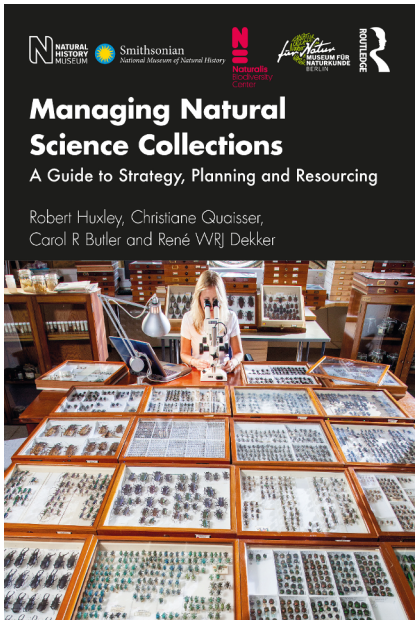
SAMPLE CHAPTER

Virtual Collections

Routledge
Taylor & Francis Group

www.routledge.com

Contents



1. **Virtual Collections**
By Robert Huxley, Christiane Quaisser, Carol R Butler, and René WRJ Dekker
from Managing Natural Science Collections: A Guide to Strategy, Planning and Resourcing

7 Virtual collections

> A digital collection is the twin of the physical collection but unlike the physical has the potential to be aggregated, linked and used as part of one very big worldwide collection which exists apart from institutions.
>
> (Vincent Smith, Natural History Museum, London)

Introduction

The demand for access to digital data and images is growing rapidly. While scientists, both academic and amateur, have known how to find their way into our collections for a long time, the general public, however, was mostly dependent on our exhibits to see and learn about our collections. But now the general public has become an important client for studying or simply enjoying the beauty and value of parts of our collections to which they never had access before. In addition, through rapidly booming citizen science projects, we ask the public to help us answer important biodiversity questions. For that reason, but not only, it is our desire and obligation as museum managers to make collections under our care accessible to everyone, with as few limitations as possible. Hence major institutions are digitising their collections on a massive scale, with increasing speed and using newly developed fit-for-purpose techniques. Our desire to expose our objects through the virtual world is offering new opportunities to present collections to an ever-growing audience more quickly and in different ways than ever before. The rapid development of specialised software and emerging technologies such as pattern recognition has helped us to achieve this. Simple databases no longer satisfy user communities' information needs and they demand digital images. Effective digitisation to meet these needs requires complex decision making on methods, standards and legal considerations. Museums are building upon the experience of others and now apply international standards to present images of the highest quality. By using additional tools, virtual objects can show much more detail than the physical object will ever reveal to the naked eye. Digitisation has also become a major competitor for funding, whether it is from within budgets or external, including crowd-sourcing or crowd-funding. Involvement of "the crowd",

whoever or wherever they are, is a powerful way to add content, and hence additional value, to our collections.

Despite the many advantages of digitisation, there are multiple pitfalls. Virtual collections do not, as many believe and even suggest, replace physical collections (see Chapter 2). One of the fundamental arguments is that virtual objects do not allow analysis of the physical material from which it is made or its genetic structure. They do not allow for any future analysis of characteristics using techniques that we are not currently aware of. One cannot "take a look" inside a photo as one can do inside a physical object. Virtual collections are not a replacement but an addition to the corresponding physical collections.

Digital techniques are changing rapidly over time. For this reason this chapter is not about techniques, hardware and software, but will focus on *strategic* management, drawing upon experience and examining factors to take into account when embarking on virtual collections projects. More technical and up-to-date information can be found in specialised journals, on websites and directly from institutions that are in the process of digitisation.

Questions

- Why should you digitise your collections?
- How do you create a digitisation strategy and plan?
- What should be digitised, how do you prioritise and who should do it?
- Which standards, software and hardware should you use?
- How much does it cost?
- What digital initiatives exist around the globe?
- What are the risks?
- What is the future for digitisation?

Why should you digitise your collections?

The advantages of digitisation are multiple. Information linked to the objects as well as the objects themselves, in fact entire collections, can be consulted on a global scale by a single push of a button. In terms of reduced risk, there is often no longer the need for the external and internal user to visit a museum or collection and physically handle fragile and valuable objects. For the collection manager there are none of the tasks associated with managing visitors, such as screening and assisting visitors or checking the collection after a visit. The institution will benefit as there will be no disturbance of the climatic conditions in the collection storage facilities because of the presence of multiple visitors. Much – but not all – collection management and research can be done away from where the objects are stored. Access will partly, or even largely, shift from physical to virtual collections. This however, asks for the end-users' confidence in the correctness and completeness of the data transcribed from the labels or captured as an image. Under pressure of time

and the desire to be cost-efficient, errors are easily made, especially when for example 19th century handwriting has to be transcribed or when transcription is done by individuals not familiar with the collections and/or poor knowledge of geography. Validation of these data by your own collection managers or researchers or collaborators is of utmost importance. It needs to be emphasised that if digitised data are not trusted and users detect errors or omissions, it will not be used by the scientific community, and if it is used, errors might lead to incorrect reproduction and false conclusions in scientific publications. As managers of scientific collections, therefore, we have a responsibility to ensure the correctness of the digitisation process and hence the balance between efficiency, cost and output. Virtual collections are not only for the general public as a virtual "exhibit", but have to serve scientists, historians and other researchers, including amateurs, around the world and hence have to meet the highest standards of accuracy.

Strategic benefits of digitisation include:

- Access – virtual collections give remote, global access to objects and related collection information, such as field notebooks and letters, which are otherwise hidden in archives and museums.
- Linking of object and collection information – by gathering, improving and combining digital information patterns and links can be made visible which otherwise are difficult to discover. Digitisation makes it possible to combine data quickly and easily on various parts of a specimen such as skin, skeleton, measurements, stomach contents and parasites as well as images, drawings and publications. This would otherwise be a laborious task.
- Validation of data – access to digital data allows the global community of scientists and amateur scientists to validate, correct or add data to objects which improves the quality of research. There are, however, serious risks associated with this as errors can be introduced (see below under "risks").
- Validation of identification – high resolution images, for instance, can reveal new information or information otherwise difficult to capture, which can also assist in the identification of similar objects in the collection. Automated wing-pattern recognition allows the identification of bumble bees in drawers of hundreds of objects and inconsistencies such as misidentifications can be resolved (see http://wing-id.naturalis.nl/).
- Retrieval of missing objects – digitisation helps to retrieve objects that were either identified incorrectly (see above) or have been misplaced among other objects, in drawers, cabinets, etc. in the past.
- Access to virtual collections – the effects digitisation will have on the number of visitors to physical collections or requests for physical loans and the consequences to updating information on the original object has not been answered unequivocally yet but is something you will

have to take into account when making strategic decisions. This will be discussed in Chapter 8.

- Unite distributed collections across multiple institutions – aggregate data to be statistically significant, adding scientific rigour, and supporting "Big Data" questions.
- Part of disaster preparedness – images of objects and labels do not replace the original object but document and save at least the status, condition and data. This is important in the case of deterioration processes, such as aging of mounting media in the microscopic slides, which makes objects less visible or blurs details through time. An image documents the status at the time of digitisation and can help future studies of the original status of the object. In case of damage or loss, caused for example by agents of deterioration such as insect pests, theft, vandalism, flooding, fire or war, information about these objects remains retrievable. Think of what happened with historical collections and objects in the wars in Syria and Afghanistan, much of which was destroyed or sold on the black market, and the devastating fire which destroyed millions of objects in the National Museum of Brazil in Rio de Janeiro in 2018. With images, at least a virtual collection remains. Similarly images can act as evidence of origin in the case of stolen artefacts offered for sale.

The new dimension: the history of mass digitisation

Why mass-digitisation? Natural science collections are vast, counting tens of millions of objects in the major international museums. It is simply impossible to digitise them in a one-by-one traditional way as was done in the early days. It is a simple calculation that, when it takes you five minutes – and that is an optimistic estimate – to handle, database and capture an image of an object the traditional way, you will do 480 objects per week or "only" 21,600 objects per year (45 working weeks). A collection of "only" 1 million objects will take you 46 years to complete or 46 staff for the duration of one year. That is no longer an acceptable option for the data-hungry user community and so we had to look for new, alternative business models, such as partnerships with external, more commercial partners, and we needed bigger budgets. This new approach was revolutionary only a decade ago.

Current projects are built on a legacy of digitising initiatives of varying scales and degrees of collaboration developed from the late 1980s. Often individuals would see the benefits of creating databases of specimens of taxonomic groups that they were interested in. Others might use databases as a new and better means to create catalogues of type specimens. A number of important initiatives began in the botanical community, such as barcoding of objects initially to record and track those going on loan and by extension to link label data to databases. Capturing of digital images of type and historical specimens became commonplace. These various data capturing

projects used a variety of technologies and database platforms, often varying within institutions. Bringing these disparate datasets together and merging them was a major challenge even before making them accessible via the internet. This process was made much easier by the data standards developed from 1985 by the efforts of Biodiversity Information Standards (TDWG), a group of mainly collections-based biologists. The sharing of experiences with different methodologies and technologies has led to the standardisation of image resolution and database fields often incentivised by large-scale funding such as the Global Plants Initiative which has required common standards and used relatively simple and affordable technologies to capture and make available much-needed images of plant type specimens in the world's herbaria. These factors have led to the mass of specimen data and images that are now available to all.

Large-scale digitisation of natural history collections really started in 2010 at the Muséum National d'Histoire Naturelle (MNHN) in Paris with the 2D scanning of ten million herbarium plants. It was revolutionary at that time. This was followed later that year by a 13 million euro project to digitise, almost at an industrial scale, a targeted 7 million zoological, paleontological, geological and botanical objects in Naturalis Biodiversity Center in Leiden, The Netherlands. On average, excluding money for management and staff, 1.85 euro was available per object. During the five-year project, processes were refined, new techniques developed and manuals written. Cost efficiency was crucial, leading to efficient digitisation of as many objects as possible. These initiatives triggered other institutions who visited Paris and Leiden to discuss, learn, copy and improve their own processes. But what seemed state of the art only five to ten years earlier has now become common practice in many collections worldwide. Outsourcing parts of or even entire processes, as in the digitisation of botany collections, to commercial companies has led to new business models, higher standards, higher output and lower costs. The same will happen over the next decade(s), for example with mass digitisation of insect collections which still set real challenges due to the complexity of label scanning and the fragility of the objects. The Museum für Naturkunde (MfN) in Berlin which, in 2018, was granted about 90 million euro for the digitisation of its entire collection of 30 million objects, might be best prepared to take up this challenge!

How do you create a digitisation strategy and plan?

Effective digitisation and access requires strategic decision making and in-depth discussions, advice on methods, standards and legal issues and benefits. So what is the strategic approach to take, especially by management and staff with less exposure to collections digitisation, such as those from a non-science or non-collections background?

Strategic approaches in a rapidly changing world

Increased access to collection data, standardised data storage according to international standards and linking the virtual with the physical object lead to multiple benefits, including:

- embedding digital management in workflows, leading to quicker and better collection services, lower cost of management and maintenance of physical collections and easier decision making;
- better data quality on websites, facilitating research among a wider audience;
- increasing availability of data of focus groups such as endangered, invasive or harmful species, allowing for rapid assessments;
- the advantage of delivering easily accessible statistics for scientific and management purposes;
- improving the exchange of information compared to the costly and potentially harmful exchange of physical objects;
- more efficient use of storage facilities through better insight into the size, composition and usage of your collection.

The question that is frequently asked by finance officers, funders and others is "What benefit, including generating income, will we see from this expenditure"? This is a difficult question to answer as sometimes benefits may not be realised immediately. In this way, digital collections do not differ from traditional collections. We invest in their preservation and accessibility knowing that we are providing a research facility that may lead to scientific discovery. It is not necessarily linked directly to the work that we and others currently do on them. On-demand digitisation, however, generating data for third possibly commercial parties, can generate income and encourage cooperation. It will help third parties instantly without them having to spend much time and effort in gathering these data from scratch or in any other way.

Exercise

A series of benefits of digital collections have been given above. What other benefits can you add, especially and specifically for your own collections?

The digitisation process

The digitisation process can be divided into three stages:

1 *Mobilisation*: connecting the physical to the digital. Mobilisation is the process of transforming data on labels for example, to digital form

and includes equipment, methods of transcribing data such as crowd-sourcing, optical character recognition (OCR) and enhancing data with georeferencing for example.

2 *Systems*: once the information is in electronic form there is a need for improving data quality and putting it into context such as climate research and resolving rights of use and access. This is mainly internal to an institution.

3 *Access and exploitation*: once the data is cleaned and contextualised it needs to be made accessible. This phase includes all the factors associated with use and exploitation, such as rules and regulations on its use, advertising digital services and exporting data to global aggregators such as GBIF.

Key challenges across these three stages are:

- sustainability and collaboration;
- culture;
- the right staff.

Sustainability and collaboration

Large specialised organisations are perhaps more likely to use best practice in digitisation than smaller and multidisciplinary institutions. The latter might, for instance, not be aware of global facilities and organisations such as GBIF, TDWG, CETAF and others. Furthermore, they might lack the specialised IT staff, such as data engineers and architects, who have the knowledge to make data easily portable and sustainable. However, this knowledge may be available in other institutions and national or international initiatives. Careful planning, using international standards and best practice and learning from others who have been through this process will help to avoid the pitfalls that others have experienced. There have been many collaborative digitisation projects, many of which have had a degree of success. But there are still the challenges of different institutions working in different ways to gather and share data. There does seem to be a gap between people and institutions involved in national and international projects on one side, and on the other, those managing collections on a day-to-day basis discussing the same issues that others discussed ten years previously. This is partly due to poor engagement with digitisation but is also due to staff either not seeing this as part of their career paths or not having the opportunities to engage with digital projects. Collaboration offers opportunities for smaller institutions to overcome this and ensure the sustainability of their data. Perhaps it is the responsibility of larger institutions with access to specialist staff and understanding of good practice to help smaller institutions.

Culture

The problem of sustainability relates to culture. Even though nearly all collection and research staff are now involved in digitising, willingly or not,

they are not necessarily part of that culture. There are many staff who are reluctant to engage. Proof of the value of digitisation, particularly to their own projects, might bring them on board. Staff selection and having staff who can act as interfaces between the technical staff and researchers is a possible mitigation. In large institutions digitisation is core business. It does not, however, receive the financial support one might expect. In some ways, small institutions have the advantage of scale and can transform discrete datasets into virtual sets using volunteers without competing for resources with other areas of museum management. Large institutions may have big budgets but they are divided across a wide range of arguably equally important functions such as exhibitions and education. There is also a cultural variation between those who want maximum (versus basic) data on each object no matter how long it takes. For example, if the policy is to 3D digitise a collection at two objects per day, it will take considerable time before a usable dataset will be completed. During the time it would take to digitise a comprehensive collection on that basis, technologies might change considerably, potentially rendering earlier work redundant and not matching up to current data standards. A simpler dataset can be completed in far less time and allows further detailed data to be added in later. Different collections ask for different business models as basic data needed to do sound research or efficient management differ between collections. This decision might also be influenced by demand. If, for example, an external research project requires 3D scans of 100 human skulls and there is funding to hire additional staff and equipment then it will surely go ahead.

The right staff

This topic leads on from culture and culture differs between institutions, big and small, multidisciplinary or with a single focus. There is a need for interfaces between subject matter experts and data technicians and architects, if at all available. These could be seen as translators who understand the subject and have a broad understanding of the digital world. The skills needed to make digital projects happen in a *sustainable* way are different to what we are used to in museums. There are a number of challenges:

- attracting data architects and engineers away from highly paid corporate posts;
- obtaining core funding rather than relying on externally funded posts with the associated lack of stability and continuity;
- a stable core of staff with the right skill sets (see Chapter 5).

One advantage of the virtual work environment is that staff, but also volunteers, can work remotely and through the latest communications technology they can be supervised, supported and remain in constant communication with the team. Over the years, commercial companies have learned about our collections needs and have invested in specialised technologies.

Nowadays, large herbarium collections of 1 million or more sheets of dried plants are imaged, barcoded and databased faster, cheaper, safer and of a better and constant quality by commercial companies than institutions can do it themselves. A company that developed and improved its working methods with Naturalis Biodiversity Center in Leiden between 2010 and 2015 is now digitising herbarium collections in multiple European countries and the USA, showing that making use of what others did before can be a very rewarding way of completing what once seemed impossibly large tasks.

What should be digitised, how do you prioritise and who should do it?

Is waiting to digitise your collection still an option? Not being the first, avoiding the law of the "handicap of a head start", might pay off. But does it really? Do you want, if finances allow, to be a pioneer and work with others to develop new techniques for collections which challenge you to get objects digitised efficiently or do you want to be last in line? Prepare your collection in such a way that it is ready to be digitised. It is many times more costly and time consuming to digitise a collection when it is not stored systematically or in any other logical sequence. The poor state of any collection might be an argument not to digitise it as there is a risk that it will be damaged even more through the digitsation process. But is the risk worth it? If in a botanical collection the specimens are unmounted, loose and more likely to be damaged by the digitisation process than those fixed to herbarium sheets, it could be argued that the risk is worth it if the collection is going to be digitally available rather than sitting in a box safe but unused. In such instances it might be better for your qualified staff to handle it once and get it digitised than if it will be used on a regular basis by researchers who might be less careful handling these precious objects. Also think about whether to combine different processes by the same person at the same time or try to organise it in such a way that it becomes an automated routine single focus process for those who do the job, even though this might not be that inspiring for your staff and volunteers. Rotating your staff in such a production line on a regular basis so that they get to do all parts of the process, may make it more inspiring though. The New York Botanic Garden shows how to deal with this: http://sweetgum.nybg.org/science/information-management/digitization/digitization-resources/.

Exercise

Consider the advantages and disadvantages of the various options with regard to efficiency and value for money for the various collections in your institution.

If collections are vast, say between 100,000 and 1 million objects, and organisations have restricted budgets or staff resources then choices need to be made as to what should be digitised first, what later and what not at all; this should be part of your strategic plan. Arguments for digitisation can vary, even within an organisation, ranging from content-related to efficiency-related reasons: for example a choice between important historical or scientific collections which are under active study and less important collections which might be stored remotely. One argument for the latter is that it will enable collection managers and researchers to consult such collections at low cost without having to travel to off-site storage facilities. Choices can be made between basic versus in-depth digitisation, from simple databasing and imaging of labels to high quality 3D or scanning photography. Or scanning of whole boxes, drawers and jars in which multiple objects are stored allowing quick assessment of what is stored where and how without checking or opening the containers physically. It will also give visitors a preview of what you have stored before actually visiting you. Apart from these internally driven priorities, your clients in the "outside world" might request a collection to be digitised (digitising on demand) or let you decide from a strategic point of view what to digitise. Once choices have been made, it does not exclude additional choices at a later stage.

The entire process from prioritisation based on research needs, collection management priorities and external requests, through to making the data available to users can be laid out in a flow chart. Figure 7.1 illustrates the schematic overview drawn up for the large-scale digitisation project that ran from 2010–2015 at the Naturalis Biodiversity Center. The key steps can be followed: prioritisation of collections, handling of the objects, (re)identification where necessary, the practical process of digitisation (how and where), storage, indexing and finally disclosure of the virtual data.

The *what*, *how* and *by whom* questions need careful consideration within the budget available. The question of who should do the digitisation and annotation, whether it is part of the job description of museum staff, done by volunteers within or a "crowd" outside an organisation, or even executed by commercial companies, depends on factors such as size of the collection, budget available, techniques to be used and how quickly the data are needed by the users. Expert users of your digital data might be able and willing to help you by adding additional data such as geo-references to your objects as they work on them for their own study and publications. Consider what options you have and whether your institution would benefit from either a slow start involving your staff right from the beginning, making errors but learning from them as a natural process or, whether you might want to jump ahead and do what others have been doing as long as you do not forget your strategic considerations. But remember, relevance and the justification for maintaining a collection is supported by use. Today, most uses include some virtual aspect, so choosing to *not* digitise at all may be choosing to become irrelevant.

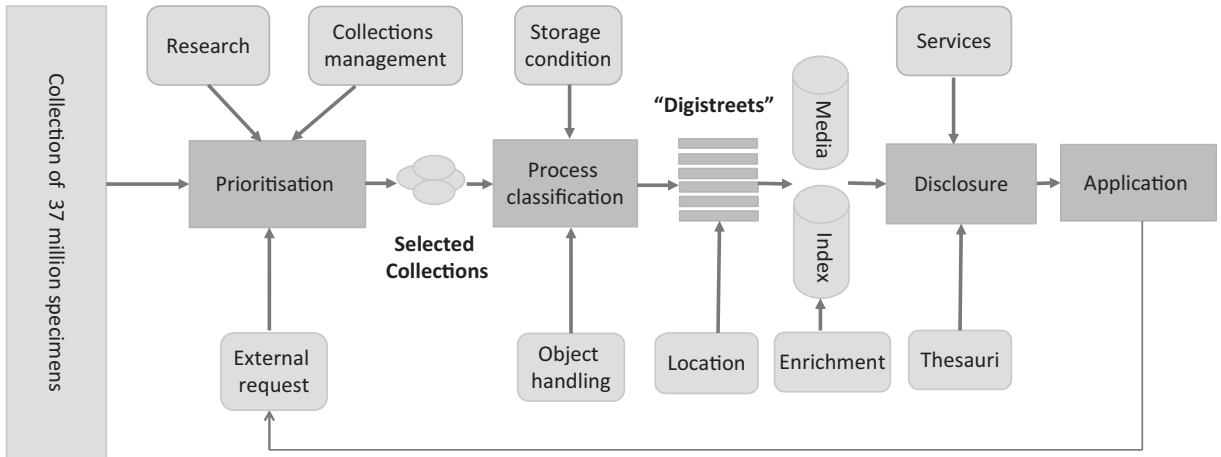


Figure 7.1 Schematic overview of the Naturalis approach to digitisation process organisation. "Digistreets" are a production line for preparing and capturing images and data

Source: Naturalis Biodiversity Center.

Exercise

- Which of the collections you manage is your preference to be digitised?
- What criteria would you apply to select this collection over other collections?
- Which of your stakeholders need to be involved?
- Who will digitise your collection and how are you going to organise it?
- Make a decision tree to document your decisions and possible alternatives.

Which standards, software and hardware should you use?

Digitised collections, whether large or small, need to be created according to common standards and be interconnectable. In the world of natural science collections, the Taxonomic Databases Working Group (TDWG/Biodiversity Information Standards, www.tdwg.org, and Integrated Digitized Biocollections, iDigBio, www.idigbio.org) are platforms for this. The TDWG community's priority is the development of standards for the exchange of biological/biodiversity data. It is a not-for-profit scientific and educational association that is affiliated with the International Union of Biological Sciences, formed to establish international collaboration among biological database projects. Their mission is to develop, adopt and promote standards and guidelines for the recording and exchange of data about organisms and promote the use of standards. iDigBio's mission is, among others, to develop an infrastructure by overseeing implementation of standards and best practices for digitisation.

As software and hardware are continuously being improved, changing rapidly over time, we are not going to advise on which to choose. Let yourself be advised by institutions who have gone through that stage and are still working on or have recently completed the digitisation of such collections; you can do this by direct contact or through appropriate networks. For reasons of efficiency and uniformity, it is wise "to follow the leader" and look at initiatives taken by CETAF, SPNHC, TDWG and others and start from there and make the necessary improvements if needed or advised based upon their experience. The same is true for data storage. There should be no competition between institutions as sharing our virtual collections is a common goal as shown by the Distributed System of Scientific Collections (DISSCo) initiative for instance.

Common principles on digital data

There are global initiatives that provide guidelines or common principles on digital data. One such initiative within the EU-funded SYNTHESYS

project aimed, among others, to develop and adopt common principles for management of digital data produced by external users to balance their costs and benefits. This document, published in December 2016, is available on the SYNTHESYS website at http://synthesys3.biowikifarm.net/syn3/NA2/objective1/task1/cmp. It provides help in maintaining the highest standards for the management of digital collection data and creates a framework for the management and conservation of digital data produced by external users of natural science collections hosted by European institutions. The Wiki-Handbook *Recommendations: Management Policy on Digital Collections (MPDC)* gives advice to institutions on how to set up institutional management policies on digital collections. It lists subjects that need or can be covered in an MPDC and provides explanations as well as definitions on these. Institutions can choose the elements applicable to their digital collections in compliance with institutional as well as national regulations. It is a freely available online Wiki-Handbook, can be updated, added and amended by its users and so will continue to be up to date.

If you decide to add 3D images to your digitised data, the *Handbook of Best Practice for 3D Imaging for Natural History Collections* might help you: http://biowikifarm.net/v-mfn/3d-handbook/3d_Imaging_Handbook:About. The scope of this handbook is to present different digitisation techniques currently in use in European institutions. It provides guidelines by showing which techniques can be used for a certain collection or object. In addition to the general information on particular techniques, it provides a thorough workflow, an equipment list and a test case per institution on how to create a good digital replica of an object.

Open access

In 2019, nearly 100 natural history institutions and related organisations committed to promoting open access have signed the *Bouchout Declaration for Open Biodiversity Knowledge Management* (June 2014) (http://bouchoutdeclaration.org/declaration). The Declaration seeks to unlock the legacy of biodiversity knowledge, make digital data about biodiversity openly available and put it to work for the benefit of all. It calls for legal and technical reforms to grant all users the right and ability to freely access, copy, distribute and re-purpose published biodiversity content, accompanied by the respectful attribution of the content source as demanded by scholarly tradition. However, while open access is the default position for many museums, we still have to honour the agreements about access that were made when we acquired the object which underscores the importance of digital access to legal records so staff can verify terms of use. See also: "The FAIR Guiding Principles for Scientific Data Management and Stewardship" (Wilkinson et al., 2016).

How much does it cost?

The costs for digitisation will differ significantly depending on many kinds of variables, such as:

- simple transcription of basic data versus capturing all data including remarks and additional notes from labels;
- variable handling time, for example, from easy to handle herbarium sheets to labour intensive wet collections in jars;
- time spent on transfer of objects from storage space to digitisation lab – the closer the cheaper (and safer);
- the legibility of information on labels, e.g. hard-to-decipher 19th century handwritten versus easily-readable 20th century printed labels;
- the decision on whether or not to image objects and if so whether to choose 2D, 3D or microscopic imaging;
- the duration of the experience and efficiency gained during a project and the number of digitisers involved, making "conveyor belt-like" work possible or not.

As an example, Table 7.1 shows the average costs in euros (including overheads) calculated for digitising different types of specimens in the Naturalis Biodiversity Center in 2014.

Careful planning can save you a lot of money and will enable you to digitise much more within your budget.

Exercise

Make a cost estimate for the collection you selected in the previous exercise.

Table 7.1 Costs for digitisation: average cost in euros (including overhead) per object at Naturalis Biodiversity Center in 2014

Collection	*N Realised*	*Euro/object*
Wood samples	125,000	1.27
Herbarium sheets (Leiden)	2,600,000	1.29
Molluscs	640,000	1.37
Herbarium sheets (Wageningen)	830,000	1.47
Entomology	530,000	1.51
Microscopic slides	510,000	1.57
2D objects (books, journals, etc.)	450,000	1.87
Geological specimens	50,000	1.90
Vertebrates (dry)	170,000	2.37
Specimens preserved in alcohol	60,000	4.65

Crowd-sourcing as an example

In 2013, Naturalis Biodiversity Center in the Netherlands experimented with online digitisation of collection data through crowd-sourcing. The aim of the project was to have labels of 100,000 glass slides transcribed and validated by a community of online volunteers within a time span of six months, at a cost comparable to that of the internal process of digitisation of glass slides. The project was carried out on an existing online platform for transcription of written heritage: https://velehanden.nl.

Each glass slide was transcribed by two different volunteers to check for and avoid errors and validated by a third volunteer, amounting to 300,000 treatments in total. On average, the labels of 711 slides were transcribed per day with a peak of 1,913 slides. The quality of the work was very high and did not differ from in-house digitisation. Nevertheless, the project duration was exceeded by three months, mainly due to an overly optimistic estimate made prior to the project.

A total of 497 volunteers signed up during the project with a broad demographic spread and the majority aged over 55. The really active group consisted of only 203 volunteers, while approximately 60% of the whole volunteer force transcribed less than 10 slides. Of this active group, 26 individuals were responsible for 90% of the work and the two most active members together accounted for over 68,000 transcripts. This limited number of active volunteers may reflect cultural differences and perhaps different results would have been observed in countries where volunteering was more embedded in society.

The total cost of the project was 33% higher than the originally estimated 60,000 euros, mainly because it lasted for nine instead of six months. The cost per slide turned out to be 0.79 euro which is 9 cents higher than in-house transcription.

Given the limited scope of this pilot and the somewhat specialised target group, the recommendation was to investigate how crowd-sourcing activities might attract a more diverse audience. A good example is provided by the "herbonautes project" of the Muséum National d'Histoire Naturelle (MNHN) in Paris. Instead of transcribing objects at random without a direct relation to one another, as in Naturalis, the MNHN stimulated volunteers to be responsible for a specific but limited part of a collection, giving volunteers a sense of ownership and responsibility for "their" section (see http://lesherbonautes.mnhn.fr and www.researchgate.net/publication/295688898). The "herbonautes project" has been implemented at other herbaria, e.g. the Botanischer Garten und Botanisches Museum Berlin (BGBM) (see www.herbonauten.de/).

What digital initiatives exist around the globe?

The rapidly increasing number of national and international initiatives presenting and linking biodiversity data on the web makes an extensive

overview of such initiatives out of date as soon as it is published. The reader is advised to search the web to see what is available and what is new, from early initiatives, such as the Global Biodiversity Information Facility (GBIF), VertNet and Europeana, to the more recent Innovations Consolidation for Large Scale Digitisation of Natural Heritage (ICEDIG) and Distributed System of Scientific Collections (DISSCo). The aim and mission of these initiatives as extracted from their websites are given below as an example.

Examples of large-scale digitising initiatives

GBIF (www.gbif.org): "The Global Biodiversity Information Facility is an international network and research infrastructure funded by the world's governments aimed at providing anyone, anywhere, open access to data about all types of life on Earth. GBIF provides data-holding institutions around the world with common standards and open-source tools that enable them to share information about where and when species have been recorded. This forms the basis of GBIF's index of hundreds of millions of species occurrence records".

VertNet (www.vertnet.org): a "collaborative community effort to bring together the expertise of biodiversity scientists and information experts making biodiversity data free and available on the web. VertNet is a tool designed to help people discover, capture, and publish biodiversity data. It is also the core of a collaboration between hundreds of biocollections that contribute biodiversity data and work together to improve it. VertNet is an engine for training current and future professionals to use and build upon best practices in data quality, curation, research, and data publishing".

Europeana (https://pro.europeana.eu): an online collection of millions of digitised images containing material, including natural history, from European museums, libraries, archives and multimedia collections.

ICEDIG (https://icedig.eu): "The Innovation and Consolidation for Large Scale Digitisation of Natural Heritage is an EU-funded project that aims at supporting the implementation phase of the new Research Infrastructure DiSSCo (Distributed System of Scientific Collections) by designing and addressing the technical, financial, policy and governance aspects necessary to operate such a large distributed initiative for natural sciences collections across Europe".

DISSCo (http://dissco.eu/): "The Distributed System of Scientific Collections is a pan-European Research Infrastructure initiative with a vision to position European natural science collections at the centre of data-driven scientific excellence and innovation in environmental

research, climate change, food security, health and the bioeconomy. 115 organisations across 21 countries have already committed to DiSSCo. Its mission is to mobilise, unify and deliver bio- and geodiversity information at the scale, form and precision required by scientific communities: transforming a fragmented landscape into a coherent and responsive research infrastructure".

What are the risks?

Digitising collections and creating a virtual "new" collection is not without challenges. There are many factors which will have to be taken into account in the planning process. First of all, virtual collections do not replace physical collections. In fact, it adds a new collection to your collections: the virtual collection.

Some, both within and outside institutions, still believe that once a collection is digitised, the physical collection can be discarded (see Chapter 2). This might be the case with collections such as newspapers or journals which are not unique, but not with natural science collections consisting of unique individuals and samples or with man-made objects such as artworks. Virtual collections do not allow for analysis of, among others, DNA, isotopes, chemicals, minerals, materials used or anything else we might be able to discover and examine in the future.

Virtual and physical collections are different types of collections which need their own strategy and management. Furthermore, virtual collections require specialised staff, IT support, software, hardware and storage facilities and hence imply significant, structural costs. If not well planned, pitfalls lie within each of these components. They will differ between organisations of different sizes and collections based on budget, capacity and opportunities. This does not mean that only institutions with large budgets or status can afford to digitise. Smaller museums can draw on volunteers and the crowd to capture data for use within their institution and local user community, but can also make their data available to the wider community by contributing data to aggregators such as GBIF. To all organisations large or small there is a further, less obvious risk. In this era of digital transformation, you may be risking your collection becoming invisible, unused and irrelevant by not ensuring digital access. This could result in less funding or even closure of collections and institutions.

One general issue which is yet to be resolved is the continuing addition of annotations, such as new names to specimens by scientists and the "crowd" to databases and virtual collections. Visitors to the collections and borrowers of physical specimens are mainly experts in their group of plants and animals and add valuable data on current names, misidentifications and type status on attached paper slips for example. The same specimen may

or not have information updated on either or both its physical or virtual representation. Virtual visitors may or may not return changes and there is often no mechanism for them to do so. As more users consult virtual rather than physical objects the problem will get worse. The risk here is that the physical collections will become out of date and of less value and accessibility. The information users get from the virtual object will therefore differ from what one gets from the physical object which lacks these annotations. The problem is magnified if the digital records available through portals are not kept current with the institution's digital records. Solutions to this can only be reached at an international level, by multiple parties involved.

There is always a risk of inaccurate data being added. If you plan to digitise your collection through crowd-sourcing, how do you mitigate against this and ensure accuracy? How do you avoid the introduction of errors and inconsistencies by your crowd? There are several ways to validate or cross-check these data, as illustrated in the section "Crowd-sourcing as an example" above.

Other risks include:

- Damage to and loss of objects during the digitisation process.
- A shift of attention to virtual objects resulting in less routine inspection of physical collections through access and less detection of insect pests for example. These risks are balanced by possible reduction in handling damage.
- Dilemma of division of distributing limited resources between physical and digital collections resulting in suboptimal care for both.
- The link between the physical object, the voucher and digital object becomes diluted or lost. The digital object will "live its own life" no longer linked to the original object resulting in separation of information between the two.
- Unaffordable costs of updating and maintaining data leading to "orphaned" and out of date virtual collections.
- Data being released without authorization.
- Obsolete media and digital records being held on unsustainable platforms.
- Highly unique data structures that prevent or slow down aggregation making digitisation efforts futile.
- Inability to meet users' needs for data because of inadequate planning.

Exercise

What are the pitfalls regarding digitisation for the collection you selected and the choices you made in the previous exercise?

What is the future for digitisation?

Much of what was considered state of the art in digitisation only a few years ago, has improved, been made more efficient and has become cheaper. Waiting to digitise is no longer an option. Do not wait for it to become even cheaper as you will miss opportunities, partnerships in international projects, visibility, etc. Digitisation is accelerating and becoming more cooperative. It has become one of the top priorities of museums and herbaria and not only in natural sciences. Data quality will continue to be improved and additional links will be created between objects and whatever is considered essential to these objects. Links can be made between the virtual objects and publications and websites in which these objects have been illustrated and described going back 200 years or more. Links can be made between virtual objects and additional information from whatever sources about the object, collector, artist or geographic origin, the latter for instance via websites that provide satellite views. Virtual reality technologies have created virtual exhibitions, based on high-quality 3D models (Bruno et al., 2010), while artificial intelligence and machine learning are "very powerful tools and are more accessible than ever before. In the hands of museums, these technologies will inevitably lead to interesting discoveries, rich data, and new paths into your collection" (Ciecko, 2017). Options are multiple and seemingly endless, illustrating the ever growing discrepancy between enriched virtual objects versus much more static physical objects.

Summary

The increasing demand for collections data in web-accessible digital form is creating new virtual collections which, like the physical collection, require careful management and curation to ensure currency and long-term value. Data can be of varying quality and one major challenge is to ensure that your data are of such quality to be usable by the community.

Virtual collections are not a replacement but an addition to the physical collection. They are collections in their own right and perhaps need to be treated as such. Virtual collections are a great benefit to collection management tasks and improve access to collections, raising it to a new level offering new opportunities to use and link collection information. Compared to physical collections, to which access is controlled, online virtual objects can be annotated by a crowd of anonymous users linked only to the virtual, but not the physical, object. There are associated risks; for example discrepancy in information linked to the same object causing possible misinterpretations or worse, errors in scientific communication. However, the additional data linked to a virtual object greatly increases its research value.

Planning digitisation projects to guarantee optimal access to your data requires complex decision making on methods, standards and legal issues. Resourcing is a key issue and options such as crowd-sourcing, dedicated digitising teams and new technologies are reviewed. Fortunately there are many examples of digitising projects available from which you can learn.

The key strategic questions, which can be different for each institution, are: why should digitisation be carried out; to what level should it be done and on which collections?

Start simple, step up production when you can, add and improve later. Visibility of your collection can come first by broad and shallow data gathering. You can prioritise which collections need to be digitised in more depth later on the basis of end-users (of which you as holder of the collection are one) – demonstrating what is really of importance. This can be from a scientific, public, educational and managerial point of view. If you need to add images, add images. If you need additional data, add additional data. And if certain collections prove to be important, you might be able to generate additional funds more easily and from unexpected sources, even crowd-funding.

References

Bruno, F., Bruno, S., De Sensi, G., Luchi, M.-L., Mancuso, S., and Muzzupappa, M., 2010, From 3D Reconstruction to Virtual Reality: A Complete Methodology for Digital Archaeological Exhibition. *Journal of Cultural Heritage*, 11: 42–49.

Ciecko, B., 2017, Examining the Impact of Artificial Intelligence in Museums. *MW17: Museums and the Web 2017*. https://mw17.mwconf.org/paper/exploring-artificial-intelligence-in-museums/ [Accessed 15/0819]

Wilkinson, M. D. et al., 2016, The FAIR Guiding Principles for Scientific Data Management and Sewardship. *Scientific Data*, 3: 160018. (doi: 10.1038/sdata.2016.18)

Further reading

Ang, Y. et al., 2013, A Plea for Digital Reference Collections and Other Science-Based Digitization Initiatives in Taxonomy: Sepsidnet As Exemplar. *Systematic Entomology*, 38, 637–644 (*and references therein*).

Blagoderoc, V., and Smith, V.S. (eds.), 2012, *ZooKeys 209: No Specimen Left Behind: Mass Digitization of Natural History Collections*. Sofia-Moscow: Pensoft.

Enghoff, H., 2017, Does Digitization of Natural History Collections Reduce the Need for Physical Access and Physical Loans? Report on Subtask 3.2.1 Under Task 3.2: "Facilitating Access beyond SYNTHESYS3".

Heerlien, M., Van Leusen, J., Schnorr, S., Jong-Kole, S. de, Raes, N., and van Hulsen, K., 2015, The Natural History Production Line: An Industrial Approach to the Digitization of Scientific Collections. *Journal on Computing and Cultural Heritage*, 8 (1): 3. (doi: http://dx.doi.org/10.1145/2644822)

Leusen, J. van and Heerlien, M., 2012, Digitaliseren van de entomologische collecties van Naturalis Biodiversity Center. *Entomologische Berichten*, 72 (5): 259–262. www.repository.naturalis.nl/document/453900/ [Accessed 15/0819]

van den Oever, J.P., and Gofferjé, M., 2012, "From Pilot to Production": Large Scale Digitisation Project at Naturalis Biodiversity Center. *ZooKeys*, 209: 87–92.