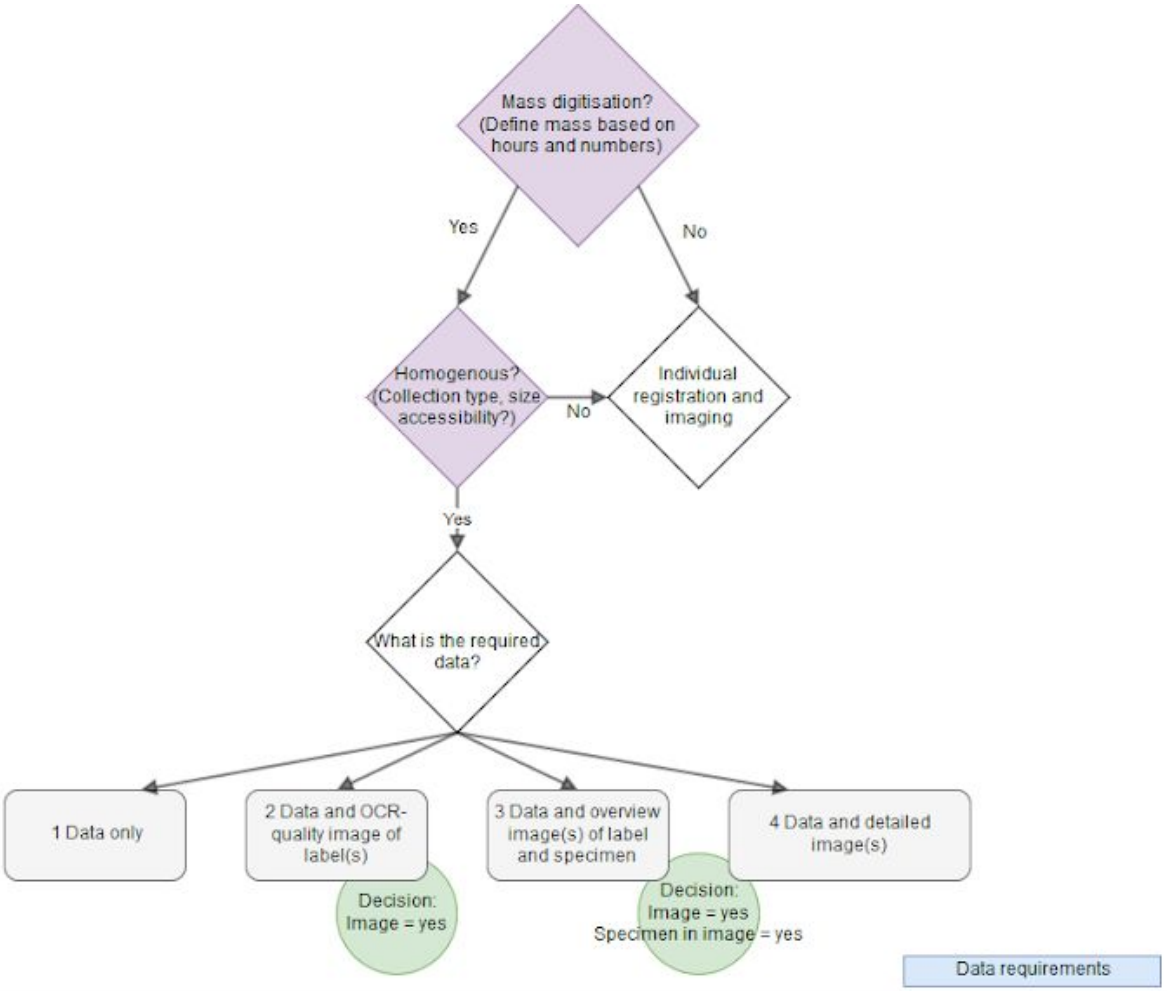
Without attempting to define an exhaustive definition, "mass" could have the following considerations:

- **Quantity**: the number of objects to be digitised needs to be large enough to justify a special approach. This number also needs to be part of the goal of the project. Ideally this quantity is not only expressed in number of objects but also in throughput time. Some processes require for relatively low numbers a significant

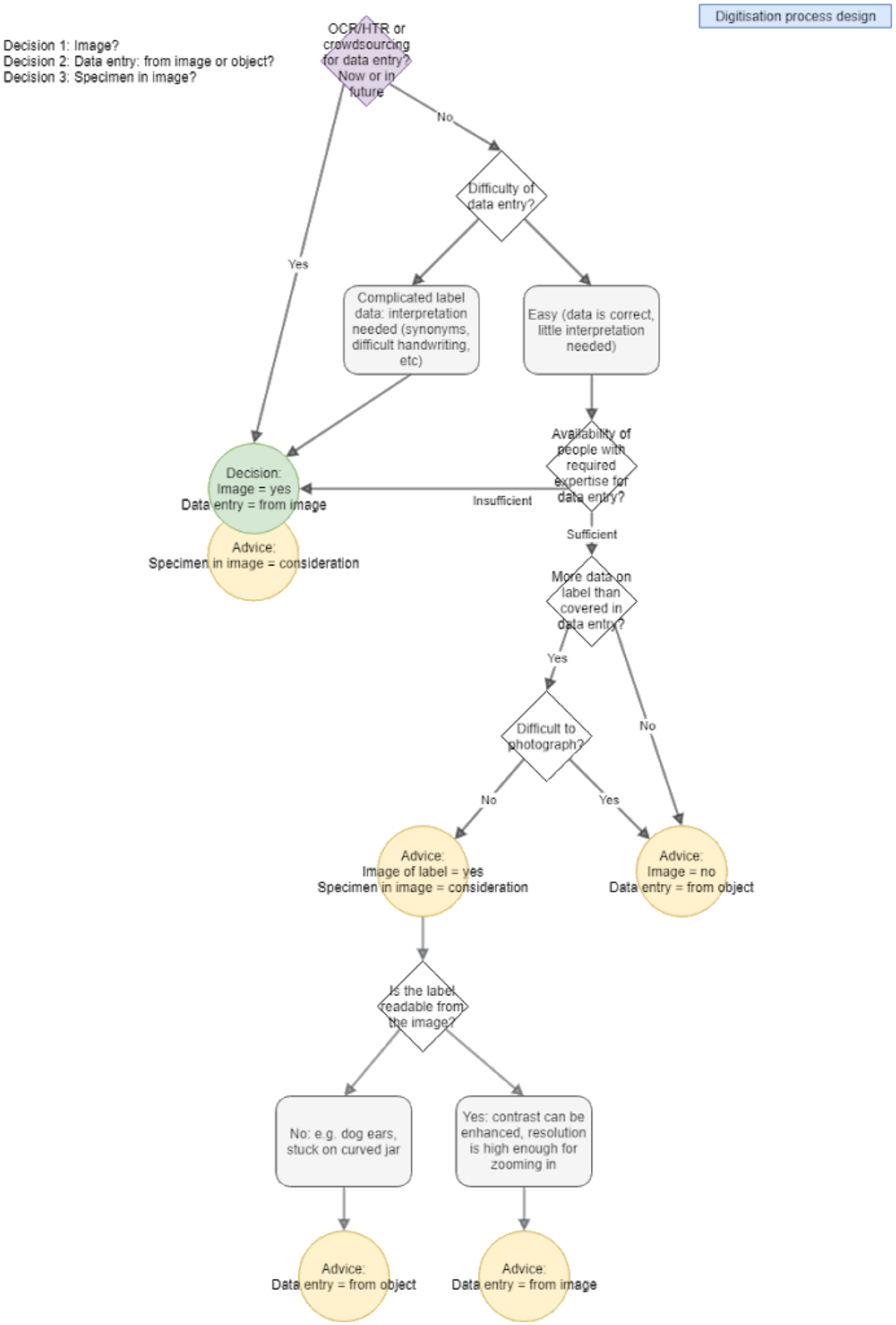
Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

amount of effort, while other processes make it possible to achieve an efficiency gain when treating larger numbers.

- **Goals**: the required output of digitisation needs to make a process-based approach possible, as well as produce usable results.
- **Process**: it needs to be possible to design a process that makes standardised and large-scale treatment possible. The number of exceptions needs to be limited. This also includes safety: the process needs to be safe for both staff and specimens.
- **Approach**: without a project-based approach mass-digitisation can't be done. Project structure requires a plan with goals, description of processes, risks and responsibilities and sufficient capacity. It requires dedicated staff hours that is sufficient to achieving the goals, organisation of facilities and work processes which have been tested through pilots. There are control moments which result in intervention when goals are not being achieved.
- **Homogeneity**: the material needs to permit digitisation in consistent ways with only limited exceptions. For this similar sizes, accessibility and handling are crucial, as well as reasonable consistency of label data for data entry.
- **Capacity**: the required expertise needs to match the availability of expertise, some scaling up or down needs to be possible, required training needs to be limited.



Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483



Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

Final thoughts

No groundbreaking automation solutions were discovered or were found to be under development in digitisation programs during the work for this report. In conclusion, no miracles can be expected for these types of collections. Digitisation of three-dimensional dry objects needs to be done, so the best that can be achieved is to gain efficiency through large scale projects or long term programs.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Glossary

- AI (artificial intelligence): area of computer science involved with development of machines that mimic "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".
- OCR (optical character recognition): mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, or other.
- HTR (handwritten text recognition: also HWR (handwriting recognition). Ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition) or intelligent word recognition.
- Catalogued = records exist on in-house collection management system
- Digitised = specimen information in collection management system with partly or fully transcribed labels
- Fully digitised = specimen information in collection management system, with fully transcribed labels and images)
- Depth of Field (DoF): distance between the nearest and the furthest objects that are in acceptably sharp focus in an image. The depth of field is determined by focal length, distance to subject, the acceptable circle of confusion size, and aperture.
- Focus stacking: digital image processing technique which combines multiple images taken at different focus distances to give a resulting image with a greater depth of field (DOF) than any of the individual source images. Can be achieved by changing focus or by changing distance between subject and camera.
- Pixels per inch (PPI): measurement of the pixel density (resolution) of an electronic image device, such as a computer monitor or television display, or image digitising device such as a camera or image scanner. Pixels per inch can also describe the resolution, in pixels, of an image file.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

References

Ariño, A. H., & Galicia, D. (2005). Taxonomic-grade images. *Digital imaging of biological type specimens: A manual of best practice*, *41*, 55. In: Häuser et al. (eds.): Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Results from a study of the European Network for Biodiversity Information: 41-55. Stuttgart.

Australian Museum (2011). Rapid digitisation project: final report. https://www.ala.org.au/wp-content/uploads/2011/10/Australian-Museum-digitisation-project-final-report.pdf Accessed 12-03-2019.

Beaman, R. S., & Cellinese, N. (2012). Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, (209).

Berents, P., Hamer, M., & Chavan, V. (2010). Towards demand driven publishing: approaches to the prioritization of digitization of natural history collections data. *Biodiversity Informatics*, *7*(2).

Blagoderov, V., Kitching, I. J., Livermore, L., Simonsen, T. J., & Smith, V. S. (2012). No specimen left behind: industrial scale digitization of natural history collections. *ZooKeys*, (209), 133.

Cook, J. A., & Light, J. E. (in press). The emerging role of mammal collections in 21st century mammalogy. Journal of Mammology. https://doi.org/10.1093/jmammal/gyy148

Heerlien, M., Van Leusen, J., Schnoerr, S., de Jong-Kole, S., Raes, N., & Van Hulsen, K. (2015). The natural history production line: an industrial approach to the digitization of scientific collections. Journal on Computing and Cultural Heritage (JOCCH), 8(1).

Hudson, L. N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B. W., ... & Smith, V. S. (2015). Inselect: automating the digitization of natural history collections. *PLoS one*, *10*(11), e0143402.

ICEDIG (2019a). Milestone 44. Technical capacities of digitisation centres within ICEDIG participating institutions. (Part of work for ICEDIG deliverable 7.1 which will be available at https://icedig.eu/content/deliverables, Q3 2019)

ICEDIG (2019b). ICEDIG deliverable 3.6. Best practice guidelines for bulk imaging of herbarium specimens. (Will be available at https://icedig.eu/content/deliverables Q3 2019)

Kalms, B. (2012). Digitisation: A strategic approach for natural history collections. Atlas of Living Australia, CSIRO Ecosystem Sciences, Australia. 95 pp.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Krishtalka, L., Dalcin, E., Ellis, S., Ganglo, J. C., Hosoya, T., Nakae, M., ... & Thiers, B. (2016). Accelerating the discovery of biocollections data. *GBIF Secretariat, Copenhagen, Denmark*.

Naturalis Biodiversity Center (2013). (e)Vertebraten handleiding E1.0.

Naturalis Biodiversity Center (2015). Uit het depot, op het web: twee eeuwen nationaal natuurhistorisch erfgoed in het digitale domein. (https://www.repository.naturalis.nl/record/588732)

Naturalis Biodiversity Center (year 2018). FES NCB Naturalis Zelfevaluatie. (https://www.naturalis.nl/system/files/inline/FES_NCB_Naturalis_Zelfevaluatie.pdf)

Nelson, G., Paul, D., Riccardi, G., & Mast, A. R. (2012). Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys*, (209).

Nieva de la Hidalga, A., Rosin, P. et al (2019a). ICEDIG deliverable 3.1. Quality control methodology for digitization operations. (Will be available at https://icedig.eu/content/deliverables Q4 2019)

Nieva de la Hidalga, A., Rosin, P., Sun, X., Van Walsum, M. and Wu, Z. (2019b). ICEDIG deliverable 3.7. Rapid 3D capture methods in biological collections and related fields. (Will be available at https://icedig.eu/content/deliverables Q4 2019)

Saarenmaa, H., Agosti, D., Dillen, M., Egloff, W., Gagnier, P.-Y., Groom, Q., Hardisty, A. and Raes, N. (2019). Milestone 35. Report on implementation of open access policies in collection institutions. (Part of work for ICEDIG deliverable 6.5, https://icedig.eu/content/deliverables Q4 2019.)

Synthesys3 (2017). D4.5 - Digitisation on demand - a report on feasibility of a digitisation on demand service for natural history collections.

Taylor, H. (2005). A photographer's viewpoint. In: Häuser et al. (eds.): Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Results from a study of the European Network for Biodiversity Information: 41-55. Stuttgart.

Van den Oever, J. P., & Gofferje, M. (2012). 'From pilot to production': large scale digitisation project at naturalis biodiversity center. *ZooKeys*, (209).

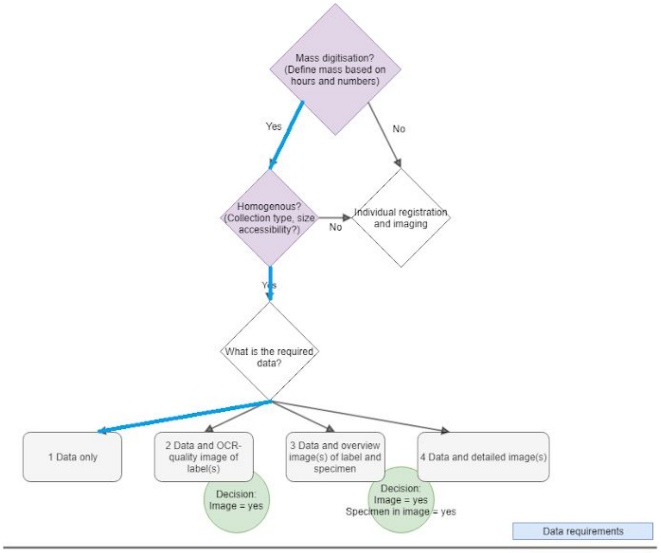
Van Walsum, M., Van der Mije, S. and Wijers, A. (2019). ICEDIG deliverable 3.4. State of the art and perspectives on mass imaging of liquid samples. (Will be available at https://icedig.eu/content/deliverables Q3 2019)

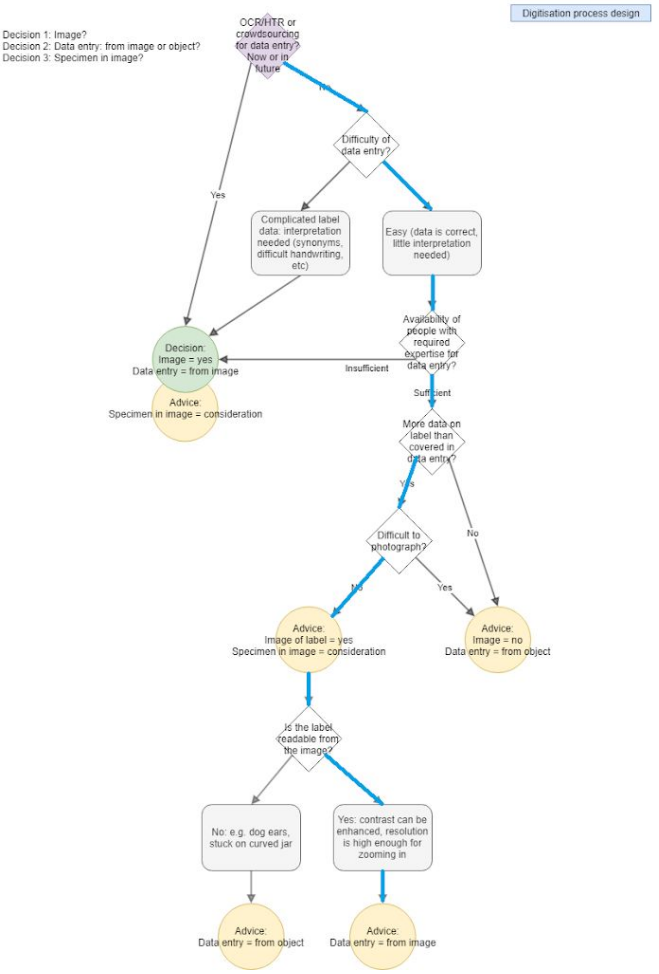
Vollmar, A., Macklin, J. A., & Ford, L. (2010). Natural history specimen digitization: challenges and concerns. *Biodiversity informatics*, *7*(2).

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Appendix 1 Examples to demonstrate the decision trees

Example 1 - research question driven: It is expected that most digitisation projects aimed at very specific research questions will find their endpoint at the start of the decision tree due to not being amenable to mass and not fulfilling the homogeneity requirement. It is possible that only the label data is required, then the flow continues on. If the research question calls for images of specimens, then it should be done with stricter quality control. The decision for imaging and including the specimen in the image are then set. The only decision which remains is whether to do data entry from the image or from the object.





Example 2 (see flow through decision tree in figure at right): the whole dry bivalve collection within the Mollusca group of an institute, comprising 200.000 samples is proposed for digitisation. The goal of the project is to perform data entry and add barcode labels. Data entry will be conform MIDS level 1: persistent unique identifier plus taxonomic name and geographical region. They are stored in drawers, there is some variation in size, but because size of the specimens is correlated with taxonomic groups which are also basis for the physical organisation of the material, this can be accommodated. Because the wet collection is stored elsewhere, the preservation type is fixed. With these properties, this collection fulfills the mass and homogeneity conditions.

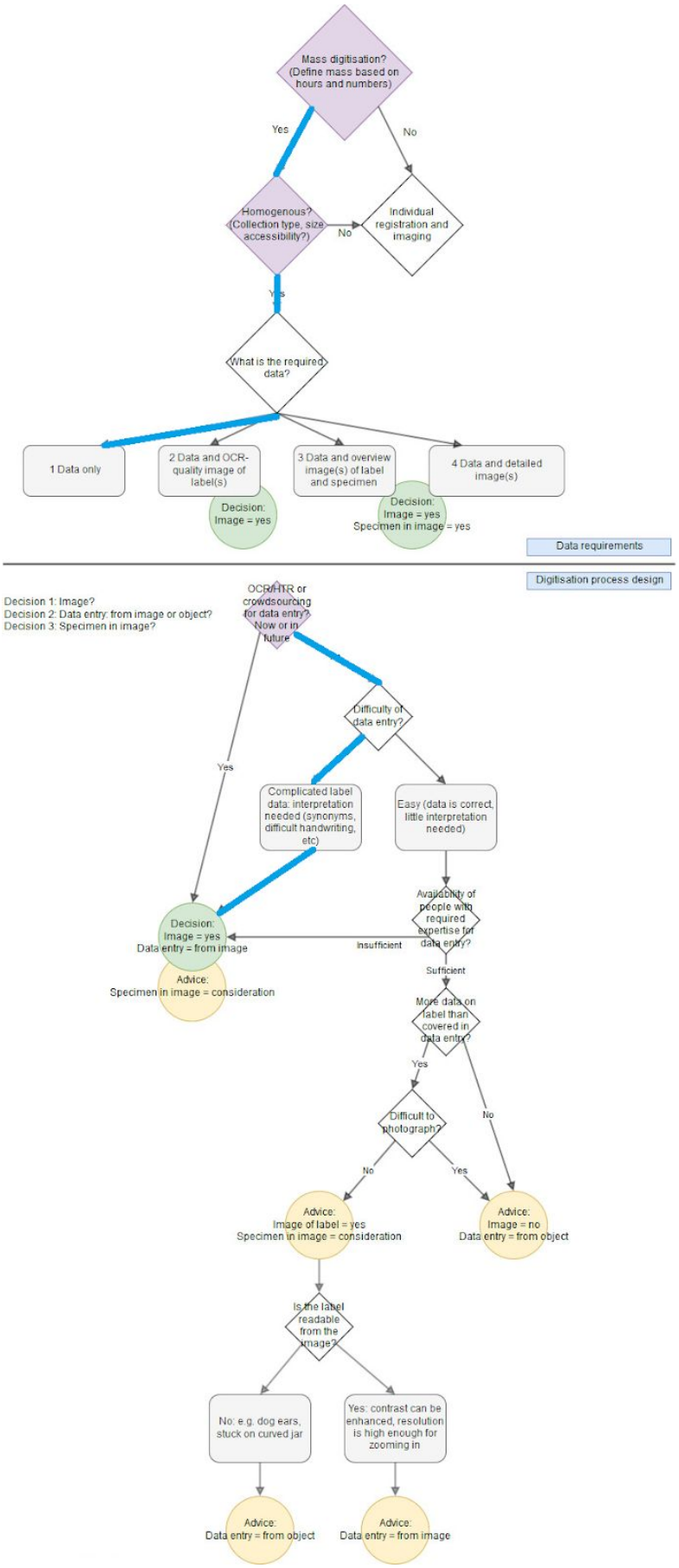
The project goals mean that images of high quality with specific views are not needed for research purposes. After this digitisation project, there may be requests for research quality images, which may or may not fulfill the mass and homogeneity requirements. OCR and citizen science are not considered for data entry. The data entry will be easy because the labels are not too difficult to read and does not need much interpretation: the taxonomic data and

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

geographical region are up to date. The limited amount of interpretation needed can be handled with a small team of operators and team lead with some knowledge of the collection. On the labels, there is more data than will be entered into the database, such as collector, collection date, detailed locality information. After a pilot digitisation project, it has been determined that photography of objects is relatively easy and results in a time increase of factor 1.5. While the aim of this project is to quickly digitise the collection with a minimum of data, the importance of the other label data is acknowledged and justifies the time and cost increase of imaging. In image form the data is then available when needed and can, at a later stage be transcribed and added to the record. Because of the images, the options for OCR, crowdsourcing, volunteers or dedicated staff are open without the need to access the material again for further data entry. Originally there was no great demand for images of the specimen, but because imaging is already taking place for each label it is decided that it is worth an additional increase in time. This way there is a timestamp for condition, and internal and external users will be able to assess which items to request for further study. Based on how well the labels can be read from the image, the choice can be made to do data entry directly from the object or from the image. In case of curved jars and creased labels, data entry from the object may be preferable. However, due to zoom and digital contrast and brightness adjustments it may be easier to transcribe from the image.

Example 3 (see flow through decision tree in figure below): the whole collection of bird skins needs to be digitised, fulfilling the mass and homogeneity requirements. There is no need for specific research driven imaging. But the labels contain a lot of handwriting, out of date names and places. This means that the data entry needs to be done by a small number of experts with sufficient experience in interpreting these labels. If these experts are the only ones working on it, the project will run much longer than if imaging is done by lesser experienced staff or volunteers and the experts use the images for data entry. Because images of the specimens help with spotting blatant errors and with curation, it is decided to also include the specimens in the images.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483



Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483