INITIATING A COLLECTION DIGITISATION PROJECT

Christopher K. Frazier[1], John Wall[2] and Sharon Grant[3]

Abstract:

This document is designed to give the reader the confidence to get started and to make the right decisions when planning a natural history collection digitisation project. The authors have years of experience working with collections and they have instilled this expertise into this paper so one can more efficiently ask the right questions and make the appropriate plans prior to committing any resources to the task.

GBIF
www.gbif.org

[1] Dept. of Biology, University of New Mexico, Albuquerque, NM 87131, United States of America
[2] Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, United Kingdom
[3] Natural History Museum, 15 Cromwell Pl, Kensington, SW7 2LA, United Kingdom

Initiating a Collection Digitisation Project

Published by:
Global Biodiversity Information Facility
http://www.gbif.org/

Digitisation of Natural History Collection Data ISBN: 87-92020-05-4

Recommended citation format:

Frazier, C.K., Wall, J., and S. Grant. 2008. *Initiating a Natural History Collection Digitisation Project*, version 1.0. Copenhagen: Global Biodiversity Information Facility. 75 pp.

While the authors and the publisher have attempted to make this book as accurate and as thorough as possible, the information contained herein is provided on an "As Is" basis, and without any warranties with respect to its accuracy or completeness. The authors and the publisher shall have no liability to any person or entity for any loss or damage caused by using the information provided in this book.

Comments and suggestions can be submitted to:

GBIF Secretariat
Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark
Telephone: +45 3532 1470
Fax: +45 3532 1480
E-mail: gbif@gbif.org

Cover image © Derek Robertson 2006
Fringilla coelebs Linnaeus, 1758

Initiating a Collection Digitisation Project

Contents

Section 1 Introduction 1

Purposes of this document 1

What do we mean by digitisation? 2

Target audience & scope of this document. 2

Document Overview 3

Section 2 Why should we digitise our collections? 3

Section 3 What you should do before you get started 5

Planning is important 5

Identify your goals 6

What are your current limitations and resources? 11

Produce a Tentative Business Case 17

Pick a database solution 18

Develop an Action Plan 18

Running the project. 24

Section 4 Organising Information and Representing Data 26

Object Information (as you see it) 26

Reference and Ancillary Information 28

Data (as the computer sees it) 28

Data Views and Storage Formats 32

Standards 33

Data Quality 35

Language 39

Intellectual Property Rights 39

Section 5 The Data Model 39

Introduction 39

The Base Unit of a Simple Data Model 40

The Focus of a Data Model 41

Modular Design of the Data Model: The Information Management System 45

Implementing a Data Model 46

Data model solutions vs. programmatic tools. 48

Complexity 48

Section 6 Deciding on a particular database solution 51

Which database(s) should you use? 53

What are characteristics of a good database solution? 55

Does it have the right focus and features for you? 60

References 65

Appendix A: Business Case Considerations 69

Why digitise your collection? 69

Identify your goals 69

Appendix B: Action Plan Issues 69

This paper is equivalent to Chapter 2 in:

Global Biodiversity Information Facility. 2008. *GBIF Training Manual 1: Digitisation of Natural History Collections Data*, version 1.0. Copenhagen: Global Biodiversity Information Facility.

Section 1: Introduction

Purposes of this chapter

Deciding to undertake a digitisation project can be a daunting process. By now, you've probably already noticed that "everybody is doing it." You've probably developed a strong suspicion that you are in danger of getting left behind in this information age if you don't start something soon. Maybe you're being pressured from one or more user groups, your administration, or colleagues to provide electronic data. Or maybe your main reason for showing an interest is that you've run up against real shortcomings in fulfilling your mission that you feel could be addressed better if your collection were digitised.

Regardless of your collection size or your reasons for wanting to start a digitisation project, your core interest probably isn't learning a huge amount about bioinformatics, system design, or information theory. You want to get it done, but where do you start? What do you really need to know in order to find a solution that meets your needs and resources without having to get another degree in computer sciences along the way?

There is simply no getting around the fact that you will have to consider a number of factors and make a large number of decisions if you want to get a workable solution that meets your needs. We know from experience that the first step most people go through when they are starting out is to think they'll just do what everybody else is doing. If only it were that simple. There are currently dozens of solutions in use, some home-grown, some commercial, some open-source. Some are tailored to your kind of situation, some are more generalized but customizable, and virtually all are going to be limited in some way that is ultimately important to you. Some are complex and powerful, some are simple, some are well-documented and some are not. And it's not just enough to know what is out there because some people or institutions are really happy with what they use, some are not, and most are somewhere in the middle.

This document is designed to give you the confidence to get started and to help you make the right decisions as you plan a digitisation project. We, the authors, have years of experience working with collections large and small and we have tried to instil this into this document so you can more efficiently ask the right questions and make the appropriate plans prior to committing your resources. Our goal in this paper is to provide you with the information that we wished we had at the outset of our digitisation projects. We've tried to give you the information that you might not get by talking to representatives or advocates of a particular solution.

Both information technology and informatics theory are evolving rapidly and finding the right solution for your needs is something of a moving target. Specific solutions that were in vogue just a few years ago now seem quaint or antiquated. In some cases, there are new possibilities on the horizon that might seem just what you need, but will they deliver as promised? Should you invest your time perusing the latest conference proceedings rather than reading through this document?

We have strived to make this document more relevant and less time-dependent by focusing on a simple fact we have learned about all IT systems. No matter how powerful, cutting-edge, or expensive they might be, an IT system is only as valuable as the quality of the data it contains. In the future or now, you will want to find a solution that allows you to enter, maintain, and output quality information. Thus, our emphasis in this document is to help you determine how to do that given the resources you have and your particular goals. This, in itself, is an extensive topic with many ramifications which may not be anticipated by those just starting out. There are several areas which have already been well covered in previous works and, rather than repeat those discussions, where appropriate we provide suitable references that should be consulted by those needing fuller understanding of the topic. The ideas exposed in this paper are the best we can come up with given our current experience working with a variety of digitisation projects and we hope they are helpful to those researchers in a similar position to ourselves.

Though in some cases we will recommend particular solutions in this document, overall our intent is not to provide a set of "best practices." It is our opinion that the appropriate solution for a specific institution or collection is highly dependant on the unique circumstances of each institution, so no single best practise exists. Getting the correct answer for yourself is the first step in implementing your own successful project.

What do we mean by digitisation?

Digitisation refers to the capture of information in electronic form. The basic information of concern to our community comes from checklists, field notebooks, collected specimens or may be extracted from publications, documents or other media. The basic unit of information may concern physical objects, like specimens in an herbarium, or events, such as observations of birds singing in a forest on a particular day or the collection of a number of organisms in a pitfall trap left overnight. The results of digitisation can be stored or expressed in a variety of ways such as formatted or marked-up documents, spreadsheets, web pages or web sites, flat-file or relational databases, maps or GIS systems.

Another important distinction for our community is that digitisation may refer to the electronic capture of an *image* of an object or it can also be used to refer to the capture of *textual information* about an object or extracted from an object that contains text. While both of these may be referred to as digitisation, we prefer the term "imaging" for the former. The specifics of imaging specimens is not dealt with in depth in this document. A good source of information on imaging is the Global Biodiversity Information Facility (GBIF) paper "Digital Imaging of Biological Type Specimens: A Manual of Best Practice" (Häuser et al., 2005). The term "databasing" is preferred to describe the process of capturing text-based information. In addition, many digitisation projects involve creation of an information system that holds both images and text-based information. In this case, "databasing" or "digitisation" refers to the creation of this system to hold information.

Target audience & scope of this chapter

This document is aimed at anyone who is interested in starting a digitisation project for the first time. It is hoped that it will also be useful to more experienced digitisers as well. The central focus of this document is to assist managers of specimen-based biological collections. Even so, it is recognized that managers may want to digitise their collections for a variety of reasons and with different goals for their digital information. Some, for example, may simply want to create an electronic catalogue of what they have, but others may be trying to create detailed records of their collection events, and still others may be looking for a way to integrate their voucher collections into an information system supporting project-based research at their institution. In this document, we provide information that we hope will be useful in planning a digitisation effort regardless of the specifics of one's particular goals.

Digitisation projects can vary in size from a single person digitising a single specimen to an institution wide program recording millions of specimens. Many of the issues discussed in this document are independent of the size of the collection or scope of the digitising effort. In other cases, we seek to provide guidance specific to the scope of the effort.

This document does not assume any particularly high level of literacy in database design, computer technology or natural history. However, in some cases, the reader may be directed to ancillary documents to help promote a better understanding of the concepts discussed here.

Chapter overview

The next Section "Why should we digitise our collections" is a short introduction to the reasons for undergoing a digitisation effort and the benefits you can expect to obtain. Section 3, "Before you begin" is broken down into several major parts. The first four sections are concerned with setting out your goals and analysing your current situation. Once you have done this, you can then begin to consider the specific details of how you are going to implement your project, covered in 'selecting an appropriate database' and 'writing a tentative action plan'. The final part addresses putting your plan into action. Section 4 discusses information (as you see it) versus data (as the computer sees it). This distinction is used to explain how the information you want the system to hold will have to be manipulated in order to be entered and maintained in a database. This Section also discusses other data issues including standards, data quality, language and intellectual property rights. Section 5 discusses the concept of the data model, starting with simple catalogues and ending with the modular design of information management systems. The Section then continues into how data models are implemented in computer systems and discusses some of the basic issues involved with implementation. Section 6 concludes with more detailed discussion and advice toward how to evaluate and select a particular database solution that will meet your needs. After having read through these Sections, the reader is encouraged to begin building the business and action plans to coordinate the development process. Appendices A and B include short outlines of the processes for developing these plans.

Section 2: Why should we digitise our collections?

Before computers, natural history collections were the physical databases from which information could be laboriously transcribed (Lane, 1996). Unless the data were then published, transcription made the data available to only one person, the researcher involved. With the advent of computers, and increasing access to data via the Internet, new ways of utilising the potential of your natural history collections has become possible.

Digitisation can have a lot of benefits for your collection, for your staff and workflow and for the potential users of your data. Nevertheless, digitisation incurs a real cost and it is important to understand and address, explicitly, the reasons why you are undergoing a digitisation effort. It may not be obvious to administrators or potential funding sources why the effort and expense are justified unless you give specific reasons. After the project is underway or completed, these reasons can then be used to generate benchmarks and other criteria used to evaluate the efficacy of the digitisation project.

An extensive review of the uses of digitised data can be found in Chapter 1 of this *Manual*, which is based on Chapman (2005c). Some common reasons for embarking on a digitisation project include:

Wider dissemination of data

Primary specimen information is typically restricted to the data embodied on the specimen sheet itself. It can therefore only be made available to whoever currently holds the specimen. Without digitisation, passing data between institutions requires either a personal visit from the interested researcher or the specimen must be loaned out, at a potentially high cost in transport and curatorial activities. Digitised data can be disseminated in many ways, primarily using the Internet, enabling many more people to access and utilise the data.

Enable your data to be studied in different ways

Once you have digitised your collection, you can then query the data in ways that were not easy to do before. For example, you can arrange the data by collector and collection date, allowing you to track the progress of collecting trips. Trying to do this in a collection arranged by family is virtually impossible. Digitised specimen records also play an important part in the estimation of species diversity (Meier and Dikow, 2004; Chapman, 2005c). So long as the relevant data is recorded in a well structured database you have the potential to view it in whatever manner you require.

Enhance curatorial activities

Digitising your collection can aid the day to day activities in your institution, usually by reducing the amount of book keeping required. It can also keep track of the status of the collection by tracking the loan status of a specimen. The quality of the collections is enhanced by identifying inaccuracies; 'lost' specimens may be rediscovered (Peterson, 2002) and standardising the terminology used on the specimen labels. Digitising quickly illustrates the absence of useful data, typically when you are trying to study the data in different ways as mentioned in the previous point. Few other activities will enable you to get to know the true depth of your collection as digitisation will (Peterson, 2002).

Protect your specimens

Digitisation inevitably requires some referencing to the original specimen. Once this is done, there is a reduced requirement to handle specimens, as the specimen data can be transported instead. By reducing handling you will increase the longevity of the original article. This is especially important for irreplaceable items such as type specimens. This does not, however, preclude nor should cause a restriction of access to the specimen, as many forms of research will still require physical examination of the item itself. Digitisation, even if you include an image of the specimen, can only reduce the frequency specimen handling, it cannot replace it. The digital collection also acts as a form of disaster management. Should the worst happen and the original collection be destroyed, the digital collection will continue to provide a valuable resource.

Aid research by reducing future transcription time

Once the specimen data is transcribed it need not be repeated for future projects which also feature the same specimen. This allows later projects to be more efficient, reducing the cost requirements.

Raising institution/collection profile

Institutions are interested in being able to access data from a broader range of sources than merely their own collections. There is also increased pressure to allow greater access to the institutions collections, resulting in improved resources (financial or otherwise) for research projects. Many new projects require online access to the resulting data, pushing forward the creation of digital data. User appreciation of good quality data leads in turn to better appreciation of the original collection (Lane, 1996), which enhances the importance of your collection. Digitising also enables you to monitor the size, growth and usage of your collection, which is very useful when pursuing funding for new projects. Digitisation can also satisfy the CBD requirement to repatriate specimen data to the originating country (Meier and Dikow, 2004).

Enhances the ability of the institution to contribute in areas beyond its traditional remit

Traditionally, natural history institutes have acted to preserve specimens and aid researchers in nomenclatural research. When specimen data is made available, it need not only be used to supply taxonomic researchers, new areas of interest can be catered for. Data could be used in education or to increase the general public understanding of the work done by the institution. The data can also be analysed to identify gaps in the collection and create collection guides to aid future collecting trips. Many more potential uses for the data exist, which can be easily implemented once the digitised data is available.

Legislation

Making data widely available is increasingly required in some countries ‘access to information for publicly funded institutions’ legislation.

Section 3: What You Should Do Before You Get Started

Planning is important

Clear planning is vital to deliver a suitable database. In practical terms, setting up a digitisation effort at any scale is a project in and of itself. Application of formal project management techniques will improve the probability that a given project is successful. It is highly recommended that the implementation of a digitisation program follows the principles of good management; however a thorough discussion of project management techniques lies beyond the scope of this paper. Project management texts are widely available and it is recommended that the reader consult those before initiating a project. In this paper, we will discuss three topics that directly derive from project management. These are the business case, the action plan and risk analysis.

The business case sets out what you wish to do and establishes the benefits you expect to gain from undertaking the suggested work. It also includes an assessment of the resources required to implement the project as well as identifying which resources are currently available. Any shortfall in resources should be clearly identified and the associated costs be stated. Setting these facts out in a single document allows a clear judgement to be made on the feasibility of the project and helps to clarify why limited resources (even if only one person) should be used on digitisation.

The action plan details how the business case should actually be implemented. It contains practical information such as how many computer(s) and database(s) will be required. It also considers the number of staff; training and how the digitisation work will proceed (commonly referred to as workflow). The action plan also details where funding should be sort to cover the resource shortfall.

Risk analysis documentation is a part of the action plan which aims to consider what to do if something goes wrong. Simple examples include what will happen if a computer fails or if funding is not secured for part of the project. Consideration also goes into how to minimise the risk of an event happening. Regularly backing up the data, seeking alternative funding, having a spare computer available are all simple ways of risk mitigation considered in the risk analysis documents.

It is possible that the business case outlines an overall goal that is too large to be completed in any one project and so is broken down into several smaller projects with their own business cases, action plans and risk analyses. This is perfectly acceptable practise and the action plan for the overall business case should then outline the separate projects and how they link together to provide the overall goal. Working in this way allows large and often long term goals to be achieved in small stages without losing sight of the overall vision.

It is important not to rush the planning phase as correcting problems once the database has been released can be both difficult and time consuming. For an institution, the planning phase could easily take six months to a year to implement correctly. This may seem discouraging at first, but taking the time to properly understand your requirements will avoid disappointment when an inappropriate package is rolled out.

It may be the case that a short term solution may need to be in place before the planned database is ready. In this case your plans should include time to migrate the data from the short term database to your permanent solution; otherwise the short term solution may become your de facto permanent database!

Identify your goals

The general principles of why a digitisation project should be undertaken have been identified. While these are good general reasons, it is important to identify exactly why your specific digitisation project should be undertaken. This section aims to raise appropriate questions that every digitisation project needs to answer at some point during its lifetime. Many projects have not considered all of these questions before the project actually began, typically causing substantial additional work when changes to the project have been required. Once the questions in this section have been answered, you should have a much better idea of your projects resources, requirements and restrictions.

Institutional versus individual

Does your project cover the entire institution or is it just a one-man project? Acknowledging the scale of your project defines many of the restrictions you will face when the project goes live. For example, if your project only involves one person, the required workflow, computing and physical space requirements commitments are much smaller than for a full institutional system. Similarly a small project can only digitise a relatively small number of specimens compared to a whole institution. For an institutional project, it is very important to get the staff on board, so careful attention must be applied to the way the project interacts with the day to day work in the institution. Training and education on the project’s aims and procedures are vital; if these processes are not followed from the start the chances of your project being successfully completed are significantly reduced. As the scale of the system increases the underlying database tends to have to become more complicated and more strictly designed, with less room for ad-hoc creation of new fields to respond to the needs of individual researchers.

Who are the principal clients of your solution?

There are many possible users of your data. In terms of the initial data generated from the specimen, users can include taxonomists, managers, researchers, technicians, collectors, environmentalists, non-governmental organisations, pharmacologists and the general public (Chapman, 2005a; see also Chapter 3 of this *Manual*). Inevitably there will be a small group of target users who are your principal audience. Typically for a natural history digitisation project the target audience will be one of the following:

- Individuals on a specific project,
- Researchers generally, and
- Curatorial staff at the institution.

Single person systems are usually very simple to implement and can easily be configured to the unique requirements of that individual. They usually only require a simple form of data entry and a querying system designed to produce a single target result such as a paper, flora or checklist. However, such datasets may be of limited use beyond the initial project unless a specific effort is made to make the dataset widely available. At an institutional level, it is often useful to impose a requirement on internal projects that all recorded data is made available to a central location. This thereby allows others to find previously recorded data and reduces duplication effort across projects.

Researchers may either be external or internal to the institution. Each of these two broad categories of researchers will require some form of interface to be able to access the data. External researchers will typically use a web site to access the information. Internal researchers may also use the same web site the external users have, but you may wish to allow them to access additional information (the location of the specimens in the cupboards would be a simple example). Concentrating on making data available to all rather than individual researchers has two useful consequences at an institutional level. Firstly, the data will need to be standardised across research projects this also helps guarantee long term storage of the data, making the data available to future researchers. Funding bodies often require projects to have some form of information dissemination objective. Planning to make the data available to external researchers will generally serve to fulfil this requirement. One way of making data widely available is to join a data provider such as the Global Biodiversity Information Facility (GBIF). GBIF has its own specialised form of connecting to the database which would need to be implemented.

Digitisation can also be used to assist best practice in the institution, particularly in the areas of loans and accessioning. In order to be able to do this, the staff will need a system that is available within the institution but there is no requirement for a globally available interface.

Of course, it is entirely possible that you will wish to enable many other users to access the database, each with their own special requirements which may include additional information to be captured by your digitisation effort.

Do be aware that the more target audiences you wish the database to serve, the more complicated the database will become. In the case of the clients above, each type of client needs to be able to access the data in a slightly different way, possibly requiring 3 different interfaces (data entry, internal access, external access) to be built. Take the time to discuss specific requirements with several representatives of each client group. This will help ensure that your goals match up with the needs of your target audience.

What language(s) will you support?

The more languages you need to present information in, the more complicated your dataset and interface will inevitably become by a significant margin. For example, the interface alone will have to be presented in more than one language and the data will have to be translated into each separate language. At the very least, your database system should be able to handle unusual (in English) characters such as diacritical marks.

How much data?

While this can be a manageable task for small collections it becomes progressively more difficult as the size of the collection grows and eventually digitising all specimens can become impracticable except as a very long term goal. Targeting specific parts of a collection is frequently a better strategy. This can vary based on the immediate requirements of the institution but typically will focus on easily defined groups such as families or a specific geographic area. One valid technique for digitisation is to take the most important specimens (usually the types) and concentrate the digitisation effort there.

Even if the digitisation effort focuses on one small project it is very important to find out what quantity of specimens are going to be digitised. This number is the primary measure to consider how much time needs to be spent recording the information on the database yet is the aspect that most assumptions are made about when setting up a project. In fact, project estimates may be half the true number of stored specimens. This can cause many problems (not least a shortfall in available resource) which can make a project only partially successful or even fail completely. Even if this is the only thing you do to validate the specimens before the project starts, get to know how many specimens your project will digitise before the project begins.

What data quality?

It is has been said that “quantity has a quality all of its own”. There is always a natural desire to produce as many records as possible in a database. Being able to give the number of records gives an easy metric which can be used to judge the success of a project. However, a simple listing of specimens may not be valuable to most users. Without suitable supporting data it is likely that significant subsequent work would have to be carried out to make the data useful.

Clearly, for each individual specimen, it is more efficient from an institutional perspective to completely process the first time it is digitised. However, there is the question of whether the funding body of a specific project would pay for the recording of information not directly linked to that project. Should the funding body refuse to pay, consideration should be given to matching resources to the shortfall in order to complete the work. If this is not practical, and with the limited funding available to most institutions it may well not be, recording the most commonly required data and the unique data specifically required by the specific project is a reasonable compromise between a complete data record and the limited needs of the current project. The mostly commonly required specimen data can be summarised as the accession number or barcode, collector, collection date, collection location, collection determination and the current determination.

Data capture or data interpretation?

The data recorded on the specimen is typically derived from the original collector’s notebooks and, as with any writing, can be full of errors. The question naturally arises of whether the data should be captured as written (giving a historical perspective to the data), or corrected to give a more current interpretation (possibly correcting spelling errors or updating the country name to reflect political changes since the specimen was taken). Either way is an acceptable practise so long as it is recorded somewhere and the practise is consistently applied across the dataset.

One particular issue with original data is the area of taxonomic interpretation. Invalid names are frequently entered as a determination (a common example of this is citing the wrong author for a species name). In 2004, Meier & Dikow found that 62 – 73% of all determinations of *Euscelidia* were misidentified, so it is clear that this is by no means a small problem.

When recording the data this really leaves two options. From a historical perspective, this data should simply be left unchanged. This does make the data less useful for taxonomic research and so there is strong reason to correct the data. Doing this for every determination can be quite time consuming and so it is recommended that where possible a new determination is made with reference to a published source of names such as International Plants Names Index (IPNI).

Another potential aspect of digitisation is the addition of useful data not actually available on the specimen. Probably the most common example of this is the use of Geographical Information Systems (GIS) to provide location data. To do this specimens have to be geo-referenced (finding the place the collection was made and assigning it a latitude and longitude). On collections from the last 10 years it is usual to find latitude and longitude provided by GPS systems, but earlier collections rarely have this data, hence it must be added. This is a worthwhile endeavour but can be very time consuming as it requires additional research. If all the available collection location data is recorded it may be better to leave such value-added data to be added later. This can have potential advantages as it may allow a geo-referencing expert to concentrate on that one area.

Enhancing existing practices in the institution?

One important potential reason for digitising your specimens is to enhance the curatorial activities in the institution. Often this would require fairly minimal data from the specimen itself (often just the name), but significant additional data referring to the curatorial activities in the institution. A standard feature that is added to aid curatorial work is a unique barcode or accession number used to differentiate different specimens. This enables a user of the database to identify which particular specimen of *Quercus robur* it is out of a collection which may have multiple specimens that cannot otherwise be separated easily.

Imaging

In the vast majority of cases an image is a huge benefit, as it captures information that may not be easily recorded any other way. Occasionally imaging may not be appropriate; algae and bryophytes are particular examples where the value of imaging is debatable, as high levels of magnification are required to differentiate the characteristics of the specimen. Of course, it would be possible to take a high magnification image as well as the image of the specimen label, but this is an additional resource overhead. Imaging does have significant associated costs but these are typically outweighed by the benefits of having a good quality image. It is not the purpose of this paper to discuss imaging in detail as this has already been covered in a previous paper published by ENSCONET (Haüser et al, 2005).

Understand what digitisation will not do

Many digitisation projects fail to achieve their goals simply due to unrealistic expectations being placed on the database. So far, this paper has discussed examples of what a digitisation project can do for you, but it is also important to be clear about what it cannot achieve (McLeod and Winans, 1991). This section is not about what the databasing project will do, it is about ensuring that impossible goals are ruled out.

Databasing is not a money saving option.

Although certain activities can be made more efficient and less expensive, increased access to the data results in more queries to the institution. Introduction of information technology also has a commensurate cost in terms of computer equipment and maintenance (both of the machines and of the digital collection itself). Someone has to actually database the specimens, which requires a short term cost. Careful planning can allow the increased costs to be partially offset by the efficiency savings, but there is likely to be an increase in cost to match the increase in capacity created when you digitise your collection.

Digitising your collection will not create new information for you.

If the information is not already present on the specimen, it will take additional work to locate suitable references to create it. If the data are incorrectly written on the specimen, it will most likely be incorrectly entered into the database, even down to spelling errors in some cases. These deficiencies can prevent the system working as expected, causing the project to fail. Fortunately, databasing your specimens makes it much easier to identify these shortcomings in your data and will enable you to take preventative action. Exploring the uses of the resultant data by comparison with other records can add valuable additional data to that obtained directly from the specimen.

Specimens will still need to be physically stored and handled.

Although requests for individual specimens may fall due to the availability of the digital content, increases in data access will likely result in an increase in requests for specimens. No matter how detailed a specimen image is, there are still physical attributes of a specimen an image cannot hope to record.

When do you want the dataset be to be available?

Most projects are created in response to a perceived shortfall in the data already available, so it is not surprising if "yesterday" is your instinctive response to this question. In practise your goals may require a great deal of resources, particularly if you are working towards an institutional digitisation project. In many cases goals can be broken down into short, medium or long term phases. Chapman (2005a) breaks project goals into the following categories:

- **Short term**. Work that can be completed over a six to twelve month period.
- **Intermediate.** Data entry over approximately an 18 month period.
- **Long term**. Any project lasting longer than 18 months.

Given that many funded projects have attached deadlines it is often practical to map your digitisation goals to the deadlines of the projects at your institution. Naturally, if you are running a small project it is highly likely that your project deadlines are your longest term goal. For all projects though, defining practical short term goals is very useful as it allows you to confirm you are making progress at an appropriate rate. This could be completing a specific subset of your research or supplying data to another institution (very useful for testing the practicality of your chosen method of data exchange).

Although an individual project usually has clearly specified endpoints, an institution usually has to define its goals for a longer period than can be funded by any one project. Setting goals and deadlines is still a useful activity as it helps ensure that new projects fit in with the institutions requirements.

Inevitably, whatever databasing work you do now has the potential to be useful in the future and ideally will be structured to reflect that – otherwise there will be a cost to the institution when it has to re-key the data for a specimen. However, it is difficult to know now what data will be important in 10 or 20 years time, so what do you choose to record? The most future-proof system would be to record everything, but this is also the most resource intensive option. As discussed under data quality, it may be more practical simply to record the most commonly required fields and accept that there will be a future requirement to record additional data.

Future requirements

Once a collection has been digitised, then there is a requirement to maintain the data over time. Quite simply, if the data are not maintained, there is a danger that its relevance and utility will decrease to the point of uselessness (Wheeler, 2004). For specimen collections, the primary example of this change is progress in the area of determining taxonomic names. To avoid this, the data should be maintained in the same way as any other collection. The data collection should be added to as the physical collection is added to, to maintain the relevance of the dataset in curatorial activities. Ideally, this would also include value added information such as geo-referencing data, so improving the usability of the data.

Your institution may have other databases holding additional information related to your specimens, such as seed, cultivation or genomic information, so a natural evolution is to integrate these datasets together. This will enable a better understanding of the total collection holdings of the institution and also enables researchers to access a much wider range of data than was previously available.

Technology is always moving forwards and integration of new techniques will always be an issue. A current example of this would be the use of mobile computing to record field collections. Trying to build new features like this onto a fully developed database is considerably more difficult compared to creating the facility when a database is set up. Considering the functionality you may wish to have in the future, even if they are not to be immediately implemented, will reduce any “growing pains” associated with extending your system in the future.

What are your current limitations and resources?

Having identified your aims, it is now time to consider the practical elements that you have available to deliver your project. It is entirely possible that insufficient resources are currently available to actually implement your desired project, but identifying the shortfall will enable you to decide how you will remedy the situation. This latter action is defined within the action plan. Many elements will be immediately obvious to the project planner, but are covered here for completeness. It may be that some of these limitations will actually be advantages for your project or institution, but for most projects some additional resources will have to be committed to allow you to start your project.

Staffing

You will have several different roles that will have to be filled to help ensure your project will succeed. If a suitable person is available to your project it is a great help, but it is highly likely that some personnel will need specific training, even if it is just how to use the database and how to conform to the specimen handling requirements of your institution. This will naturally take some of the time allocated to your project. Assessing the degree to which you will need additional training will inform the level of time and resources required to complete your goals.

When thinking about staff and staffing costs, don't just think about the digitisers themselves. As with any other project, they will need managing, which takes time and hence has an associated cost. It is also good practise to include some data quality checking to ensure the work is being completed to the appropriate level required by your project. Depending on the size of your project, each of these roles may require a full person or may be incorporated into a single post. A good ratio would be one manager/data checker post per five staff or less. You may also need to consider including a database manager role. You must also decide who will handle the IT issues, even if it is just a case of buying a computer, installing the database and ensuring it all does not break down!

The typical roles associated with projects are:

Digitisers. There are several potential personnel pools you may be able to draw on to get your project done. It is quite likely that as project manager you will use one pool of staff to get the majority of the work done, but don't disregard the opportunity to use any chance you get to add data to your digitisation project.

> **Curatorial staff as part of their regular work.** This is a very useful option if the database is primarily to be used to support curatorial work. It is also very practical when capturing information about loans as they enter or leave the institution. However, the curatorial team already have full time jobs to attend to, so the rate of data acquisition will inevitably be comparatively slow, if steady.
>
> **External contract staff/company.** Passing data to an external team for digitisation can be a risky business. Despite the contactors promises there is no guarantee that the work is being done correctly until the results are returned to the hiring institution. This means that extra care has to be taken both in preparing the target collection before it is sent away for digitisation, and following digitisation when extensive data checks are required. Also, when dealing with external companies, making corrections to the data becomes more expensive, in terms of repeated shipping costs and the potential to be charged for the time required making changes to the specimen. It is rarely a free service, but is frequently cheaper than hiring project staff for the institution and has the added bonus of taking up no office space at the institution other than that of the project manager and data quality checker. However, the remote digitisers are rarely trained in handling natural history collections, so may not be able to interpret complicated data as well as staff in your own institution. If the data are recent and can be easily extracted from a printed file or label then this can be a valid way to go forwards.
>
> **Volunteer staff.** Many institutions have volunteers who wish to get further involved with the ongoing work there. They can seem to be ideal candidates to get to do the digitisation work but this can be a double edged sword. Volunteers are usually eager to get involved but are doing so of their own free will and if they get bored will likely stop their involvement, perhaps without explanation. Sadly, while digitisation is very important, it cannot be described as the most exciting work to be under taken at any institution. Volunteers also expect to work to their own timetables and likely will not want to spend a full working week digitising specimens. Naturally, this means that volunteers will not achieve as much as full time project staff. They also will need office space for them to be able to do their work. For a small scale project, particularly one without great time pressure, volunteers can be a valid approach to digitising your

collection. Do be aware that there is an increased likelihood of staff turnover for your project, which may prevent you from making steady progress as you try to recruit and train new volunteers.

Visiting researchers. Depending on your institution, visiting researchers may be rare or common and they may want to look at any number of collections, making their own notes as they go. There are several issues here. The visitor may not want to look at the collection being worked on; they may not wish to enter the same quality of data that you expect and will almost certainly need it in the form they wish it for their own projects rather than yours. All of this tends to make researchers more suited to institution wide projects and will only be of very limited use to smaller projects.

Project staff. Making use of full time paid project staff is typically the best all round way to get your data entered at a good consistent quality, having the advantages of specially trained staff concentrating specifically on your project. It is also usually the most expensive and resource intensive option as you will have to pay more for trained staff (typically coming from a natural history background). Trained staff do have the advantage of being able to interpret the data themselves and be able to operate more independently than unskilled workers, thus requiring a lower level of management.

Students. At academic institutions, students may be available to do data entry. This can be a relatively inexpensive solution, especially if there are work-study programs involved that subsidise student work. Students may also find digitisation an entry level position that allows them to interact with, learn from, and become involved in the museum community. Using students may also make the project more worthwhile in the eyes of the institution and, in some cases, may earn release time for faculty to work on the project. In some cases, higher level students can be used for more advanced aspects of the project. For example, biology graduate students may be able to help with QC or specimen sorting or review prior to digitisation. Fine Art students may be useful in helping with the imaging process. Turn-over with students is an issue as is boredom and the potential distractions caused by student life.

Data owners. Researchers, other institutions and commercially available datasets can be a big help in providing ready made data, even if it is just the provision of standard values to populate lookup lists. Some care is necessary though, as Intellectual Property Rights (IPR) can become an issue, restricting the ways you are able to use the data.

Data experts. Curators and specialists in various fields will always be required for consultation purposes. It is almost inevitable that someone will be required to answer questions that arise during the project. Trying to divert someone who is not part of the project into helping you can a major hurdle, so getting them to agree to devote time to the project is a major asset.

Technical staff. Technical staff range from the IT support person who comes to setup your computer to the person who designed the database/spreadsheet you will be using. They all need some idea of the importance of your project and will probably have a role in maintaining the systems you use. If you are using a database system larger than Microsoft Access (and perhaps even then), one very important person will be the systems administrator for your database. He is the person who makes sure that the database is up and running, although he is not responsible for the actual data content.

Project management. There are two roles associated with project management. One is the project manager, who is responsible for the day to day running of the project and the second is the data manager, who is responsible for checking the data quality as well as maintaining the data. The data manager is also frequently the person who will continue to be responsible for the data once the project has finished. It may also be useful to have a project champion, a senior figure in the institution whose role is to support the project at an institutional level.

Data entry procedures

You must also consider how many can database at once. The total number of digitisers operating at any one time has several effects, notably the amount of physical space required for the project and the number of practical resources needed, such as the number of computers. It also affects the level of complexity required of the database and potentially of the I.T. infrastructure required for the project.

Options include:

> **One person working at a single database.** The easiest option but also the slowest in respect of project deadlines. For one person, you only need one desk, one computer and sufficient space to lay out the specimens to be digitised. Depending on the level of data security required there may be no requirement for any additional IT infrastructure. There is the risk, though, that if the individual computer fails, the entire dataset accumulated to that point will also be lost, likely causing the project to fail. Some form of data backup up is essential.
>
> **Several people using individual databases.** As per the first option, this is a simple case of multiplying up the resource for one person, although the need to protect the data increases as does the likelihood that one or more of the computers will break down and need to be replaced. There is almost certainly an additional step required, which is to merge the separate databases together to form one single dataset at the end of the project. This means that using several people will reduce the time required to digitise a collection by a factor of the number of people involved (compared to a single person doing the work), plus the time required to set up the project and the time required to unify the data at the end of the project. The risks of losing large amounts of the data are somewhat mitigated by having the final dataset broken up into several parts. However, backing up the separate parts of the data is still recommended to prevent the need to re-record data. This will add an overhead to the project proportional to the number of people involved in digitisation.
>
> **Several people sharing the same database**. This option combines the advantages of several people working on the same project, without the overhead of having to merge the data together at the end of the project. Protecting the data is also much easier as there is only one database to back up. In order to achieve this, some form of IT network will have to be in place, potentially adding to the resource burden of the project.

How big is your collection?

Knowing the total size of the collection will help you to judge the length of time it will take to digitise everything. An accurate assessment is important as it is easy to underestimate the actual size of a collection. It is also important to include a realistic assessment of the rate of acquisition of new specimens, as this must be included in your estimate.

Is access to your data restricted in any way?

As countries have become more aware of the potential wealth accruing from exploiting their native flora and fauna they have become much more protective of dissemination and use of the available information. This has resulted from and in agreements such as the Convention on Biological Diversity (CBD). These require permission to be sought for each country before specimens are collected. These Memoranda of Agreement (MOA) can have many different restrictions, including who may view the data. This may be restricted to one institution or even just one department in an institution. When sharing data, intellectual property rights (IPR) have to be respected and sometimes this requires institutions to sign legal agreements before data can be released. Due to the difficulties of agreeing legal agreements between countries, mostly due to lawyers refusing to sign legally binding agreements based on laws enacted in other countries, many IPR agreements are enacted as Memoranda of Understanding (MOU). No matter if the agreement is legally enforceable or not, these documents must be respected when you release the data. This is also why you should not harvest data from published websites without permission. The originating website may have had permission to publish the data but you may not have permission to use the data as you wish, which means you may be breaking the law. At the very least, you are abusing the trust of the publishing institution by not telling them how you are using large quantities of their data.

You may also wish to unilaterally conceal some of your own data. Rare species, such as those listed on the ICUN Red List, may be a target for commercial exploitation. Revealing data such as geographical location (especially as accurately as a GPS reference) can risk immeasurable damage to the wild populations. Hence you may choose to not to release this data, or only release a very broad definition of the location to protect the native population. There is some argument on this last point that the data are already available in other forms, such as digitised duplicates of the specimen or other collections from that location, but surely anything that makes the destruction or illegal exploitation of native species more difficult is a good thing. Although there is some progress towards a standard approach to sensitive data (see Chapter 6 of this *Manual*, which is based on Chapman and Grafton (2008)), currently the is no agreed-upon international standard, so you must follow the dictates of your conscience when implementing your project, in consultation with your colleagues.

Does your institution require you to use an existing system?

Should your institution already have a central database it is not unreasonable for the institution to expect you to add your data to that which others have already contributed. This can have the advantage of already having useful data for you to use and a ready made data entry system. However, it may restrict the way in which you can interpret and record the data. If this is the case, it may be better to use an individual system to record your project data, but it is vital that the data are provided to the institution at the end of your project in a format that can easily be imported into the central database. This means that any standard reference data requirements the institution requires should be followed in your database. In most cases, having a predefined database actually makes your project easier to deliver, so following institutional standards will normally be the best course.

Do you have legacy data (electronic or paper)?

Pre-existing research data may already exist in the institution, without being incorporated into a centralised database as described in the previous point. This data may enable a large volume of data to be established very quickly, although some time may be required to raise the data quality to an acceptable level or adapt it to your chosen system. Pre-existing paper systems are often easier to digitise (especially if typed) and are a strong candidate for external digitisation as little interpretation of the data is required. It is well worth taking the time with legacy data to check the recorded data against the physical specimens, in order to check both data accuracy and for any annotations that have been made on the specimen since the recorded data had been taken.

What are your physical requirements?

Digitisation does not take place in a vacuum. Staff and volunteers will be involved in your project will need a place to work. Things to consider include:

Where will the digitisation take place? There are three basic options for the location of the project work.

- **Digitise in the collection itself**, which has the advantage of working close to your specimens but typically there will be limited space to work, meaning only a few digitisers can work at any one time. If the collection is popular, the digitisers will be regularly interrupted with the consequence of a reduced work rate.
- **Establishing a dedicated area for digitisation** tends to remove the issues discussed above, at a cost in office space to the institution. However this does allow dedicated and uninterrupted work for a group

of digitisers. Moving the collection specimens to the digitisation area may prove a problem but careful planning can usually solve this.

- **Digitise in an entirely different location.** It is more difficult when the digitisation area is in a different building as moving specimens usually requires particularly careful handling. In this case, imaging the specimens in the collection and then using the image for digitising purposes can be a valid way of solving the transport problems.

All of these have been successfully used in the past to implement projects so long as proper attention is paid to the requirements of the chosen location.

Existing IT infrastructure. Proper attention to your technological requirements should be made. A full sized computer (PC or Macintosh) is better for daily work than a laptop and if already available will affect your choice of database. However, if you are going to be making collecting trips then a laptop is a practical requirement as you can then record observations in the field. Will you be imaging your specimens? If so, a camera or scanner will required. If you wish to connect to the internet or anywhere beyond the confines of your desk some form of network will be necessary, even if it is just a telephone and modem. Consider what you have and what you will need and make a record of them.

Do you already have project deadlines?

If you have already started a project then you will already be working to deadlines that will affect your project. No doubt you will be very conscious of this limitation but recording the deadlines will help underline their importance to others, especially if you need their help to deliver the project.

Will you be working outside your institution?

Whether travelling to other institutions or collecting data in the field, working outside your regular place of business places special requirements on your project. Although you could record specimen data on paper and transcribe them later, it would be far more practical to make use of some form of mobile computing, such as a laptop or possibly a Personal Digital Assistant (PDA). If you wish to digitise specimens directly into your database, it will need to have one (or more) of the following pre-requisites:

be portable or

have a module that can handle data entry which is then imported into the main database or

has a web-based data entry system

The last option is rarely practical when considering fieldwork, as it presumes that you have a reasonable connection to the Internet no matter where you are. Even in today's world of powerful satellite communications it is rarely practical to have a long running connection when working outside of settlements (and sometimes not even then).

Funding

What funding do you already have, or which funding bodies could you realistically seek funding from?

Is there the will to change?

If there is no support for the digitisation work within the institution it will be much harder to implement any project. It is very useful to have a key person to act as a champion of your project, and gathering supporters within the institution will also help. Without such support it will be difficult to run a successful large-scale digitisation project.

Produce a Tentative Business Case

Putting together a business case enables you to clearly set out your goals and limitations in a clear fashion. It should present to others why your project is worth undertaking. One of the most useful steps in getting a digitisation project up and running is to get the rest of your institution to accept and get involved with your project. This often means they will be willing to volunteer their time to look at elements of your project that cover their specialisations, enabling you to improve your data quality. If your curatorial team accepts the importance of your project, the task of getting hold of the appropriate specimens becomes much easier. Having staff willing to contribute resources to completing a project makes its successful completion much more likely and also increases its chances of attracting funding. You should also go beyond the simple statement of goals and limitations to answer practical questions such as:

What do you gain from doing this project?

Explain why it is worth doing for your institution and the rest of the world.

Is your project feasible?

It is important that you believe that your project is practical and achievable. Knowing you have set out all the arguments makes it much easier to do this. Many large scale projects are forced to answer ‘no’ to this question, in which case the goals outlined here should become inspirational targets that need to be broken down into smaller, short-term practical projects that attempt to meet the overall goals of the business case. In these cases, it is suggested that the business case is drawn up to cover the entire group of projects and a more detailed business case is set up for each smaller project.

Do your goals exceed your limitations?

Presuming your project is feasible, you are likely to have to estimate what additional resources you will need to complete your project. A few very lucky projects will have sufficient staff, time and equipment in advance to complete their project, but most will have to find additional resources from somewhere. Outline here what additional resources you believe you will need. Precise details of how this resource is supplied must be included in the action plan.

Rising above your limitations

Having considered your issues and decided that it is appropriate to proceed with your project you can finally outline a course of action, which will be filled out in detail in the action plan (or plans, if you choose to break a project down into several stages). Answering the following questions will help you supplement your existing resources.

> **Can changing working practices free up time to work on your project?** Don’t expect changes in working practises to provide a universal answer to all your issues. There is a reason why people work in the way they do and it is rare that simple changes in practises will release massive resources that can be repurposed to digitisation. However, it might be possible to gain sufficient time to allow a specialist to help with particular parts of your project. Changes in working practise can help in freeing up the needed physical space for a digitisation project. This is one area where gaining wide spread acceptance from the

rest of the organisation is a massive boon, as it is natural for staff to be distrustful of being forced to change successful practises in the face of new demands on their time.

Can other nearby institutions help out? It is entirely possible that you could arrange a joint digitisation effort with nearby institutions, using the resources of larger bodies to assist you (Snow, 2005), or sharing implementation costs with similarly sized institutions.

Who might fund your project? Many funding bodies exist, both on a national and international level. Bodies such as GBIF, the Andrew W. Mellon foundation or the Gordon and Betty Moore foundation all fund a number of projects every year. It may also be possible to gain funding from commercial bodies. It is important to be realistic about your funding opportunities as all funding bodies have many more applications for funding each year than they can actually fund. It is highly advisable to carefully research the requirements of funding bodies in order to shape your proposal accordingly.

Should your project be broken down into distinct units? If you believe your project as a whole is feasible it would not be surprising to say 'no' to this question. However, it is much easier to get a small project funded than a large one, so it is still sensible to give this question careful consideration. Again, exposing your business case to the rest of your organisation will help you quickly assess the viability of your ideas.

Can we do a proof of concept? Running a trial of any digitisation project is a very useful exercise. It will give you lots of practical information to help plan your project. It may not always be possible to do a proof of concept, but it is highly recommended.

The checklist in **Appendix A** should help you summarise your arguments and make writing the business case easier. Allow yourself time to reflect on the business case and seek constructive feedback from others, as it is always useful to question the underlying assumptions of any project. Some of the hardest feedback to take is from colleagues who may not initially support your project, yet in many ways this is the most useful. After all, it is easy to convince those who already believe it is the right thing to do, it is far better (and more satisfying) to convince doubters. Once you have updated the business case to take full account of the opinions of others, you can be reasonably sure you have a viable business case.

Pick a database solution

Now that you have a business case and the agreement of your colleagues, it is time to consider in detail how you intend to reach your goals. Although creating a business case may take considerable time, it is still important not to rush into the implementation phase. Poor decisions made at this stage of your project could still be with you years into the future, so a little time, patience and a willingness to test your decisions will be repaid by a smoothly running project when it finally begins. Do not worry if after implementing this section and section 6, you feel the need to come back and refine your ideas, this is will result in a better project over the long term.

Selecting a suitable database is a complex decision which has to take account of many different factors. These are discussed in depth in the "Deciding on a particular database solution" and it is recommended the user reads that Section when making a decision about the database they will use.

Develop an Action Plan

Having written out a business case and selected a database to use, you will no doubt have many ideas about how you will implement your project. We should now try to prove those ideas will work when your project gets implemented. Simply start by writing down your project ideas in light of the issues raised in the business plan. Then check your ideas against the following points, altering your plans where necessary to take account of any issues raised.

Many of the issues discussed here interact with each other, so it is a practical precaution to go through the list at least twice and assure yourself that changes made as you work through the list still answer the issues raised earlier. Once you have completed this task successfully you can be confident that you have an achievable project that is as robust as you can make it.

Does your chosen solution match your goals, limitations, and resources?

If so, then you are in an excellent position to complete your project. For most though you will require additional resources to match the shortfall between your resources and your solution. Many of the questions here serve to detail that shortfall and to provide a baseline cost for your solution. Once you have that, you can either seek the funding or modify your solution to fit your available resources.

Will your solution handle your future requirements and what if it doesn't?

Ideally of course your solution will leave room to adapt to changing needs. Trying to design for all possible solutions can be very expensive in the short term (even if in the long term it is beneficial) and the costs may easily become prohibitive.

How many staff do you need?

This is somewhat dependant on how quickly you want your collection databased and is limited by the amount of physical space you have. Taking into account your deadlines and the size of the collection you want databased you need to get an idea for the number of staff you will use. It is better to overestimate the number you need rather than underestimate as over estimating making the project run more quickly while underestimating carries the risk of the project failing. A careful balance must be maintained here, if your project becomes too expensive then it will not get funded and your planning will be for nothing. The most effective way of judging the number of staff you need is to undertake a proof of concept, with staff actually digitising a significant number of species for real, using your planned workflow. Be very careful of using the estimated rates promoted by other institutions or database vendors, they are very often trying to make their institution/company look better than it actually is. As a very rough rule of thumb, assume a rate of 100 specimens a week per digitiser if you are going to database in great detail, with high resolution images. If you are not imaging, take a rate of 200 specimens a week and 300 specimens per week if you are neither digitising to a high level nor imaging. These figures are conservative, but you will be thanked for delivering or exceeding your project goals, which these figures should help you achieve. Don't forget to make allowances for staff holidays and absence from work due to illness, as these will affect the overall quantity of specimens you can digitise during the lifespan of your project.

How will you train your workers?

Training workers to use your system will take time and resources. Workers will need to know how to implement your workflow, how to use the computer and database software, and, most importantly, how to translate the specimens into digital information consistently, efficiently, and accurately. It is not enough to get data into the system; it has to be "good" data. This will only happen if the proper people are selected to do the work and that they are afforded proper training in how to do it.

How long will it take to build or implement?

All projects have some degree of lead time, often six months or more simply to find funding. However, buying and setting up computers, getting a working area and recruiting staff all take significant time. Outline what your lead time and add it to the duration of the practical phase of your project. For larger projects using professional staff, two months is the bare minimum time it will take to hire staff and will frequently be longer. This is broken down as follows: 1 week to get an advert written and published, 2 weeks for applications to arrive, sift and arrange interviews. Another week to hold interviews and finally a month for the successful candidate to work out any existing notice. Usually it is possible to get the physical requirements in place during this time, unless you also have to hire someone to complete those tasks, in which case another month is recommended. It is also wise to allow a little time at the end of the project to account for delays and complete any outstanding tasks (such as writing a project report). If you are going to include a website to show your data, the website development will run much more smoothly if it can take place following the main digitisation effort. If you do not do this, you risk unexpected developments in digitisation altering the website work, delaying it beyond the lifespan of your project. Website design can vary massively depending on the IT infrastructure of your institution, but to design a website from nothing, serving up large quantities of data will typically take around four months for one programmer.

What are you going to have to buy?

If you are recruiting new staff, you will most likely need new desks, chairs, computers, telephones and all the accoutrements needed for an office environment. You may need to buy a server to store the dataset on and a link to the outside world. If you are imaging, how are you going to image? Again, try to follow best practise for your specific specimen collection. Hopefully your institution will have much of this ready for you to use, but don't assume this is so before you start your project.

How much is it going to cost?

As you may begin to appreciate, running a digitisation project is not cheap. Make certain you include hidden costs to your institution and don't forget to budget for things like travel and subsistence when working abroad.

What will your workflow be?

In other words, what is the most effective method for your staff to proceed when digitising. The following factors will affect your workflow plan.

Number of digitisers. Larger numbers will allow you to specialise staff into particular roles (imager, data-baser, geo-referencing and quality checker are potential examples). Be careful though as unless the specialisation is that digitiser's chosen vocation, it can cause boredom amongst your staff, which will slow down their rate of work.

Collecting and returning the specimens. This is a job that can typically be batched up so that a day or a week's worth of specimens can be collected at once, depending on the size of the specimens. If you are working in the collection you may even decide to collect each specimen as you need it. Consider how the curatorial staff will find a specimen you have taken away for digitising should they need it while you have it and agree this with the curatorial staff. Also consider if any preparation work must be done to prevent damage to the specimens from insects or poor handling, and factor time for that into your plans.

Specimen handling. It is also a good idea to document proper specimen handling procedures as your digitisers are unlikely to start as curatorial specialists trained in handling your specific collection.

How long will the specimens be absent from the collection? As the specimens may be in demand from other projects, minimising the length of time specimens will be unavailable is to be encouraged.

Location. If your staff is working in locations separate to your main collection, then the difficulty and time taken to transport the specimens to the digitisers must be considered. In this case, it may be better to transport digital images and database from the images.

Data Quality. Quality has to be considered in terms of the data you are going to record. For example, is it checked against recognised standards? Handwriting can be very difficult to decipher, adding a great deal of time to the digitisation process. Chapman (2005b) notes that it is far cheaper to capture data accurately than to correct errors later.

Adding value to the original data. Value can be added to digitised data in several ways. You can compare it to published standards, you can interpret the data to accepted lists (handwritten collectors almost always have to be interpreted this way as their signatures seem designed to confound the digitiser) or you can add valuable data such as latitude and longitude. Doing all of this takes time that should be allowed for, and the specific process recorded.

Imaging. If you are going to image the specimen, how are you going to include that in the workflow? Will it be before or after the specimen is recorded? Will a specialist do this work, or will the task be shared?

Data order. It may seem more efficient to follow the order of data as listed on a specimen label and enter it onto the database, but the order the data fields are presented in the database may not match that on the label and labels rarely have a completely consistent format (consider yourself lucky if yours do!). Experienced digitisers will develop their own preferred way of working, but when training give clear guidance on the most efficient way of entering your specimen data into your database.

Data checking. Always include time to check the quality of the data is up to your planned standards. It is very easy to spoil otherwise high quality work by simple data entry mistakes that can be quickly spotted by simply checking a random sample of data and checking in detail where needed. One way of doing this is to view the data in a table so it may be sorted so similar entries are grouped together, allowing mistakes to be more easily spotted. Redman (1996) noted that an error rate of up to 5% should be expected and it is far easier to correct those mistakes while it is still a simple matter to return to the original data (Chapman, 2005b).

Can procedures be overlapped? It may be that several stages of digitisation can proceed at once. This may be hard for an individual specimen, but working on several specimens at once may be possible. To take a simple example, a specimen could be being scanned while another is being databased, all controlled by the same digitiser.

What effect will staff absence have on your workflow? If you have a key member of staff on holiday or absent through illness, will this stop the rest of the digitisation procedure from happening? This may be unavoidable for a small team, but usually it is possible to make contingency arrangements to circumnavigate the problem.

Are there any bottlenecks in your plan? Careful planning will allow for a smooth process, but be careful to calculate the proper time allowance for each step in a process. Take this hypothetical example: It takes a digitiser 10 minutes to database a specimen, which is then handed over to be geo-referenced, taking another 3 minutes. If this was one digitiser to one georeferencer, there would be no delay in the process, and in fact the georeferencer would be under utilised. If there are four digitisers to each geo-referencing member of staff. This would mean that there is a bottleneck, as the georeferencer can only complete 3 full records before the next batch of four are ready, resulting in a georeferencing backlog of 2 records every 30 minutes. Considering how to handle that backlog in advance can save time when the project is up and running.

As can be seen, there are many things to keep track of when working out the practical process of digitisation. It is recommended that once you have worked out the process details you write it up as a user manual for the digitisation staff, as it will make the staff training much easier.

What happens when the project is over?

Once the digitisation has been completed the data still exists, so your planning should include an outline of what will happen to the data. You could use your database on another project, possibly adding content to your existing dataset. Ensuring the data remains relevant is also important and ideally should be maintained following the end of the project. This could be done by having someone specifically responsible for updating the data or having the data widely accessible in the institution so it can be maintained by the curatorial staff in their day to day work. For every new database, this can take up to 0.3 full time equivalents (FTE) of a curators time (Snow, 2005).

Will your solution provide the appropriate level of data quality?

Decide what level of data quality you will accept. The use of published standards can greatly improve the quality of your data and can make data entry easier. Standards can be used for many areas as simple dropdown lists, but some caution must be observed, especially with older specimens as the terms used on the specimen may no longer agree with the current naming convention, countries are a particular example of this problem. In these cases, an ability to record the original name is very useful, but inevitably adds to the complexity of the database. Many standards exist and sometimes several may cover the same topic. Ultimately though, so long as one standard can be converted to another then it will not matter too much which one you choose.

Planning for the human element

Databasing is not a purely computerised system; it has a human element which must not be neglected. Digitisation requires training, not only in the process as described as above, but in the actual data entry system. Learning to interpret complicated data takes time and proper training. Poor training can be seen as the cause of a significant proportion of data error (Chapman, 2005b). Gaining experience in digitisation also takes time during which your staff will not be working at optimum efficiency, so allow for this when calculating the number of specimens which can realistically be digitised during your project.

How long will it take to database your collection?

There are two basic techniques to data entry. Detailed data entry means entering data carefully, using a maximum of lists, data checking and appropriate structuring of data to maximise accuracy and ease of retrieval. This produces the best data quality but at a high cost in time. Rapid data entry indicates the ability to enter data quickly and easily. However this implies a reduction in data checking and frequently less highly structured data. This can cause an increase in data entry errors and hence a lowering of data quality. Later correction of this data typically requires a high commitment of resources. You do not have to take up either of the two extreme options described above but achieving an acceptable balance between them is not easy and what may be acceptable for an individual project may not be suitable for an institutional system. Specimen quality can vary greatly and so it is difficult to set an average time required to database per specimen. This will also vary depending on the level of data quality that will be recorded and the level of accuracy that will be accepted. Once again, trials are the only practical way of developing realistic estimates of digitisation rates. Quality assurance (QA) is also a vital part of any databasing operation, allowing for error correction at an early stage, hopefully reducing the long term costs to an organisation.

Prioritise efforts

Undoubtedly the best possible result for any digitisation effort would be to digitise all the available data in the entire collection. However, this may take longer to complete and be too resource intensive to do in any one project. In order to maximise the effectiveness of your resources you will have to prioritise your efforts, as discussed in the business case. To briefly recap, you can reduce the total number of specimens you will digitise by targeting particularly important specimens or a specific family or species. You could also reduce the amount of data captured by focussing on key data fields (typically basic collection information and determination information). Of course, you may need to combine both techniques to reach your goals or create a practical project. Here too, long term considerations will have an effect. If it is your intent to make the data available in the long term, then capturing a maximum amount of data per specimen will be more efficient for your institution than maximising the total number of specimens captured, however useful that target is in the short term.

Contingency planning/risk analysis

What happens if your project doesn't go as planned? By considering what might go wrong and preparing for it, you can help to ensure your project will run with a minimum of disruption whatever happens. Risk analysis is a project management tool and the reader is recommended to consult suitable project management texts to decide precisely how you will implement your risk plan. Here are a few simple things you will need to consider:

Staff loss or extended staff absence. How will you respond to the loss in digitisation rate? You may be able to hire new staff, or you may have to change to goals to cover the shortfall. It may be possible to extend the project slightly to allow others to complete the work. It may be politically difficult to ask for this time, but most funding bodies are sympathetic to unavoidable problems so long as the issue *was* unavoidable.

How will you address the issue of not making the required digitisation rate? This is a significant issue. Your project has a number of specimens it must digitise in a fixed length of time (if you don't have this, it is well worth setting a realistic target by which you can measure your performance). This implies a digitisation rate that must be set as a guideline. It will take time for the staff to be trained, so don't expect to hit the rate immediately, and if your targets are realistic, it will be exceeded at peak progress (allowing for normal staff absences). If they don't hit the target, then you should review your work processes to see if there are any avoidable delays, or consider getting additional staff involved. Paid overtime is a possibility if the project budget will run to it. Unpaid overtime is a sign of a badly run project as you clearly needed resources you did not factor into your project proposal. You may hit your targets, but your digitisers are less likely to wish to continue into another project, meaning you have lost a trained resource. If you cannot get more digitisers, reduce the number of specimens you will digitise until you have a realistic target.

What happens if your computer breaks down? At least, get a service agreement to replace your computer, but do be aware this can take time. Ideally, have funding to cover a backup computer, although this will rarely be practical on a small project. Also allow some time in the project to cover these delays.

Backup strategies. Protecting your data from hardware failure is highly recommended as part of any action plan. If your only computer fails and you cannot recover your dataset, your project has failed. This is not something your funding body is going to be sympathetic to, meaning it is unlikely you will be able to attract further funding in the future.

Malicious alteration of data. This is a rare occurrence, but is more likely if using an online system that could be externally hacked. Simple password security helps, which requires some form of database administration. Tracking basic information such as who last edited the data will also assist in the quality assurance process.

What will you do to document your solution/implementation?

Documenting what you do makes working during the project much easier, especially when training new staff. Proper documentation can make your data entry process more consistent among workers. It also enables you to look back and learn from your first project and apply that experience to your future projects. Nevertheless, creating documentation takes time and can, itself, become a bottleneck, if the staff preparing it have other duties in the project. Appropriate planning in advance for your documentation can go a long way to making your workflow go smoothly and ensuring that your timeline is met.

Rating your project

Consider what indicators you will use to judge success or failure of your project. Don't simply consider how many specimens were completed in what time, but also include quality measures. Include a staff morale measure, as an effective project has the potential to yield trained staff for future projects, but if they no longer want to work for your institution you will have lost a valuable resource.

Is your solution a good return on your investment?

As stated at the beginning of this section, you should review your action plan several times to ensure all the various issues you need to resolve work in harmony during the actual implementation of the project. When this phase is completed, consider the project as whole and particularly consider the resource requirements. You must consider if the results will be worth the effort you will put in, as this will likely be one of the criteria a funding body will use. For a well designed project the answer should be a resounding 'yes'. If you are not certain, look again at the amount of work you are trying to do and try to focus more on the most important parts of the collection you intend to digitise, putting the rest off to a later project.

Eventually you will have a well planned project with a strong chance of being funded, allowing you to achieve your goals. When this is all written up, you should be able to run the project as soon as you secure your resources and/or funding. Now all there is to do is put it all together.

Running the project

Finally we have reached the point where you can actually implement your project. It probably seems like a lot of work has already been done without any visible results, but the reward is in a smoothly running project from the very beginning. During this period, there are still a number of things to do even before the project actually begins, but all are practical tasks designed to deliver your project.

Test your assumptions

If this is your first project, you will not know if your assumptions relating to the digitisation workflow and digitisation rates are valid. As discussed in earlier sections, practical proof is the only way to get this knowledge. Doing a small prototype project of maybe a hundred specimens (enough to gain some expertise with the chosen system) will help give valuable insights which can be used to refine your action plan. Having practical expertise with the system also enables you to better train your digitisation staff as you will be much more aware of the issues associated with digitisation. For this reason, it is highly recommended that whoever will manage your project staff undertakes this task. If you are building or altering a database as part of your project, this could form the first phase of data entry, in which case monitor the work very closely and be prepared to alter your workflow early in the project.

Should this alteration be left too late, it is entirely possible that the project will not recover the lost time.

Seek funding

All projects have a cost, whether it is met by the institution or by external bodies. A well developed project may take six to ten weeks to properly prepare (Snow, 2005). Finding suitable funding is something this paper cannot practically discuss in detail, as available bodies vary by country and by the exact nature of the collection being studied. Properly building up the business case and action plan can only enhance your chances of writing a successful project proposal, so hopefully getting funding will be straight forward if you have followed the suggestions in this section.

Build your database

Depending on the database you have selected, you may need to put time aside to have your database altered or even created. It is recommended that a modular design process is adopted (discussed further in later sections), enabling parts of the system to be tested, or even used practically, as the rest of the database is developed.

Hire your staff

Remember it may take a couple of months to get staff to arrive, so this is the first thing to do once you have secured your funding.

Develop documentation

Practical documentation such as a training manual and the design of the database are very important for future work. Suitable documents make it easier to develop experienced staff and maintain the database systems. Take time to develop good documentation before the project properly begins.

Arrange suitable office space

Your staff needs somewhere to work, so ensure it is ready for them when they start. Don’t forget to allow sufficient space for any specialist equipment they may need above and beyond the normal office desks and chairs.

Buy equipment and set it up

Similarly to office space, this needs to be ready when your staff arrive as it is a waste of resources to have them waiting for the equipment to arrive.

Train your staff

No one arrives at an institution immediately ready to digitise. At the very least you will have to train them in your institution’s procedures and probably in using the database. This is where the time previously invested in documenting your procedures will pay off, as it will enable your staff to begin working effectively much sooner.

Start digitising

At last, the main part of your project can begin.

Continually monitor the project

Be aware of the unexpected events, which, hopefully, you have prepared for by undertaking a risk analysis. Even having prepared by undertaking a prototype project, things can go wrong. Using short term goals and reviewing your achievements on a regular basis you can quickly react to unforeseen issues and adapt to overcome them.

Review the project

Once the project is over, whether it was successful or not, review it against the success criteria you laid out in the action plan. Consider what worked well for you and also what could be improved. This allows you to apply the lessons learnt from this project to the next one.

Once you have completed your first project, it is then time to consider what the next project will be. This should be easier having gained experience of the issues involved in a digitisation project, and you probably now have experienced staff and a database to make use of. It is still recommended to take the time to develop or update your business case and create a new action plan, as new techniques and opportunities will become available as time moves on.

Section 4: Organising Information and Representing Data

Which came first, the information or the data? For every definition of 'information' that uses the word data to define it there is one that defines 'data' using the word information. The discussions are as with the egg and the chicken - circular. Losee (1997) discusses some of these at length. Thankfully, it is not so much the definitions that are important in this context but that you have a clear understanding of three concepts:

You know things about your objects of interest, which we will call **object information**.

You will improve and extend the quality of your object information by the use of **reference and ancillary information**.

A computer can store and represent your object, reference, and ancillary information in certain ways. What it uses to do that we will call **data**.

Object Information (as you see it)

The information that you have access to, either already existing in digital format or waiting to be digitised, will fall into one of two categories: primary or secondary information. Primary information is assigned to identify an object or is taken directly from the object (i.e., from the label, tag, or field notes associated with a specimen). These are the things you know which should never change regardless of opinion (as long as it is correct in the first place). Secondary information is used to describe or categorise the object or to associate the object with other types of information. Secondary information reflects what you know at a particular point in time and so will vary over time due to changes in opinion and increased knowledge.

Typical primary object information

Object identifier – The identifier that uniquely separates one single object out. This could be an accession number, barcode, LSID or any other method of assigning a unique value to a specimen and would be of the type *unique identifier*.

Collection event – Made up of collector, collection number and collection dates. These would respectively be of the type *person, identifier* and *date.*

Collection event location - A statement of the locality where the object was collected (*text* type). May or may not include coordinates (e.g. latitude (*GPS* type*)* and longitude (*GPS* type)), land ownership or management (*text* type), and elevation (*numeric* type).

Collection event method – A statement concerning how the object was collected.

Descriptive information about the object– The collector's description of the living specimen (*text string* type). Information about the specimen preparation (*text string* type). Notes or remarks included with the specimen by the collector or the preparer (*text string* type).

Environmental – A description of the characteristics of the locality where the specimen was collected e.g. substrate (*text string* type), vegetation (*text string* type), associated species (*text string* type) & physical geography (*text string* type).

Donor information – The donor person (*person* type) and/or institution (*place* type), contact details (*place* type) and any special terms that might apply to the donation (*text string* type).

References – made up of title, date, journal name, collation, author

Typical secondary object information

Geographical (spatial) – Political geography e.g. country, province, district. Georeference(s) of the collection locality.

Taxonomic and nomenclatural - Includes the collector's initial determination and any number of later determinations or revisions, along with the determiner's name and date determined. Determinations are considered to be *taxon name* types, the determiner is a *person* type and the date is a *date* type.

Storage location - Which institution (*place* type) owns the specimen, its barcode or accession number (*unique identifier* type) and where it is stored (also a *place* type) within the institution. May contain a trail of various owners and storage locations over time.

Molecular – Although rarely recorded at the time of collection, it is increasingly common, and is good practice to associate DNA samples and sequences with voucher specimens. It would be considered to be of the *sequence* type.

Status markers and labels – This is a catch all for information about your object that relates it to other information of interest. Conservation status markers indicate that the object has some designation as rare, endangered, threatened, or sensitive typically through its name (i.e., it is a rare species). Type status designates the specimen as a type according to the rules of a particular nomenclature. Transaction markers can associate a specimen with one or more loans. Other markers may associate the specimen with projects, publications, mark it as part of some subset of the overall collection, designate its fitness for particular uses, or mark it for some particular consideration.

Remarks/comments – This is catch all for remarks about the object or about the information concerning an object that are not part of the object itself. Examples: "The jar seems to be leaking slightly," "Not sure the ID is correct," "The collector name was illegible, but I think this is an A. Smith collection," "I can't find this specimen in the collection – Felisa Jones 5/1/1999".

Recording all of this potential information has a high short-term cost (Armstrong, 1992) but a long-term benefit is that the work never needs to be repeated. Partial recording of the primary information is cheaper in the short term but more expensive in the long term as, when the additional data is required, additional resources must be used to repeat parts of the digitisation workflow. In larger institutions this activity alone can take up a third to one half of the total digitisation process, meaning the net digitisation cost per specimen is significantly increased.

Which of these types of information you handle within your system is determined by your purpose and the level of detail of each will vary accordingly. As a general rule try and make sure that the primary information is recorded as fully as possible and then be selective about the secondary data.

Reference and Ancillary Information

It is unlikely that your digitisation project will be confined exclusively to recording information directly associated with your objects of interest ***and*** that you will enter that information free-hand for each object. **Ancillary information** is all the electronic information that you will manage as part of your digitisation project that is not directly associated with the objects themselves. When ancillary information is used to constrain the values you can enter about your object it can be thought of as **reference information**.

Ancillary information can be as simple as the pair of values you will use to designate presence/absence of a feature, a list of values you might enter as labels for an object (e.g. male, female, hermaphrodite or sterile) or a much more complex set of information like all the fields associated with publications. Ancillary information often involves information about the reference information. For example, if the reference information is the name of institution where an object is held, the ancillary information gives you the name of the curator and her contact information. If the reference information is the Federal Status Designation for a specimen, the ancillary information could include the date that status was published in the Federal Register, its publication reference, and the populations to which it is applicable.

Typical Ancillary Information

Nomenclature

Geography

Morphology

Person Names

Projects

Institutions

Publications

Transactions

Data (as the computer sees it)

The important thing about data is that represents your information correctly and that you can get the same information out that you put it. Specimen information is usually stored in what we will call base units. These base data units are conceptual tools, which allow related data to be stored together and which can be joined in an unlimited numbers of ways to digitally represent your information. Later sections give examples of the different ways in which this is done in reality and your choice of database solution will take this into account.

These are some common data base units:

Person – first name, last name, initials, title

Taxon name – rank, epithet

Place – latitude, longitude, altitude, name, address

Date – day, month, year, century

Sequence

Chromosome number

Reference – Title, collation, year, Publication name

Basic data types

Knowing what data types best data is an important step in building or choosing a database solution. In essence there are three types of data: numbers, characters and dates. The choice of which can impact the way your database works on more levels than you might at first think:

Text/String fields are the simplest of data types and for the novice provide quick and simple data entry. What you see is generally what you get. They allow the user to type in both characters and digits, be aware though that although you are typing a number on the keyboard it will not be handled by the computer in the same way as in a true number field.

- **Pros:**
  Raw data looks like it did when you entered it
  Sorting produces expected results with letters
  Easy to use in table format
  Easy to format data for output
  Copy and paste works as expected

- **Cons:**
  Text fields take up more space than number fields
  Text based lookup tables may be inefficient and slow if strings are long
  You may have to type the same thing over and over again
  Using the data for a purpose other that the one it was entered for usually involves reworking the data and may produce unexpected results
  Sorting can produce unexpected results where numbers are involved or be impossible on very large fields

Parameters:

Length – this indicates how many characters can be accommodated in a field. Setting this to be too small will truncate your data. When importing data to a text field not all systems will tell you that this has happened and even cut and paste can be problematic. Setting it to be too long can have a serious impact on your ability to sort columns in some cases you cannot do this at all.

Padding – some text formats set a field size to a particular length regardless of the amount of actual data in the field. This can seriously inflate the size of your database so you should carefully check this is required.

Memo fields are a form of text field and are mentioned separately because they have some specific limitations. These types of field have no size restriction and therefore allow the user to be very verbose. They are especially good for description fields such as habitat. However, because of their potential size they are rarely sortable and difficult to search. They do not handle text formatting such as bold and italics at a character level and are erratic with regard to carriage return. They should be used sparingly and never for data which is searched often.

Numeric fields are even simpler than text fields, allowing only numbers. However, they are often meaningless on their own and can be more restrictive than text fields. There are 2 types of numeric field: Floating point fields allow decimals, integer fields allow only whole numbers.

- **Pros:**
  Number fields take up less room for storage
  Allow you to use lookup tables efficiently
  Easier to code form based data entry systems
  Make it easier to atomise data

- **Cons:**
  Usually require a text field as well to qualify them so you have to use 2 fields instead of one to handle your data.
  Have to use lookup tables to handle text
  Increased atomisation of data which makes exporting data more complicated
  Data is not easily readable in table format

Parameters:

Length – determines the highest number you can enter into a field. It is usually expressed as the number of bits required to store the number.

Base – numbers do not need to be stored as Base 10, but may be in hex for example (Base 8). The computer and the programmer will love this but the user may find it difficult to interpret. Stick to Base 10 if you can.

Date/Time fields are modified number fields and should be used with some care. Most but not all store an integer which represents a particular date from a known start date which is 1, using this integer you then use the built in formatting to display that date in a variety of ways. However, different packages use different start dates, so you must check that the start dates are the same and that you are transferring the date and NOT the number that represents it when exporting or transferring your data.

Boolean fields are designed to reflect a dichotomous choice such as yes/no, true/false, present/absent, is/isn't. Typically, the underlying field is populated with either a one or a zero and entry is via a label such as the choices given above or via a check box, although in some solutions the actual entered value will be the same as the label. Care must be taken with Boolean fields with respect to default values and null values. For example, suppose you use a field "isSterile" to designate that a given herbarium specimen has no fruits or flowers. If no value is entered into the field, your particular solution may translate this as a zero, indicating that it isn't sterile. Alternatively, it may enter only a 1 or null in which case there is no way to distinguish not sterile from no value entered. The final possibility is that it will store a trichotomy of choices, 1, 0, and null (=nothing entered), although it may be more difficult to clear a field after a yes or no is entered.

BLOB (Binary Large Object) or Container fields allow you to store files such as images, sounds, videos, documents and other binary data within your database. You may have to define in advance the largest file size you can store in a given BLOB field. Of course, storing files within your database can dramatically increase the size of your database file which can affect performance. Some solutions allow links to be stored instead of the files themselves. In this case, the database file doesn't grow so big, but moving either the linked file or the database may cause the link to be broken.

Calculated fields

Often you will want to handle information which is actually calculated from fields which you have measured and have already been entered into your database. As always there are many ways to do this but the things to bear in mind are:

1. how often the formula and its parameters are going to change.
2. how often the measurements that are used to calculate it will change and
3. the final use of the results,

If the formulae and parameters are relatively stable then it may be worth considering creating fields in your database which store the results of a calculation rather than hard-coding (typing in by hand).

If the measurements are subject to change then you need to consider how you are going to keep track of the changes in the database. Often a linking table will help you out.

If the results are going to be widely used and add value to the original data then they are probably worth storing in your database for others to use. If not then you may be as well to do the calculations in a familiar environment and not keep them in the database or do the calculation as part of you final display. It is not wrong to use packages other than your database for calculations.

Functions

Calculating fields and using formulae in your database will bring you into contact with functions and this is where your data type becomes important. In most cases the functions that you can use on a text field cannot be used on a number field and vice versa. Remember that although Date/Time fields appear to behave as text they are really numbers underneath which is why you can add and subtract them. Using functions to write equations and formulae can significantly reduce data entry time but require some investment in terms of coding and working out the logic.

In general if you can code it then do so, it will mean that all entries are always calculated in the same way and you only have to do it once. Of course if you don't have the programming expertise to hand or the time to learn a new language then re-use someone else's code (that's how the real programmers do it). If all else fails enter it by hand at least you've got it.

Special characters and encoding

Special characters include diacriticals, accents, mathematical symbols and non-latin letters. Where data is entered from different original languages these are important, however, inconsistently used they can be confusing. For example are Wurdack and Würdack the same person? If you are entering data from scratch decide *a priori* whether your data entry staff will use them or not and stick to it.

Encoding is a how programmers map special characters. There are different methods and you will need to know which your database solution uses especially if you need to import legacy data.

Both topics are discussed here:

http://www.nada.kth.se/i18n/iab-charsets/terminology.html

Data Views and Storage Formats

Most database solutions use a combination of four ways to present your data to you. The data entry view, the storage format, export formats and multimedia displays. Any of these may be presented to the user as rows, i.e. a table or as a screen of data entry boxes i.e. a form.

Explaining the difference between these things to the users is a useful exercise and will avoid much confusion especially where your data entry personnel are also responsible for the way the data are presented to other audiences but are not programmers themselves.

The data entry view

The data entry view is the place where data is input to your database. Introducing a new data entry system where one exists already (particularly ones created by the users themselves) is often the make or break of a digitisation project. It is not impossible but getting these individuals on board early in development is crucial. There are often very good reasons that an existing system behaves in the 'quirky' way that it does! Don't be dismissive be prepared to listen to the logic they may know something you missed. Only then explain your more elegant solution, twice if necessary. As long it's sensible and doesn't slow them down too much they'll come around. Of course where nothing exists you will need to assess your data-entry staff and select something that maximises efficiency but does not compromise correctness.

The more structured and atomised the underlying database structure the more likely it is that the data entry view will be a form-based solution. Data entry is often slower in these systems but data checking can be very stringent. To increase speed of entry some systems use Rapid Data Entry (RDE) tables which can them be imported but these have less stringent data checking. Which of these you opt for depends on who is doing your data entry.

It is important to realise and/or explain that this view is not really meant for displaying/presenting data and it should be streamlined for data entry. If you are specifying a data entry system you should bear this in mind. Don't try to over-engineer it just to make things look nice. It is more important to be flexible and to allow for as many types of user as possible to get data into the correct parts of your underlying database. You'll never get it a single data entry solution that is right for everyone.

For example: It may be the case that your database is held in SQLserver and have a MSAccess front-end for routine data-entry within the institute and a web form for outside data-entry.

The storage format

The storage format is how the data are stored in the database. If you have a suitable data entry view users will not need to be concerned with this format. However, in general, the more complex the data model the more likely the storage format will make use of link tables, record IDs and objects. In these cases very little can be gleamed from seeing the actual data. Of course for a simple flat data model this format may be the same as the data entry view.

Export formats

These are many and where data will need to be analysed in packages outside your database they can be a crucial criterion in choosing your solution. In these cases the database becomes a repository for primary data rather than a work area.

Multimedia displays

This is where data should be manipulated so as to target different audiences and highlight specific things. It is also the area most often confused with data entry. If your data have been entered well and in the appropriate formats you can display them pretty much as you please so it is important to distinguish clearly between how and where you enter data and what you display.

It is often the case that this part of the digitisation project is left until last with the least resource allocated and time allowed. It is the way an external audience judges your project and is as important as the underlying database itself. If resources are short it is worth considering this as a separate project with its own line of funding and staffing.

Standards

What is a standard?

A standard is a document approved by a recognized body that provides for common and repeated use, rules, guidelines or characteristics for products or related processes and production methods. One of the most common misconceptions is that there is a collections database standard somewhere which will tell you how to build your collections database or information management system. It does not exist. There are many different collections databases and other information management systems in use and they are not underlain by a common, standard data model.

Standards, however, do exist that can affect biodiversity informatics activities, including the design of collections databases and information management systems. For our purposes here, these standards can be grouped into four broad categories:

> **Data Exchange Standards**. These standards, also known as transfer or transport protocols, are used to organise and format information so that it can be exchanged or combined regardless of source. The most commonly known data exchange standards for collections data are the Herbarium Information System and Protocols for Interchange of Data (HISPID) (Conn 2000), Access to Biological Collections Data (ABCD) (http://www.tdwg.org/activities/abcd/) and the Darwin Core (DwC) (http://www.tdwg.org/activities/darwincore/). Exchange standards give the headings, fields, tags, or elements with which to organise your data. ABCD and DwC are both expressed as XML schemas. ABCD has a hierarchical structure and is intended as to be a comprehensive and detailed format to model biological collection information. DwC has a much simpler format and is designed to facilitate the exchange of the "most important" information that might be generally useful.
>
> **Standard datasets** are used to create "controlled vocabularies" for certain kinds of information. These can be extremely useful when used as the basis for lookup lists and reference tables. An example of a standard dataset is Brummitt & Powell's Authors of Plant Names (1992) which is recognized by the International Convention of Botanical Nomenclature as the standard for author abbreviations in plant names. Another is ISO 3166 which is a geographic standard for coding the names of countries and their principal subdivisions (e.g. states and counties/provinces). These codes can be very useful in constraining the values for your geographic administrative units. Saying a dataset is a "standard" doesn't mean that it is necessarily the only choice available for a certain type of information. For example, Federal Information Processing Standard (FIPS) 10-4 has an alternative listing for two letter codes for countries which is different from that of ISO 3166. Saying a dataset is "standard" also does not mean it will fit your needs perfectly or even well. For example, neither FIPS 10-4 nor ISO 3166 has an entry for England, Wales or Scotland which might be a problem if you used these standards for your drop-down list for country as it appears on your herbarium label.
>
> **Best Practice Documents** are guidelines to help standardise methodology and practices and are generally vetted by an organisation or society. Examples include the American Society of Mammalogists' Documentation Standards for Automatic Data Processing in Mammalogy (McLaren et al. 1996) and the

documents produced for the Global Biodiversity Information Facility by Arthur Chapman (Chapman 2005a, 2005b, 2008).

Technical Standards is a catch-all term for standards that do not fit in the previous categories. Typically, technical standards affect the design and implementation of systems that allow the exchange, presentation, and manipulation of data. Software developers use technical standards to build support for the interfaces and encoding into their products and services. For example, the TDWG Access Protocol for Information Retrieval (TAPIR) specifies how to use HTML to transfer XML-based request and responses to access structured data (i.e. data in ABCD or DwC format) stored on any number and type of distributed databases. Another example is the OpenGIS Web Map Service (WMS) Implementation Specification which supports the creation and display of map-like views of information that come from multiple sources.

Standards provide the common language, rules and protocols for the sharing and interpretation of information (Conn 2003). Understanding and using standards can increase the quality of your information system, streamline development, and increase interoperability of your system and information with other systems and information. On the other hand, there are many standards and it may take a high level of expertise to be aware of standards that may be applicable to your situation and to choose which standards are best for your purposes.

Standards Bodies

These are organisations which both develop and maintain standards. Increasingly they are cross-fertilising and looking for ways to link standards together in meaningful ways. There are many international standards bodies and even more that operate at a regional or national scope.

Bodies

Biodiversity Information Standards (TDWG):
http://www.tdwg.org/

OpenGIS Consortium (OGC):
http://www.opengeospatial.org/

International Standards Organisation (ISO):
http://www.iso.org/iso/en/ISOOnline.frontpage

Lists

http://www.consortiuminfo.org/

http://bubl.ac.uk/link/i/internationalstandards.htm

In addition, there are non-standard bodies that have online resources and hold meetings and workshops that serve as useful starting places to help understand standards and their role in our community:

Global Biodiversity Information Facility (GBIF):
http://www.gbif.org

The Society for the Preservation of Natural History Collections (SPNHC):
http://www.sphnc.org

Natural Science Collections Alliance:
http://www.nscalliance.org

And the many taxonomic societies.

Data Quality

What is data quality?

The key to remember it that it is all relative. The terms “*fitness for use*” (Chrisman 1983), *“potential value”* (Dalcin 2004) and “*defect-free”* Redman (2001) have all been used to describe data quality and indeed all of these should be considered as indicators of whether your data is any good or not. In the end though it all boils down to whether you can use your data to do what you want to, whether you can explain what you have to others and whether it can be used by someone else for something completely different.

Chapman (2005a) states that data quality should play a role at every stage of the digitisation process and this is crucial as it will allow you to prevent problems arising in new data and correct things in existing data. A simple way to assess the quality of both the data that you have and the data that you aim to create is to use Redman’s (2001) list to think about its:

accessibility;

accuracy;

completeness;

consistency with other sources;

relevancy;

comprehensiveness;

level of detail and

ease of interpretation.

These qualities are relevant regardless of the size of your digitisation project and so it is important to decide how, given your goals, you are going to address each. It may well be that you have to prioritise them given your working limitations, but they should be accounted for in your action plan.

Entering new data

Of course if you are starting from scratch you only have to work out what data you need to have in order to get the results you want in the time you have, but the purpose of your is project still the most important thing to know and to document.

On way to help create high quality new data is to use lookups, dropdowns and/or controlled vocabularies. These are lists of standardised data/terms from which one or more options may be selected for a particular field in a database. As the data values have already been checked, the use of these lists has the advantage that that data accuracy is improved, although it does not remove the possibility that the wrong option is chosen. Hierarchical lookups which filter themselves is another useful way to make entry more accurate. If you can use lookups then do, it will reduce the time spent error checking. There are, however, limitations and drawbacks to using standardized datasets in look-ups that should be considered at the outset of your project. First, they must be obtained, formatted, and perhaps augmented before data entry starts. This may impact your ability to start digitisation in a timely fashion. Secondly, these standard datasets may change or be updated after you have imported it into your system. It is not necessarily a straightforward process incorporating and reconciling these changes with the rest of your existing data.

An alternative is to use lookup lists in a less strict fashion, still allowing data entry for trusted users and restricting others.

The use of anything but simple lookups will, in the majority of database solutions, increase the programming overheads and may complicate the database structure itself.

Importing existing data

In the majority of cases you will have existing digital datasets (legacy data) which you wish to re-purpose, incorporate, merge or build upon to help you to achieve your goals. In other cases, your workflow may entail creation of datasets outside your system and subsequent ingestion. Whether you intend to move these data to a new system or add functionality to an existing database system, the first and most important thing you need to do is work out what you have and why it was created. It is rare that a dataset can be transferred from one system to another without some work being done to it. The first principle of practical data quality assessment is 'purpose.' Once you know what you have and how combining your datasets benefits your aims for this digitisation project you can decide what extra information you will need to achieve your goals.

Remember that legacy systems have rules all of their own and just because data is held in database software does not mean that it functions as a database. Time spent assessing each table and field to determine both its purpose and actual content will save you time in the long run. You will inevitably have to 'clean-up' legacy data and the time taken to do this task should not be underestimated.

Maletic and Marcus (2000) define data cleaning as:

- Define and determine error types
- Search and identify error instances
- Correct the errors
- Document error instances and error types
- Modify data entry procedures to reduce incidence of similar errors in future.

This is discussed in more depth in Chapter 4 of this *Manual,* which is based on *"Principles and Methods of Data Cleaning"* (Chapman, 2005b), but some things to look out for are:

Field names

Field names can be misleading in many ways.

Case 1: different disciplines use the same terms to describe completely different concepts. The term 'valid' in a zoological names dataset does NOT mean the same as in a botanical one. So despite being used correctly in both cases they are not equivalent.

Case 2: the contents of a field may bear no relation to the field name at all. This may have occurred because there was nowhere else in the system to enter the required data or because the user did not understand the field name.

Case 3: a field may change its use. A field labelled 'date' may have originally been intended to hold the collection date for a specimen but a second user thought it was the determination date. So, while the data itself looks fine there are actually two different bits of information.

Column/field order

This should not matter in a well-designed database; however, not all databases are well designed.

Example: There are 2 fields in a spreadsheet one to record determination date and one to record type validation date. You know what they are because the field to the left says Det by and Validated by respectively. However, they are both labelled ‘date’.

Because spreadsheets use cell references to identify data elements and columns this is perfectly valid. However, you will experience problems when you import them to a database package because the field names will conflict. At best it will prompt you and give the option to rename at worst it give the field an auto-name. It is better to give the columns distinctive names prior to import.

Rows vs. records

In a database table, rows represent records and each record represents a unique instance of something (e.g. specimens, people, publications, etc.). Each record is comprised of number of fields which exist whether or not they are populated with data. Each record has the same data or potential data. Data that come in from text documents or spreadsheets are not necessarily organized this way. The data may be hierarchically organized with headers that are repeated only once for each instance.

Case 1: Rows in a database table have been used to hold dividing label information. In the table below records one and five represent headers in the original document that have now been parsed erroneously over the first several fields in the database.

ID	Barcode	Collector	Coll num	Location	Name	Det date	Country	Coll date	User
1	The	Adam	Smith	Collection	1960-9				
2	98987	Smith,A.	90	BM	S. aph.	2/6/1969	Ecuador	Mar 1969	yy
3	98988	Blogg,B	1	KEW	B. perr.	5/12/2006	UK	Jun 1908	xxx
4	98989	Anon.	306 (I think this is Smith, A)	NY	E. sup.	8/8/1971	France	Sep 1701	xxx
5	The	Richard	Spruce	Collection					
6	10001	Spruce,R.	5040	BM	M.aus.	1/1855	Ecuador	12/1852	Yy
7	10002	Spruce,R.	5041	BM	M.apr.	1/1855	Ecuador	10/1851	yy

In a text document or spreadsheet, information may be repeated inconsistently from one record to the next.

Case 2: The row for record 4 had a comment inserted after the collection number.

Formatting and data types

In a database, data in a given field all have the same data type. This is not the case when importing from a text document or spreadsheet. For example, a date column in the original document might be mostly populated with data in a mm/dd/yyyy format, but occasionally have a cell with a value like “1/11-13/2001” or “Spring 196?” These will not come into your database properly. Even worse, something like “May 2001” may represent formatting on an underlying date of 05/01/2001. This will import correctly, but does not accurately reflect the true collection date information.

Never assume that the contents of a field or similar fields are formatted consistently. Unless they were entered from an un-editable lookup list they won’t be and even if there is a lookup list it’s not guaranteed. Do NOT underestimate how long it will take to reformat all the values in a dataset that need it. Automatic parsing scripts will only get you so far, be pragmatic about how long it takes to develop them, at some point you will have to do some by hand.

Combining datasets

You may have more than one dataset with different original purposes in different systems and even different formats. While it is perfectly valid and often desirable to combine them you MUST be aware that with different original focuses the limitations of one dataset may negate the usefulness of the other. Although the quality of an entire combined dataset is not as low as the worst single dataset contained within it, be aware that it is not as high as the best. Data quality may in fact be better preserved by linking datasets together rather that merging them and it may also be the case that merging is not as simple as first impressions suggest.

Example 1:

Dataset 1: is a specimen dataset created by population ecologists to look for genetic drift in a particular species complex.

Dataset 2: is an accessions register.

A field in the 1st dataset is called ‘Sequence?’ and uses a ‘y’ to indicate that there is a DNA sequence to go with this voucher specimen. Null values indicate that there is not.

The 2nd dataset does not have this field at all.

If you combined to two datasets without altering the 1st one to use ‘n’ to explicitly state that a sequence does not exist, a null value could have one of two possible meanings and the user would not be able to tell. The quality of this field is now reduced.

Example 2:

Dataset 1: A taxonomic treatment of *Corallina*

Dataset 2: A Coralline algae type catalogue

The 1st dataset records information about a particular taxon, and also the specimen which has been designated as its type. A field called ‘taxon’ is used to store the basionym of the taxa in question.

The 2nd dataset records all the specimens in the herbarium which are currently in a type folder and uses a field called ‘taxon’ to store the current name of the specimen.

Here the same field name has, quite correctly within the context of the table itself, been used to record two very different pieces of information and the rows are not analogous. Merging of these datasets while possible would not be straightforward.

Language

The default language used in a database should be appropriate to the database entry staff and to the primary users. It may be that more than one language may have to be catered for in the database. Apart from the use of common terms in a common language such as Latin, this is a huge complication. Not only do the data have to be recorded separately in each language, suitable procedures have to be put in place to display the appropriate version of the data. Maintenance of the data is also made more complicated, as two versions of the record must be updated. Automatic translation of the data can be attempted, but these are not always accurate and hence reduce the reliability of your data.

Maintaining multiple languages in a database can be done but has maintenance issues which, may out weigh the data–entry usefulness. At the very least, the use of proper encoding is imperative if any of the languages are non-Latin.

Intellectual Property Rights

The issue of intellectual property is vast, complex and outside the scope of this section.

However, it is an issue that every digitisation project should be aware of, and which, to as great an extent as possible, it should address. IPR will impact your project both in terms of using information and data to create the digital resource, and also in how that resource is disseminated to the target audience. Be aware that even though you and/or your institute may have access to and regularly use a dataset for research you may not actually have the right to publish it in its native form (either on the internet or on paper) for a purpose other than that for which is was originally given.

As a general rule always try to find out what the original source of the dataset is and document what you did to the best of your knowledge, always get permission to use it and always acknowledge the source. Check your institute's guidelines and any local legislation, as rules differ from country to country.

These are places to start:

GBIF

http://www.gbif.org/News/NEWS1174645079

USA

http://usinfo.state.gov/products/pubs/intelprp/index.htm

UK

http://customs.hmrc.gov.uk/channelsPortalWebApp/channelsPortalWebApp.portal?_nfpb=true&_pageLabel=pageLibrary_ShowContent&id=HMCE_CL_000244&propertyType=document

Section 5:The Data Model

Introduction

If you read the last section, you have perhaps started the process of identifying the primary and secondary object information that you are interested in recording and the ancillary information you are interested in maintaining as part of your digitisation project. You have some idea of how the computer is going to see your data and how you want your system to present the data for different uses. Now you need to address how that information is going to be organized and handled as data by a computer system. This is the topic of the data model.

We will start our discussion of the data model with respect to a simple case, the catalogue, and introduce the first organising concept, the **base unit** of interest. We will then move on to a more complicated situation in which both the base unit and the **focus** of the data model becomes important considerations in the model design. These points will allow us to better understand the next subject which is the more complex, modular design of information management systems.

The data model is the foundation for an implementation in which the data are entered, viewed, manipulated and made available to others. We discuss some basic concepts in data model implementation including the structure of the implementation system and the importance of distinguishing between what the model can do and what the implementation itself can do to make your data usable. We then address some key concepts involved in how the complexity of the data model impacts the way information is handled as data in an implemented information management system.

The Base Unit of a Simple Data Model

Let's start by considering the most simple data model that you might realistically use as you digitise a collection: a catalogue. A **catalogue** is a representation of a collection of objects as a list. A catalogue differs from a simple list in that it contains descriptive information about the objects in the list. To exemplify the role of the base unit, let us do a thought experiment in which the objects of interest are the arthropods in an invertebrate collection. Your goal, as curator of this collection, is simply to generate a catalogue of your holdings so that you and others will know what you have. Seems simple and straightforward so far right? Even in this simple example, however, there are a number of different base units you might choose and different ways of organising your data model.

The individual as a base unit

In the first case, we will choose the individual pinned insect as the base unit. To have your electronic catalogue be an accurate representation of your collection, each record and each insect ***must*** have a unique identifier. These identifiers fulfil two functions. First, they match each record in your database uniquely with each specimen and, secondly, they identify each specimen as distinct from all others. Note that in this data model all of the information associated with the specimen except the identifier can be considered descriptive and, to some degree optional with respect to the data model. A specimen can be catalogued, for example, so long as it has a catalogue number even if it doesn't have a taxonomic name associated with it.

The base unit in degenerate data models

In the second case, the collection consists mainly of pinned insects, but you also have collections of spiders in vials of alcohol and your thrips and other small invertebrates are on slides. The vials and slides can have hundreds or thousands of individuals in or on them and it is impractical if not impossible to assign an identifier to each individual. Instead you will assign unique identifiers to each pin, vial or slide. Your logical base unit now is not the individual, but the "preparations" in your collection be they pins, slides or vials. This data model is **degenerate** in the sense that each of your base unit preparations may in fact represent a collection of objects of interest. The degenerate model has implications for the information you record and its interpretation. The taxonomic name associated with a slide, for example, may represent the identity of each individual on the slide or it may be at a higher rank that reflects the lowest rank that all the individuals have in common (i.e. they are all members of the same family). You might have aggregate or summary fields such as the count of individuals or a list of the sexes and developmental stages represented in the vial.

Of course, the greatest conflict in a degenerate model is when you want to associate information with an individual or a subset of individual within your collective base unit. For example, you wish to publish a type based explicitly on only two of the spiders in a vial or you do DNA sampling on only one fish in a jar of fishes, but how do you record this information in your database and how do you represent these events with your preparations? The answer is not simple. It may require use of note fields or linked tables. You may need to use subunit markers like gill tags or it may be necessary to promote a subunit. So for example, take out the two spiders and put them in a separate vial with a new unique identifier.

Suppose, your collection is very large and your goal is simply to know what you have as fast as possible. You do not have tags on each of your specimens and for your purposes it would take too much time and effort to label each specimen. In this case, you might be tempted to create a catalogue based on taxonomic name. Your catalogue might contain a list of unique names as your base unit with the count of specimens or preparations for each name as the main descriptive information you record. Another alternative would be to use the drawer as your base unit. Your collection consists of a number of drawers in a number of cabinets, so why not just record each drawer, its cabinet location and the taxonomic identity of the specimens in each drawer and perhaps the number of specimens or preparations in each drawer? If it fits your purposes, this may be the way to go.

The problem with degenerate models is that, by definition, they exclude information about potential subunits. It not necessarily “wrong” to use a degenerate model if the information is unnecessary for your purposes,. Keep in mind, however, that a degenerate model may not easily be convertible to a model based on another base unit if, in the future, you decide your existing solution does not meet your needs. For example, if you developed a system based on drawer location and counts per drawer, converting to an individual or preparation based system might take virtually the same time and resources as starting the new system from scratch.

The Focus of a Data Model

In data models more complex than catalogues, one must consider both the base units employed in the model and the focus of the model. Let us use the example from the last section of a collection of pinned specimens in an arthropod collection. In a specimen based model, the base unit is the specimen (or preparation) and it is also the focus of the database. Other information such as the taxonomic name and collection information are added as attributes of the base unit specimen. But we know that, particularly for arthropod collections, when one organism is collected a large number of other organisms are collected at the same time in the same collecting event. So perhaps it would be more efficient to have the focus of our data model be the collection event itself. In this case, we give each collection event a unique identifier and associate a lot of information with that event like who was there and when and where it happened and why. We could then associate the specimens, the “what was collected” as conceptual attributes of the collection event.

The difference between these two models is profound. For example, in the first case, each of the specimens has a unique identifier, but when the focus is the collection event, this need not be the case. The collection event could be linked to a number of uniquely identified specimens or it could be linked to degenerate specimen information (1,200 thrips, 200 cockroaches, etc.) or it could posses some combination of a descriptive list of what was collected and individually identified specimens.

Note that in our simple case of a specimen based catalogue of arthropods in a collection, we still are likely to enter information about the collection event and vice-versa for the collection event based system. The difference is how that information is converted into data in the data model. In the first case, the collection event information might be entered free-hand for each specimen or it could be cut and pasted from one record to the next if they were collected during the same event. But in either case we do not have sufficient information in the system to tell us much about a given collection event. Doing a grouping on the collection event information doesn't help much either, since the information might have been entered somewhat differently for each specimen record. On the other hand, since our focus in this case is the specimen, we don't have the burden of having to enter much about the collection event in order to populate our specimen records.

In terms of data model, then, the next step more complicated than a simple catalogue is a system with an identified base unit, a single focus, and which contains additional descriptive information about this base unit. This descriptive information may include one or more lists, but these lists are treated as multiple-value attributes of the focus information. The focus is used to make data entry efficient and directed towards the purposes of your digitisation project. The focus dictates the priority you will give to recording and maintaining information in your system. The focus also determines what kind of output you are likely to generate from your system.

The following are some examples of common information products and purposes in our discipline and their relationship to the information they use and the foci of the information systems by which they are generated.

Floras and Faunas

These are **taxon based** products. The focus is the taxa within some superset of higher ranked taxonomy and implied geographical extent (e.g. the flora of North America or the mammals of Queensland):

> *Taxonomic and nomenclatural information,* providing the currently accepted taxon names, synonyms and discussion of the application of names to members of this fauna or flora.
>
> *Geographical (spatial) information*, providing the distribution information for each taxon.
>
> *Publishing information*, providing the reference information concerning each taxon or each taxon name.
>
> *Descriptive Data* for each taxon.
>
> *Remarks and comments* of the author or cited authorities.
>
> *Images* representative or illustrative of each taxon.
>
> *Voucher Information* including specimens that were observed by the author, collected by the author in support of the flora or fauna or which otherwise document a taxon's inclusion in this flora or fauna.
>
> *Taxonomic keys*, providing an example of tertiary information that is not used elsewhere.
>
> Maps

Presence Checklists

A database of checklists is **location based**. The focus is geographic areas and the taxa that are in some way documented to be present there.

Geographical (spatial) information, providing the distribution information for the demarcated areas of interest.

Taxonomic and nomenclatural information, providing the taxon information itself.

Publishing information, providing the reference information for an observation or other documentation of a taxon's presence at a location.

Status markers and descriptors for occurrence (e.g. migrating, permanent populations, historical or ephemeral; rare, occasional, common).

Remarks and comments of the author or cited authorities.

Summary information such as species richness.

Status Checklists

A list of taxa or populations of taxa based on their conservation status or, conversely, a list of taxa or populations with their status according to one or more authority.

Taxonomic and nomenclatural information, providing the taxon information itself.

Publishing information, providing the reference information for the application of a particular status to a particular taxon or population.

Conservation status markers and descriptors (e.g. rare, endangered, or threatened).

Remarks and comments of the author or cited authorities.

Collection Notebooks

Typically **Collection Event based** or based directly on **collection number**.

Collector and collection event, noting who did the collection, when and the identifier for the collection.

Geographical (spatial) information, providing the broad location details of the point of collection.

Taxonomic and nomenclatural information, providing the collectors determination.

Environmental information, giving information of the specific locale the collection was taken.

Descriptive information may also be included

Images of the collector notebook pages

Specimen Catalogues

Specimen based, like a herbarium catalogue

Collector and collection event, noting who did the collection, when and the identifier for the collection.

Geographical (spatial) information, providing the broad location details of the point of collection.

Taxonomic and nomenclatural information, providing the collectors determination and any additional determinations since then.

Environmental information, giving information of the specific locale the collection was taken.

Descriptive information may also be included as required.

Donor information. The specimen may not have come to your institution directly following a collection event, and so details of the donating person or institution will be required.

Images of the specimens

Transaction Documentation

The focus is the **transaction event** by which objects are obtained, moved, loaned or exchanged.

Object Identifier, indicating the specific object(s) involved in the transaction.

Collection Management, noting who is involved in the transaction, their contact information and the terms of the transaction.

Taxonomic and nomenclatural information, providing the current names and any new names returned with the specimens.

Donor information, the specimen may not have come to your institution directly following a collection event, and so details of the donating person or institution may be required.

Limitations, explanation of the restrictions placed upon the specimen.

Publication information, providing the document or reference for a document that was generated based on the transacted specimen(s).

Sightings and Observations

Observation, such as bird sightings

Observer and observation event, noting who did the observation, when and the identifier for the observation.

Geographical (spatial) information, providing the broad location details of the point of collection.

Descriptive Data may also be included as required.

Taxonomic and nomenclatural data, providing the observer's determination.

Environmental data is also helpful, particularly the latitude and longitude.

Images that document the observation

Publication information, providing the document or reference for a document that formed the basis of the observation.

Project Documentation

Project based

Collector and collection event, noting who did the collection, when and the identifier for the collection.

Geographical (spatial) information, providing the broad location details of the point of collection.

Taxonomic and nomenclatural information, providing the collectors determination and any additional determinations since then.

Environmental information, giving information of the specific locale the collection was taken.

Descriptive information may also be included as required.

Process and procedural information, details of the way the project curates its specimens.

Limitations, explanation of restrictions placed upon the specimen as part of the project.

Authorization, details about permits obtained to allow collection as part of the project.

Other project related data.

Publication information, providing the document reference(s) or the document(s) that were produced as part of the project.

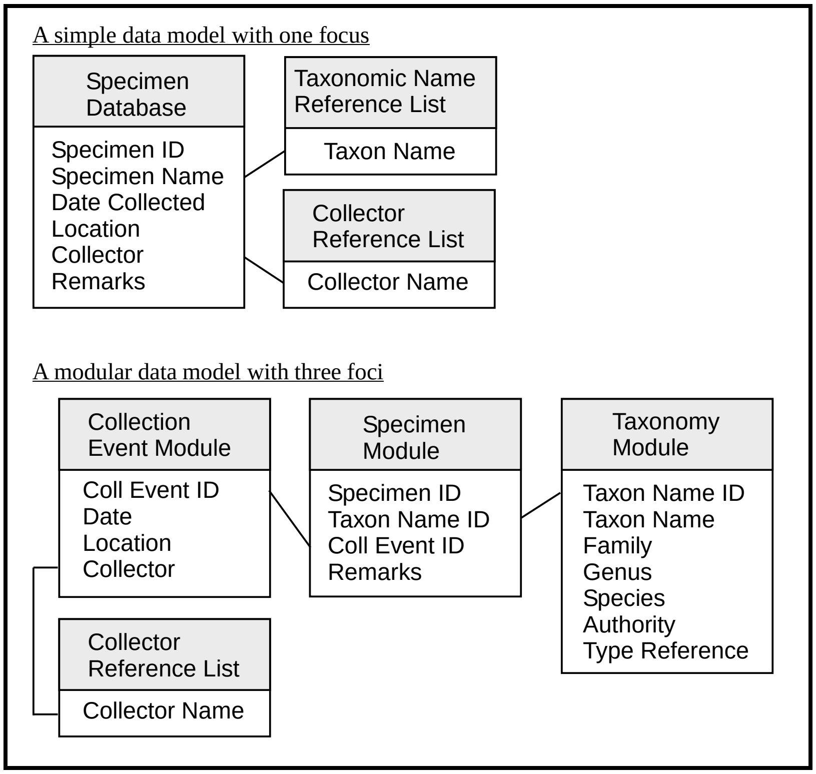
Though there are other possibilities not listed above, your ***main*** reason for undertaking a digitisation project likely falls under one of these categories. It is just as likely, however, that what you really want is a system to handle many of these different activities. You want a catalogue of your holdings, you also want to support projects, manage transactions, keep track of your collector notebook information, etc. How are you going to make that happen?

Some of these activities may be handled by simply extending a simple data model as given before for a catalogue. For example, a few fields can be added to mark whether a specimen is on loan and to whom. You can add a field for conservation status or for the project name under which the collection was made. But in all but the simplest cases, what you will need to do is develop an integrated **information management system** that allows you to organise your data and your views of the data in different ways for different purposes. You will need a data model with a modular design.

Modular Design of the Data Model:
The Information Management System

An **information management system (IMS)** allows you look at your stored information with more than one focus. For example, in a given IMS you can focus on a specific specimen record and gain access to detailed information about the collection event from which that specimen was obtained or you can alternatively focus on a collection event record and navigate to detailed information about each of the specimens collected in associate with the event. The design of an IMS is necessarily modular. In this example case, we have one module for specimens with the specimen or preparation as its base unit and a distinct separate module for collection events which has its own base unit (the collection event). The IMS allows you to focus on either the specimen or the collection event and treat the information from the other module as a type of attribute of that focus module.

In an IMS each module record consists of a base unit, in most implementations a unique identifier for each base unit, and additional information about that base unit (attributes). The data for a given attribute might be entered directly into a field in that module, or it might be selected from a reference list, or from another module. The unique identifier becomes more important in systems with modular design as can be seen in the example below for the collection event module. The base unit for a collection event is the unique combination of date, location, and collector. The only data contained in the module which reflects this base unit is the collection event ID.



An IMS with a modular data model can be a very powerful tool to help you in curation tasks and information delivery. It can also be very time consuming to develop a sophisticated data model . It may seem obvious, but with a more detailed data model there is more data to be managed and more data interactions. For example, how will you manage a nomenclature module where some of the information, the list of names, for instance, is imported from outside sources, but some of it is added in-house? How would you import an updated list of names and match it to your specimens and other in-house nomenclatural information?

Implementing a Data Model

There are two basic approaches to the design of the data model. They are the flat file approach, such as a spreadsheet or the more complicated relational database model. The advantages and disadvantages of each are considered here.

Flat file/spreadsheets

Flat file/spreadsheets have the advantage of simplicity. All you have to do is define the field names (Family, genus, collector, collector number etc.) and you can then get started with your data entry. Rapid setup time is a real advantage offered by flat file designs. Also, it is very easy to add an additional field if something has been forgotten. If data are missing you are not forced to enter any data into a particular field, which may not be the case with the relational database system. Thus data entry is usually faster using a spreadsheet, at the cost of data quality, which may cause delays in the overall process as corrections are made to the data.

This does have several disadvantages though. There is typically no form of data validation and all data is entered as text (even numbers!), which can make querying difficult. If a field requires several values to be entered (such as a determination history), this is very difficult to achieve as the data must all be placed in one field, which makes the data difficult to search. It is also more difficult to study the data as it cannot be easily queried or filtered to present only the pertinent information.

In a spreadsheet it is possible to work around these issues to some extent and introduce some simple design constraints such as drop down lists and formatting fields to hold specific types of data. However, it is still relatively easy for the user to circumnavigate these restrictions if they so desire. It is also the case that maintenance of the drop-down lists will become an issue. In short, the more complicated the data you wish to capture, the more appropriate it is to use a relational database.

Relational Database Management Systems (RDBMS)

Relational Database Management Systems (RDBMS) can be created in a similar manner as a simple spreadsheet. You create a table in which you define your fields by name and also by data type. You can also define which fields must have data in them, without which the record will be rejected. Once this is done, you can then open the table and start entering data. This method has most of the same issues as the spreadsheet approach, but allows a slightly greater degree of control of the entered data, at a slight cost in setup time.

However this is not all that a RDBMS system can do. Lookup lists can be created and maintained via the use of additional tables which are linked together via fields known as keys. Keys are defined in two parts. Firstly there is the *primary key*, a unique value, usually a sequential number, assigned to the value you are wishing to define. This is then used in your main table as a *foreign key*, which is simply a field designed to hold the primary key of the data you are looking for. This is where the 'relational' part of the RDBMS comes in, as the primary and foreign keys create relationships between the various tables. In a similar way, the problem of multi-valued fields can be solved as a separate table with its own fields, linked together by key fields.

Using RDBMS system it is also possible to query the data in much more sophisticated ways, filtering the data to more easily find, and where necessary update, pertinent information. This makes an RDBMS approach much more powerful than the spreadsheet.

The price for this additional power is an increase in complexity of the system. It is not possible in this paper to dip more than the merest tip of a toe into the very deep pool that is RDBMS systems, so the reader is strongly advised to consult your IT staff or study database design for your system before getting started. This naturally adds considerably to the setup time required for a project but the increase in data accuracy is often worthwhile. Also, the more complicated the RDBMS becomes the more difficult it is to represent the information in a user-friendly fashion. Although Microsoft Access is capable of integrating a lookup into the users view of a table, many systems are not and designing a suitable user interface then becomes an additional requirement.

Object Orientated and Object Relational databases

Object Orientated and Object Relational databases are slowly becoming more available. It is still to say that these have similar strengths and weaknesses to the RDBMS options and should be considered in the same way as relational databases.

Data model solutions vs. programmatic tools.

It should be understood that the data model is simply a representation of the way the data are stored. It should reflect the data that you require for your project but may not reflect the way the data is actually entered or maintained. You will require some additional tools, such as a find and replace system, to keep your data up-to-date and to ensure the data is as up to date as possible. For the simple spreadsheet approaches these are already built into the system you are using. To some extent this is also true for RDBMS systems, but it is usually better to add a user-friendly front end to the system. This will generally contain forms for data entry, querying and retrieval. It will also typically include tools to calculate field information and may present join together data so it is presented in a format that is easily readable. It is important that someone in your organisation has a clear understanding of the way the system works and which tools are required for the different tasks of creating and maintaining your data. Ideally, this will be written down! This person will act as the administrator or expert for your system and will help resolve difficulties should you encounter any.

Complexity

As noted in the discussion on how complex the data model should be, it is necessary to consider how data models grow in complexity as more details and refinements develop.

Multiple value fields

It is not uncommon for a single field to represent several separate entities. Collector is an excellent example. A single field could hold the primary collector or it may hold a group of collectors such as:

- T. Wajima, S. Yoshizawa & T. Kitayama

This is not particularly user friendly for the purposes of searching, so you may wish to split these out into separate values, hence:

- T. Wajima,
- S. Yoshizawa
- T. Kitayama

If the number of values is small, then it is possible to hold each of these in a separate field in a flat file, calling the fields “collector 1”, “collector 2”, “collector 3”. However, this soon becomes unwieldy and requires many fields which may not be used for the vast majority of specimens – if one record requires a “collector 10” field, then all records have a “collector 10” field, even if they never fill the data in. It is better to have the data in a separate table, where only as many collectors as required need be stored. These are linked by primary and foreign keys (the foreign key being stored in the collectors table). Two records are illustrated in the following diagram, showing the first record having three collectors and the second record has only one, co-incidentally the same as one of the collectors from the previous specimen.

Specimen table	
Specimen	**Name**

SpecimenCollectors table		
Coll.	**Specimen**	**Collector name**

key	
1	Codium latum
2	Mastocarpus yendoi

key	key	
1	1	T. Wajima
2	1	S. Yoshizawa
3	1	T. Kitayama
4	2	T. Kitayama

Atomisation

Collector’s names also neatly illustrate the issue of grouping names together. Taking our example of T. Kitayama, it could also be written as:

- T. Kitayama
- Kitayama, T.
- Kitayama
- Kitayama; T.
- T. Kitayaama

Or many other variations on the same name. This makes it impossible to present the data consistently as more variations of the name are entered. In order to solve this problem the name can be split into its various parts in a process known as atomisation, in this case initials and surname, giving the following:

Collector table		
Collector Id	**Initials**	**Surname**
1	T.	Kitayama

This information can then be manipulated to form calculated fields in a consistent manner as follows:

- “T.” + “ ” + “Kitayama”
- “Kitayama” + “; ” + “T.”

This enables the data to be presented consistently in many different formats. This does take some extra time to enter, but improves the accuracy of the data and enforces a common format for all the projects data. Such atomisation should be applied as appropriate for your needs and your workflow. Notice however this does not solve all of the problems illustrated in the initial example, as spelling errors cannot be rectified by atomisation alone. This would have to be solved by the use of lookup lists and data checking.

Normalisation

Look back at the collectors table at the top of the page. You will notice that “T. Kitayama” is repeated twice in the table. Having to do this for multiple records increases the chance that a spelling error, as seen in the atomisation example will occur. To reduce the chances of this happening, the data model undergoes a process known as normalisation. Put simply, this takes fields which have common data that may be repeated many times and places them into a separate table, which typically becomes a lookup table of accepted values. It is possible to normalise data repeatedly until it becomes redundant (for example, it is possible to separate out collectors initials into a separate list, but there is nothing to be gained from doing so), so it is recommended that normalisation is also applied with a healthy dose of realism. Normalisation allows data to be maintained more easily, as changing one value in one table can correct errors in many records. It is also possible to remove spelling errors by correcting the line, or by altering the records to point to the correct entry in the lookup table.

This is applied in practise once again via the use of keys. For a single valued field this is straight forward. In the following diagram, there are two specimens where the type is defined from a list in the type table. In this case both specimens are defined as isotypes.

Type table	
Type key	**Name**
1	Type
2	Isotype
3	Kleptotype

Specimen table		
Specimen key	**Type key**	**Name**
1	2	Codium latum
2	2	Mastocarpus yendoi

In the case of the collectors example the same principle applies. Extract out the collectors to form a table of its own, using the keys to link the two table together, providing exactly the same results as the first example.

Specimen table	
Specimen key	**Name**
1	Codium latum
2	Mastocarpus yendoi

SpecimenCollectors table		
Sp_Coll. key	**Specimen key**	**Collector key**
1	1	1
2	1	2
3	1	3
4	2	3

Collectors table	
Collector key	**Collector Name**
1	T. Wajima
2	S. Yoshizawa
3	T. Kitayama

This gives a table that is simply a list of numbers, which may seem difficult to comprehend and in truth this is the case. It is at this level of complication that the use of forms becomes necessary to hide this complication, presenting the data in a user friendly format.

It is also true that the more separated out the data is, the more difficult it is to query, as the various table must be joined together correctly to create appropriate results. The difficulties of joining tables together correctly is beyond the scope of this paper and the user is strongly recommended to study the details of their database solution before starting into this level of detailed design. Finally, the more the data is spread out over many tables, the longer it will take your computer to process the information and present it to you.

Combining multiple field values, normalisation and atomisation.

Putting the previous examples together gives three tables which look like this:

Specimen table	
Specimen key	**Name**
1	Codium latum
2	Mastocarpus yendoi

SpecimenCollectors table		
Sp_Coll. key	**Specimen key**	**Collector key**
1	1	1
2	1	2
3	1	3
4	2	3

Collectors table		
Collector Id	**Initials**	**Surname**
1	T.	Wajima
2	S.	Yoshizawa
3	T.	Kitayama

As can be seen this is quite complex, however, it does illustrate the point that the more complex the system is, the more you are relying on someone who understands the details of what the system is doing and how, when it is presented to the user there is a great deal going on behind the scenes to make the database work correctly.

Section 6: Deciding on a particular database solution

In this section, we only briefly look at the various areas you should consider when selecting a database. It is highly recommended that you look at the preceding sections before selecting a database for yourself. Whatever you do, ensure your database fits your institutions IT capabilities. Picking a database you cannot actually use is probably the quickest way to guarantee your project will fail! If you do need a database with extensive IT infrastructure requirements, you will need to include the resources to extend your IT systems, which will make your project significantly more expensive.

Whichever database design you choose, don't forget the human element. One of the most important, if not the most important, aspects of the database system you select is the user interface. If it is difficult to enter data, then data entry will inevitably be slower and your staff will be less happy in their working environment. The easier a database is to use the more widely accepted it will be and the faster data entry will proceed. This is a very difficult aspect of data entry systems to evaluate without proper testing with your real specimens and is one of the reasons why practical testing of your chosen database is considered so very important.

There are many collections databases already on the market (Berenson et al, 2003), at a wide range of scales from individual collections to full scale multi-collection institutional databases. Fortunately, databases also come at a wide range of prices and it is likely that you will find something that will fit your project funding. However, most commercial databases are also quite generalised, so you may need to consider ways of adapting the database to fit your particular needs. Should a database be close to but not exactly match your requirements, it is often possible to hire a contractor to further modify the system to reflect your needs. It may even be that you need to create a completely new database, which may not be the fastest or cheapest option, but it does give you the ability to specify exactly what you want.

When you do select your design, don't forget how complicated real world data can be. A simple example is the recording of the date a collection is made. Viewed naively, a date is a simple thing with a day, a month and a year. Collectors though seem to have a perverse wish to make things complicated and may only record the month and year, only the year, or a range of dates. Even if the date is recorded as 5/9/1815, does this mean the fifth of September 1815 or the ninth of May 1815? Reference to the collector may make this clear but that does require additional research. Dealing with data is discussed in Section 4.

The design you elect to use is referred to as the data model. For many institutions the underlying data model will depend on the commercial database package selected. If you are designing your own database, then you will be able to specify your own data model. Data models play a role in how data is disseminated outside of the institution's database, so it is highly important. Data models are discussed in Section 5. It may also play a role in migrating data into the institutional database, if this is a requirement on your project. Many data models exist, although the current emphasis is on data exchange schema such as ABCD[1] and Darwin Core[2]. If you are designing your own system it may help to gather your own ideas together by looking at these schemas.

Dissemination of data is another issue. Some commercial database packages will have mechanisms to allow external (typically Internet) access to your data. Does this fit your requirements, or will you need to create a separate application? Separate applications do have the advantage of being designed precisely for your own requirements but equally have additional costs associated with them. Allowing access to the outside world also brings its own risks, especially if you are intending to make use of data entry online. Internet attacks by so-called hackers are a significant risk when exposing your data (Morris, 2005) and appropriate measures should be taken to guard against such attempts.

[1] See http://www.tdwg.org/standards/id/81/

[2] See http://digir.sourceforge.net/schema/conceptual/darwin/core/2.0/darwincoreWithDiGIRv1.3.xsd

Once you have selected a candidate database, you can then begin to consider how you can practically implement your project, which is discussed in the following section.

Which database(s) should you use?

Put simply, you can either get an existing software package from somewhere else or you can build it yourself. There are advantages to either of these approaches as well as drawbacks. Keep in mind that no solution will be perfect. There are things it will do well, things you wish it did better, and things it won't do at all. However, with appropriate planning you should be able to obtain a solution that meets your needs reasonably well and that represents a good investment of your time, money, and resources.

Existing Packages

Existing packages can either be commercial or open source. If you have the resources to pay for it, you may find that the upfront and possibly continuing expenses of a commercial package represent an appropriate expense to get you up and going rapidly. Much or all of the design work has been done for you already. The same may be true for an open source package which may have little or no cost. In either case, it is important to look for access to technical support and documentation. How active is development; what is implemented now, and what is planned for the future? How often are new versions released and what will be involved with migrating to the next version of the program? Who else is using the program and how likely is it that the program will continue to be supported in the future? But most importantly, does it do what you want it to do and expect it to do?

One of the main advantages of using an existing package is that you should be able to evaluate it prior to a full commitment to using it. Try it out if you can or at least study the demonstrations and documentation fully before committing to use it. Keep in mind that the representatives of a given package are unlikely to focus on the limitations, shortcomings, and flaws of their software. As much as possible, you need to discover these yourself during your evaluation process. Do not assume that the package will do things it is not documented to do and ensure that the functions your are interested in actually work the way you would expect them to.

It is extremely unlikely that an existing package fits exactly what you need for your situation regardless of the sales pitch. It may be too simple and may not be able to support the full range activities you expect from it. Alternatively, it may be too complex, requiring more expertise and resources to manage and maintain than you have available. It may have features you will never need, but which you will nevertheless need to maintain and interact with just to use the program. Or it may simply have the wrong focus. You could use a package that is primarily designed to database photograph collections to database your specimen collection, but is this probably isn't the best way to go.

Existing packages should always be evaluated with respect to flexibility, customisation and ad hoc solutions. At one extreme, a package may represent an abstracted system for holding a data model with built in capabilities for views, reports, searches, etc. To use this system, you will have to do a substantial amount of development and possibly programming and its use may represent little savings over designing the whole system *de novo*. On the other hand, a system may be so rigid that it simply cannot be modified to meet the particulars of your situation or may only be modified with substantial additional cost that were not foreseen upfront. This may be particularly true for commercial systems that are highly proprietary.

Between these two extremes is support for ad-hoc customisation. Many packages, for example, have placeholder fields that you can name and use for data that were not expected in the original design. Such fields, however, may not be as easy to enter into, search, control entry, export, or even view. Furthermore, these changes may not be compatible with the underlying or intended data model, making their interpretation less than obvious.

It is beyond the scope of this paper to critique existing packages or even to give a detailed listing of them. GBIF commissioned a survey of existing publicly distributed collection management and data capture software solutions (Berendsohn et al 2003) and this may be a good place to start. Other starting points include:

- **TDWG Subgroup on Biological Collection Data: Software for Biological Collection Management:** http://www.bgbm.org/TDWG/acc/Software.htm
- **GBIF Links to Software and Tools:** http://www.gbif.org/links/tools
- **Digital Taxonomy: A Web Resource for Open Source Biodiversity Informatics:** http://digitaltaxonomy.infobio.net
- **A searchable list of herbaria that are databasing their collections:** http://www.cals.ncsu.edu/plantbiology/ncsc/type_links.htm

There are many more programs in use than are currently listed in any of these sites, although it is possible that in the future they may be more complete. As you explore packages that may be of use for your situation, you would do well to both check with existing users of any package you are interested in and also check with other collections that may be similar to yours to see what they are using.

Building Your Own Solution

If your databasing needs are relatively simple or your resources scant, it might be worthwhile to build your own solution. A collection of just a few hundred or even a few thousand specimens can be catalogued in a flat file or spreadsheet solution that is built in less than a day. Some simple preparation, consideration of quality control issues, a look at the ABCD and/or DwC schemas (see Standards in the previous section) and you could rapidly be on your way to building a system that could be of great use to you and potential users of your data. Building your own solution could also be a good starting point. If done carefully and thoughtfully, you will likely to be able to migrate your data later into an existing software package or scale up your solution with subsequent modifications should your needs or access to resources change in the future.

Alternatively, you may have the needs, resources, and access to expertise to build a more sophisticated solution. Building it yourself will give you the flexibility to tailor your solution better to your specific needs. If you have the computer and programming expertise and the time to do it, then truly building it yourself from scratch may be the best way to get just what you want. But you should be careful that it is easy to underestimate the amount of time and effort it may require to design and build an elaborate information system. Another solution would be to customise an existing package to meet your needs. If you can get hold of an open source package that it sufficiently close to what you want and which allows sufficient customisation to let you do what you want, then this may also be a satisfying route to follow.

In most cases, “building it yourself” really means getting someone or a staff to build your database or information management system for you. Depending on your situation, this can range from as simple as getting a motivated and sufficiently proficient graduate student to build it to more substantial investment, either by contracting with a commercial firm with existing staff and resources to design and implement your solution or by hiring such staff yourself. Alternatively, your institution may have existing IT staff that you can commission or task with the development and building process. There are two important considerations to keep in mind if you follow this route. First, the quality of what you get may depend largely on your ability to communicate what it is you want and need. If you can't articulate what you need from an information management system or don't have the time to convey this to the developers, you will be hard pressed to find the suitable expertise to build what you want anyway. Secondly, it is very important that your solution developers appreciate the nature of biological data and museum collections data. Database programming expertise is simply not enough. Time and time again, I have seen solutions built that work well in theory, but then collapse when faced with the reality of the data we work with and the real world work flow their solution is supposed to enhance. When this happens, strict database developers will often try to convince you to change the data to fit the data system they have built. In other cases, they will try to "solve" the data problem with elegant database solutions that invariably are not so perfect as they had intended or, more likely, are never finished, leaving you with nothing but wasted time, effort and money.

Morris (2005) discusses many of the issues involved with relational database design and implementation in the specific context of biodiversity informatics. This is certainly a worthwhile document to consult during your planning phase and also to recommend to your development staff.

One of the advantages to a build it yourself solution is that the result should be more transparent and easily modified than if you were to use a commercial product. This is not always the case. Depending on your expertise and the manner in which it is developed, a home-grown solution can be just as much a black box as any other solution. Furthermore, the question of how it is maintained and upgraded as necessary over time needs to be considered in advance. Will it still work when a new operating system comes along? Is the source code available and accessible or just the compiled version? Finally, what if any effort will go into generating documentation? Should there be no documentation or little for your database, you should strongly consider how new workers will be trained in its use should there be turn-over in the experienced staff.

What are characteristics of a good database solution?

Whether you are looking at commercial packages, open source software, or considering building your own solution, it is important to understand the criteria with which to evaluate proposed systems. Ultimately, a "good" solution is one that both matches your needs and which works well. The following list of questions is intended to help develop your criteria as you evaluate which particular solution might work for you.

What does it really cost?

Cost includes the initial price for development, hardware cost, licenses, maintenance, upgrades, and additional software requirements. Cost also includes access to the expertise to keep the system in working order.

Is it stable?

A good database solution should be stable in three ways. First, it should work for most of your purposes now with some, but an overwhelming number of features on the "coming soon" list. Be very clear about what the proposed software really does versus what it could do. Be very clear about how modifications and upgrades will be implemented, particularly with respect to how they impact workflow. Also evaluate how changed or added features will be tested. Will they be pretested with realistic sample data or will you have to find out if they work while you are trying to enter real data and hope for the best?

Secondly, is the software bug free? How hard is it to crash the program? Of particular interest is determining if or how likely it is to corrupt the data if the program crashes. If you do find bugs in the program down the road, how likely is it that you can get them fixed? Are the data backed up as part of the package or will you have to have an external back up system?

Finally, what is likelihood that your program will be supported over future changes in computer architecture, operating system, database program, networking protocols, and programming language? As rapidly as technology is changing, it is realistic to assume that your database solution will have a finite lifespan in its current form. Hopefully, this will be measured in years, but it is possible that some solutions will be out of date in a matter of months or even out of date at the time they are built.

Does it have good documentation and/or technical support?

Even the most powerful, elaborate information management system can be rendered useless if you can't figure out how to use it. Documentation should cover both what the solution can do and how to use it. Are tutorials included? Technical support should include issues with setting it up, how to use it, and dealing with and reporting potential bugs. Make sure to look into how technical support is accessed, how available it is, response time, and how much it costs initially and after the introduction period. Look at your needs for documentation and support for the whole package from the computer hardware to the backend database, the front end, and the implementation. You should also be able to view the data model in some form or another.

Does it have the performance you expect?

Few things are more frustrating than having to wait endlessly while the computer cranks through even the simplest function. There are many reasons for slow performance. It could be that your computer or server is slow (low processing speed), the hard drive is full, or your memory used up. Or it could be that your computer is plenty fast enough, but the program is still just plain slow. Poor programming can lead to slow performance because some routines are inherently less efficient than others. Some solutions may have background activities, such as record change tracking or validation processes, that slow down performance. Some implementations may need to periodic maintenance events, like manually rebuilding the indices or "vacuuming" the database, to keep performance from degrading. If a solution allows or requires network connections as part of its functioning, you need to also address the reliability and speed of your network while determining if a particular solution will work for you.

Some performance issues are scale dependent on the amount of data in the system. A particular solution may work great, for example, when you have only 2,000 records in it, but when you get to 200,000 records, you find that searches, imports, reports, or exports take forever.

What is the learning curve?

Few programs are going to be usable straight out of the box. You will have to have training and/or access to documentation to use it properly. In general, the more functions a program has, the longer the learning curve is to use it. But there are other factors that affect the learning curve. Some designs are more intuitive than others. How navigation and functionality rely on buttons, menu items and keystroke commands will affect the learning curve. How easy is it to search for or view data? Do you have to know SQL to perform searches or write scripts to generate reports? How similar is the functionality of a given program to programs you are already familiar with?

Long learning curves can lead to frustration even when the documentation is good and the proposed program is ultimately a reasonable solution to your needs. One of the hardest parts of evaluating a proposed solution is determining whether it is worth the time and effort necessary to learn to use it, even if it is a "good" program.

How do you initially populate the system with data?

You need to think, up front, about how all the necessary data will get into your system. Generally, our main focus will be information about specimens, but it could also be collection events, projects, publications, nomenclature or any combination of these or other types of information. Which if any of these types of information need to be entered before other types of information can be entered? For example, do you need a list of collector names or taxonomic names before you can enter a specimen? If so, where are you going to get these data? Are some data supplied with the system? Which kinds of information does the system treat as reference (i.e. static lists of values for a drop-down list) and which are modifiable and if so, by who and how? Which data can you get from existing sources and which do you need to compile yourself?

How do you enter, edit, view and delete data?

Look carefully at how these four functions occur throughout the system. Are they separated, for example, such that entering data takes place in one way, but editing existing records occurs differently or in a different view of the data? Who can delete data and how easy is it to do by accident? How are links maintained between modules when information is entered, changed, or deleted in one? Which information is required to be entered for a valid record and what happens if you have incomplete information? Are there shortcuts to help enter data that remains constant over many specimens? How easy is it to view data in unique combinations?

Can you navigate easily around the program?

A sophisticated relational database with multiple views of information and modules for various functionalities can be very powerful if you can use it. Ease of navigation is an important feature of usability. Data entry screens should allow for intuitive tab order and generally easy flow from field to field. Typically, there are more fields for entry or view for a given record than can fit on a single screen, at least with a readable size font, so look at how you navigate to see more complete information for a record. How do you move between views of a single record and lists of multiple records? How do you move from one module to another when the modules are linked or when they are distinct? For example, how do you access more information about a taxonomic name, a collector, a location or a collection event when you are looking at specimen records? How do you navigate between data entry functions to label generation or transaction recording modules? If the database has simple, intuitive, or even sensible navigation it will improve your day-to-day satisfaction with the program. Clunky or seemingly random navigation can, at best, increase the learning curve substantially and at worse be a constant headache that reduces productivity substantially.

How does the solution improve data quality?

For almost any solution more sophisticated than a spreadsheet you will be looking for features that improve the quality of your records. This can entail drop-down choices for some fields, calculated fields that allow data to be used in different ways without re-entry, and field-level validation to ensure that entered data meet some minimal expectations for that field.

The alternative side to this is that if the program expects higher data quality than exists, can you enter lower quality data? For example, perhaps the collection date field prevents you from entering 10/32/1964, but what if the collection date is “summer 1964;” will you be able to enter this or will it force you into some ad hoc solution? Ideally, a good program walks a fine balance and will disallow or call attention to data entry errors, but also allow over-ride or alternative entry methods for lower quality data.

Another aspect of data quality involves a built-in validation process and/or support for a quality control process. Can lower quality data be identified and retrieved for post-data entry review? Can records be marked as having been subject to administrative or expert review? Can an administrator associate particular data entry issues with particular data entry personnel? Can an administrator determine when or even if a record has been changed and determine quickly which information was involved in the change?

What kind of importing functions does it have?

Programs can differ widely with respect to expectations that data will be entered directly into the database or whether information does or can come from an import from some external source. If you have substantial legacy data how will it get into the program, how will it or must it be formatted prior to import? Is import addressed only as an initial function of setting up the new solution or will the program easily support imports of new data in the future. Some programs go even further, expecting data entry to occur outside the program into a format that is then imported. Some provide stand-alone data entry modules that can be distributed to collectors. In any of these situations, it is important to review where data quality issues are addressed for imported or potentially importable data. Will you have to address them prior to import and, if so, will you have any tools to help evaluate the data quality outside your main program? Will you have to address them during import using import logs or error reports to key you into issues with the data? Once in the system will you have markers to distinguish records as belonging to a particular import set? Are there tools to improve their quality once in the system or even remove them if errors are too great?

What kind of exporting and reporting functions does it have?

Exporting and reporting functions can also vary widely among programs. Pre-packaged reports may be included to facilitate some commonly expected output, but there should be some capability to output unique reports. Look at how reports may be formatted for printing on paper if this is important to you or, alternatively, if functions that allow print-outs can be used to generate electronic output as well. Look at the support for alternative types of output including pdf, html, xml and choice of encoding. Can you export your data in UTF-8 or ISO-8859-1 or just ACII text? Or can you even tell what encoding you are using during export?

Support for web-based access to data is important for many database users today. At one extreme all or most interaction with the database may be through a web interface, from data entry to view access for the public. All-in-one solutions may have a substantial impact with respect to security and performance. Alternatively, the main database may interact with or export data in a format suitable for web presentation on a separate server. While this solves some problems, it is important to address the additional requirements that will be involved in developing and maintaining the web presence.

There is a lot of interest today in allowing automated or semi-automated harvesting of data such that your database can act as a node in a confederacy of like-minded databases. At the bare minimum, this entails the needed capability to export into a known schema such as ABCD or Darwin Core. Most likely it will also involve support for a data portal such as DiGiR or TAPIR and the necessary scripts to maintain a current, refreshed and compatible view of your data. Some programs may package such capabilities with their program or provide support for developing these extensions yourself.

Exporting and reporting data from a relational database can become increasingly complicated when the underlying data model is complex. This can be manifested in slow export; preparing a flat-file Darwin Core export from some programs can take as much as 24 hours of processing time even when the number of records is relatively low. It can also mean generating a report or export, on the fly, is simply a dauntingly difficult task to do. Canned reports or exports are helpful, but make sure you become aware of the specifics involved with generating a novel report or export and any tools, scripts, languages, etc. that will be needed to carry it out.

What support for networking and multiple access does a solution have?

Some solutions may reside on a single computer allowing only a single local user at a time, although this is becoming less common. More sophisticated solutions keep track of multiple users or can give different classes of users different rights. With larger projects, look for support for multiple, simultaneous access so that more than one person can be entering data at the same time or, for example, so that your collection management can be using it to process loans from her office, while data entry is occurring elsewhere. It is generally a bad idea to have multiple copies of the database in circulation such that one is the “real copy” while others are distributed for others to do searches and reports and other functions.

Can you, and if so, how do you customise it?

To one degree or another, unless you build your database completely yourself, your solution is going to be something of a “black box.” There are going to be some features that you simply do not understand how they work (but hopefully they do work!). There are going to be some things you are not going to understand why they work the way they do, but probably some trade-off is involved. If X worked the way you wanted it to, maybe Y wouldn’t work so well or at all. A big part of the evaluation period and the learning curve is finding out what you have to get used to in order to use a program properly and which things are actual deficiencies of the program.

In any case, at some point changing needs or a growing understanding of your needs is going to entail some customisation of the way a program works. It may need slight tweaking to meet the needs of your particular workflow. It may need drop-down values that are compatible with your legacy protocols. It may need whole new modules of functionality to be added when more resources are available to add them. To the degree possible, you should determine in advance how flexible a proposed solution will be to such changes and what built-in features it has to allow customization. Can fields be reorganized on a particular view? Can you alter the tab order among fields? Can you create new views for certain tasks, and if so how? Can you modify the underlying data model or do you just change the way you interact with a relatively static and immutable model?

Custom built or open source programs may have the opposite problem, that it is too easy to inadvertently cause a change that disrupts the program function. Can a data entry person change a script underlying global navigation or delete a view?

Does it have the right focus and features for you?

A program can work beautifully, but if it isn't right for you and your needs, it is not going to be a good solution for you. Following the discussion in Section 5, a prospective program should focus on what is important to you. If a program is designed to handle any type of museum objects including artwork, architecture, and anthropological items, maybe it is not right for your collection of pinned insects. If it doesn't allow appropriate customization, you could find yourself spending most of your data entry time navigating around fields that have no relevance to the information you wish to record. Or worse, you spend all your time putting your information into a data model that doesn't allow you to access and retrieve it in the way you expect and need to in order to fulfil your mission.

Maybe you've found a program that allows you to record detailed information about your specimens and collection, but what you really need is something to keep track of all the details of your various projects and the literature you produce in conjunction with these projects. If the program doesn't focus on projects as you would like, you may find all of your efforts coming short of the outcome you expect.

As you evaluate existing solutions or prepare to build or have built your own solution, enumerate the data you are interested in entering, maintaining and having in your system. What is the focus of your information system or will it have multiple foci? For your objects of interest (see Section 4), what is your primary object information and what is secondary? What information is ancillary to your primary object information and what will be used as reference information?

There are many features that an advanced information management system may have or that you may be interested in them having. Some common features and data issues of interest to our community are discussed below. Depending on the nature, complexity, and focus of a system, these may be handled in different ways which are not necessarily better or worse than another way. It is important that you address, during evaluation or system development planning, the specifics of how these will be handled.

Handling nomenclatural information

Nomenclatural information is incredibly difficult to translate into data in a database system. The taxonomic name is, in one sense, just a label or attribute for the specimen; as in, X is a specimen of "Y" taxon. Even this is subject to interpretation, however, as there is the name as given on the label, which may or may not be spelled right and may or may not have a consistent format with respect to the authority or ranks given, and then there is the "accepted" version of the name cleaned up to match some expected or more correct format.

Parsing a taxonomic name is not necessarily straightforward. The name as given on the label may be at any of a variety of taxonomic precisions from something like “Unknown Arthropoda” to “Rosa alba subsp. alba forma angustifolia.” For some taxonomic groups, “species” names are more or less straightforward using binomials or trinomials, but others, particularly plants, are much more complicated. Subspecies and variety, for instance, may be at the same rank (i.e. Rosa alba subspecies alba or R. alba var. alba) or at different ranks (R. alba ssp. alba var. ternata). Hybrid taxa and cultivated taxa add additional complications as does the inclusion of nomenclatural authority.

There can also be information in the label name that is not strictly part of a name such as “Rosa fulva (sp nova?)” or “conforms favourably with. Rosa Alba.” A system may or may not even allow you to enter such information and if you can enter it, it is less than straightforward how to relate it to the same name without the accompanying modifier.

Then there is the issue of taxonomic hierarchy, as the label name represents some identification within a hierarchy of linked names up to Kingdom. The placement of a species within a particular hierarchy is subject to interpretation and there are generally multiple hierarchies that might be applied to the same species. These hierarchies do not necessarily follow the same rank structure either. If that is not bad enough, in most taxonomic hierarchical systems, there are always some taxa whose placement is uncertain and thus left unlinked to next higher rank.

Synonymy is a large issue. Names can be related to each other through hierarchy, but also as partial or total equivalents. Handling synonyms means not only holding more names in your system, but also maintaining the relationships among them and developing mechanisms for manipulating them and distinguishing and applying currently accepted status consistently throughout they system. A related concern involves common names. A given system may or may not allow one or more common names to be associated with each scientific name. Maintaining these represent an additional management burden.

The way a given solution handles nomenclature may be tied to expectations about how many potentially applicable names there are. A drop-down list might work well for a database of the mammals of Iowa, but it would never do for the arthropods of North America.

Type specimens add in additional complexity in that name for which the specimen is a type may or may not be the same as the label name or the currently accepted name for that taxon.

All of these considerations should make it clear that obtaining or developing a perfect solution to handling nomenclatural information is not realistic. Instead, look for a solution that handles your needs relatively well, has reasonable flexibility to handle unusual circumstances, and which does not add an unduly large burden to manage.

Tracking nomenclatural changes

The label name is subject to interpretation and later revision either by a later expert review of that specimen (“determination” or “annotation”) or by application of a new taxonomic interpretation of that name as codified in a published manuscript or treatment (“nomenclatural update”). A given name may also simply be misapplied or mistyped during data entry and, thus, need to be corrected. For many reasons, it may be useful to track changes to the names that are applied to a specimen and track the date and person who changed it. It may be useful to track why the change was made, for example, to distinguish between a typographic correction and an expert determination. It may be useful to keep track of all determinations or just the original and the most current one.

Generating labels or tags

Collection management programs often generate labels or tags for specimens, although this is most commonly found for herbaria. Because of their size, arthropod pin labels often contain abbreviations and other shortenings of the total information such that they are difficult to generate automatically from the databased information. If the collection management program allows label generation, careful consideration should be given to the interaction with workflow such that new labels get associated with the specimen that was databased and that the new labelled specimens are clearly marked as distinct from labelled specimens that still need to be databased.

Tracking curation and transactions

Curation information can be as simple as maintaining the current physical location of a specimen or it can represent a detailed track of an object through elaborate processing phases up to and including it's final placement in the collection. Complications arise from maintaining proper identifiers and relating them through the processing phase. An accession may comprise readily identifiable and distinct objects at the outset, but in many cases, partitioning, sorting and later, more detailed identification may be involved before the objects are placed in the collection.

Transactions involve keeping track of loans, exchanges, and specimens sent out for identification. Tracking transactions entails information about people, institutions, policies, and documentation. In some systems and models, transactions can include transfer of a specimen from one collection to another within the host institution (such as to a teaching collection), gifts to or from the host institution or to keep record of specimens that have been deaccessioned or lost.

Marking sensitive data

Records may be need to be marked as sensitive for a variety of reasons such as for species that are rare or in danger of commercial collection or because they represent vouchers for ongoing research. Marking records as sensitive allows development and implementation of data restriction policy which, in turn, determines what to can be done with sensitive data. Data restriction policy can apply to either whole records, to certain types of information within records, or to some combination of record level and field level data access.

Marking records as sensitive can be as simple as putting in a sensitivity check field in your object record table, but more elaborate systems maintain mechanisms for recording notes multiple sensitivity markers per record and notes as to who marked a record as sensitive and why and for how long the record should be treated as sensitive.

Marking records as sensitive one record at a time can be an intensive process, particularly if the record set is large. Matching records with sensitivity criteria can also be difficult as it is hard to maintain current sensitivity information and the matching process itself can be difficult. For example, while it would be nice to mark all specimens as sensitive which are listed as federally endangered or threatened, these designations are often applied to populations, not taxa per se and most data models have a hard time capturing this detail. Cascading sensitivity from the listed name to synonyms that may appear on your specimen labels is also problematic.

Tracking record changes

An ideal system keeps track of what changed, who did it, when, and why for every field. This is generally impractical solution as it substantially inflates the data maintained by the system. A creation date/timestamp and modification date/timestamp field on the records that you are responsible for maintaining is a good start. Taxon name changes and determinations should have a separate system as the information is different. A determiner changes the name in a different way than the person who enters the determiner name.

Data exchange standards like Darwin Core expect to see a Date Last Modified field for each record. Interpretation of this field gets more complicated, however, when the exported modification date is triggered by updates to fields not in the export field set.

Allowing or supporting georeferencing

Georeferencing is the process of translating a locality description into a mappable representation of that description (Chapman and Wieczorek 2006). It is increasingly useful to be able to georeference legacy data (Beaman et al. 2004). It is also important to recognise that a georeference represents a hypothesis about where a collection event occurred. As such, for a given collection event or specimen there may be multiple georeferences. Some may represent the use of different methodologies (MaNIS vs. BioGeomancer protocols for instance) or different degrees of review (i.e., initial output vs. result of expert review or review by collector). It is also useful to distinguish between the results of a formal georeferencing and any coordinate information that came in with the specimen initially.

Handling collection dates

Collection dates tend to be problematic in information management systems because they are not always to the precision of a single day. Systems that enforce entry to a single date field should be avoided as this gives the appearance of more precision than is actually present in the data. Handling collection dates in a text field allows the information to be entered verbatim, however, the resulting data are unlikely to be useful as dates and there is likely to be a range of entered values for the same date (e.g. Aug. 23, 1976 and 10/23/1976). Dates with separators have the addition problem of ambiguity as to which number represents the day and which the month. Morris (2005) discusses the date issue at more length. Generally, a good solution will need to have a number of fields that allows entry of single dates, date ranges, and textual information (i.e. “Spring 1976”), while still allowing data to be represented as dates, at least the year information, if present.

Handling geographic administrative units

It may seem straightforward to record geographic administrative units (GAUs) with country, state/province, and county/district fields, but this is not always the case. GAUs can change such as the break up of the Soviet Republic, can change name, such as Rhodesia, and worse change geographic extent (Valencia county in New Mexico was broken into a new Cibola county and a smaller redefined Valencia county in 1978). GAUs may have different names commonly in use (i.e. “United States” vs. “U.S.A.”) and are generally different in different languages. Locality information may refer to units that are difficult to place into one of these three fields (i.e., England, the island of Hawaii, Greenland). Administrative ownership of a region may be distinct from the region itself (i.e. Martinique). Some collections do not come from any administrative unit at all (e.g. “700 miles south of Hawaii in the Pacific Ocean”) and some come from features that define the boundary between units such as river between two states or a ridge that divides two counties.

One solution for handling GAUs is simply not to hold this information in separate fields at all, but to include it in a more general locality field. This is generally less than satisfactory, however, as GAUs are commonly used as search and retrieval criteria and allowing free entry in a locality field will entail much more keystrokes and the introduction of typographic error. Collections from more limited geographic scope may be relatively insulated from most of these problems. If your holdings are more global, it may be important to give more attention as to how to handle GUAs.

Other features and issues to evaluate

The above are only some of the features and data issues you might need to give detailed attention to as you evaluate whether a potential package or development plan is sufficient to meet your needs. Other areas that might warrant similar detailed attention include:

Collector names and collector groups

Separation or conflation of locality, ecological description, associated species and other collection event information

Morphology and specimen preparation information

Observations

Images

Project and voucher information

Publications and literature

Institutions and collection metadata

Security and access

References

Armstrong, J.A. 1992. The funding base for Australian biological collections. Australian Biologist 5(1): 80-88.

Berendsohn, W.; Güntsch, A. and Röpert, D. Survey of existing publicly distributed collection management and data capture software solutions used by the worlds natural history collections. Global Biodiversity Information Facility, 2003. http://circa.gbif.net/Public/irc/gbif/digit/library?l=/digitization_collections&vm=detailed&sb=Title

Beaman, R., Wieczorek, J., and S. Blum. 2004. Determining space from place for natural history collections in a distributed digital library environment. D-Lib Magazine 10. Available at http://www.dlib.org/dlib/may04/beaman/05beaman.html

Chapman, A. 2005a. Principles of Data Quality. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/DataQuality **SEE ALSO: Chapter 3 of this *Manual*.**

Chapman, A. 2005b. Principles and Methods of Data Cleaning. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/DataCleaning **SEE ALSO: Chapter 4 of this *Manual*.**

Chapman, A. 2005c. Uses of primary Species-Occurrence Data. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/UsesPrimaryData **SEE ALSO: Chapter 1 of this *Manual*.**

Chapman, A. and O. Grafton. 2008. Guide to Best Practices for generalising sensitive primary species occurrence data. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/SensitiveData **SEE ALSO: Chapter 6 of this *Manual*.**

Chapman, A.D. and J. Wieczorek (eds). 2006. Guide to Best Practices for Georeferencing. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/BioGeomancerGuide **SEE ALSO: Chapter 5 of this *Manual*.**

Conn, B.J. (ed.) 2000. HISPID4 – Herbarium Information Standards and Protocols for Interchange of Data, version 4 (Royal Botanic Gardens Sydney) http://www.rbgsyd.gov.au/HISCOM .

Conn, B.J. 2003. Information standards in botanical databases – the limits to data interchange. Teleopea 10:53-60. http://www.rbgsyd.nsw.gov.au/__data/assets/pdf_file/72707/Tel10Con053.pdf

Häuser, C.L., Steiner, A., Holstein, J. & Scoble, M. J. (eds.) 2005. Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Results from a study of the European Network for Biodiversity Information. Stuttgart. 304 pp.

Lane, M. 1996. Roles of Natural History Collections. Annals of the Missouri Botanic Garden, vol. 83 (4): 536 – 545. http://www.jstor.org/view/00266493/di995851/99p0266q/0

Losee, R.M. 1997. A Discipline Independent Definition of Information. Journal of the American Society for Information Science. 48 (3): 254-269. http://www3.interscience.wiley.com/cgi-bin/fulltext/39670/PDFSTART?CRETRY=1&SRETRY=0

Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond integrity analysis. Proceedings of the Conference on Information Quality (IQ2000): 200 - 209. Boston: Massachusetts Institute of Technology. http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf [Accessed 18 February 2008].

McLaren, S.M. et al. 1996. Documentation standards for automatic data processing in mammalogy. Version 2.0. American Society of Mammalogists. 68 pages. Available at: http://www.mammalsociety.org/committees/comminformatics/docstandards.pdf

Meier, R. & R. Dikow. 2004. Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. Conservation Biology 18(2): 478–488. http://www.blackwell-synergy.com/doi/abs/10.1111/j.1523-1739.2004.00233.x

Morris, P.J. 2005. Relational database design and implementation for biodiversity informatics. PhyloInformatics 7: 1-66. http://systbio.org/files/phyloinformatics/7.pdf

Peterson, A.T. and Navarro-Sigüenza, A.G. 2002. Computerising bird collections and sharing data openly: Why bother? Bonner Zoologische Beiträge 51: 205-212.

Redman, T. C. 1996. Data quality for the information age. Artech House Inc.

McLeod, S. and M. C. Winans 1991. Logistics and planning for computerization. Section 2 in: *Stanley D. Blum (ed.),* Society of Vertebrate Palaeontology, USA: Guidelines and Standards for Fossil Vertebrate databases. 129pp.

Snow, N. 2005. Professional Biologist: Successfully curating smaller herbaria and natural history collections in academic settings. BioScience 55: 771-779.

Wheeler, Q. 2004, What if GBIF? BioScience 54:717 http://goliath.ecnext.com/coms2/summary_0199-48075_ITM

Appendix A: Business Case Considerations

Why digitise your collection?

Wider dissemination of Data

Enable your data to be studied in different ways

Enhance curatorial activities.

Protect your specimens.

Aid research by reducing future transcription time.

Fits the institutions corporate goals.

Enhances the ability of the institution to contribute in areas beyond its traditional remit.

Identify your goals

Institutional or individual?

Institutional A database that must cope with a wide range of specimens and many people entering data.

Individual The database will only be required to handle your specimens and data standards.

Who are the principal clients of your solution?

Individuals on a specific project

Researchers generally

Curatorial staff at the institution.

Others?

What language(s) will you support?

More languages cause greater complexity.

How much data?

The number of records affects the time digitisation will take and the scale of the database required for storing the information.

What data quality?

Will you record:

Collection data

Taxonomic information

Storage location

Habitat information

Initial description

The specimen itself

Data capture or data interpretation?

Data Capture Record the data presented on the specimen as written.

Data Interpretation Alter the data to correct errors, such as incorrectly naming the specimen.

Enhancing existing practices in the institution

Document the ways in which digitisation will aid the curators in your institution.

Imaging

What will be imaged?

How will you take your images?

How detailed will your images be?

What format will the images be stored as (JPG, TIFF etc.).

Where will they be stored?

How will they be accessed?

Understand what digitisation will not do

Databasing is not a money saving option.

Digitising your collection will not create new information for you

Specimens will still need to be physically stored and handled

When do you want the dataset be to be available?

Short term. Work that can be completed over a six to twelve month period.

Intermediate. Data entry over approximately an 18 month period.

Long term. Any project lasting longer than 18 months.

Future requirements

How will the database continue after the current project ends?

Staffing

Who will do the digitisation?

- Curatorial staff as part of their regular work.
- External contract staff/company
- Volunteer staff?
- Visiting researchers?
- Project staff?

How many can database at once?

- One person working at a single database
- Several people using individual databases.
- Several people sharing the same database

Is expert help available?

- *Yes* Your project will run much more smoothly.
- *No* Consider how suitable expertise will be made available to your project.

Is suitable expertise available?

- Data owners
- Data experts
- Technical staff
- Project management

Limitations

Is access to your data restricted in any way

- *Yes* Note which fields/specimens will not be released and why.
- *No* Consider the possible consequences of not restricting the data.

Does your institution require you to use an existing system?

- *Yes* Record how will this be integrated into your project and if this limits you in any way.
- *No* A database will have to be selected along with appropriate data standards.

Do you have legacy data (electronic or paper)?

- *Yes* How will this be integrated into your project and will you be able to check the data quality?
- *No* You will be able to set the data quality standards and provide reasonable quality assurance.

Do you already have project deadlines?

- *Yes* Prioritise discovery of how long you will actually take to digitise your specimens, then work backwards to discover how much time you have to plan the project. If there is insufficient time, consider requesting a project extension.
- *No* Take your time to properly plan your project.

Will you be working outside your institution?

- *Yes* Document the effects this will have on your database and factor suitable travel expenses into your resource requirements.
- *No* More freedom is allowed in choosing your database.

Physical Requirements

Where will the digitisation take place?

- Digitise in the collection itself
- Establish a dedicated area for digitisation
- Digitise in an entirely different location.

Document your existing I.T. infrastructure.

- This will allow you greater security when selecting an appropriate database.

Conclusions

Is your project feasible?

- *Yes* Begin to consider ways to implement your plan.
- *No* Revise your plans until they are practical.

Do your goals exceed your limitations?

- *Yes* Consider the following options when writing the action plan:
  - Can changing working practices free up time to work on your project?
  - Can other nearby institutions help out?
  - Who might fund your project?
  - Should your project be broken down into several stages
- *No* Start writing your action plan.

Appendix B: Action Plan Issues

Which Database?

Pick a database solution

- Commercial
- Open source
- Modified commercial or open source
- Bespoke

How long will it take to build or implement? (include in the project lead time).

Resources

- How many staff do you need? (Digitisers, manager and other staff).
- How will you train your workers?
- What are you going to have to buy?
- What budget do you require?

What will your workflow be?

- Collecting and returning the specimens.
- Digitisation location.
- Data Quality.
- Adding value to the original data
- Imaging.
- Data order
- Data checking.
- Can procedures be overlapped? .
- What effect will staff absence have on your workflow?
- Are there any bottlenecks in your plan?

The human element

- Staff loss or extended staff absence
- Training

Contingency planning/risk analysis

- How will you address the issue of not making the required digitisation rate, if it becomes an issue?
- What happens if your computer breaks down?
- Backup strategies.
- Malicious alteration of data.
- What will you do to document your solution/implementation?
- Are there other risks you should take account of?

Conclusions

Will your solution provide the appropriate level of data quality?

- *Yes* Your data is perfect for your current project and should be useful in other projects.
- *No* Can you:
  - add resources to improve data, or
  - reduce the total number of specimens worked on, allowing more time to enhance the remaining data, or
  - Improve the data quality in a future project?

Does your chosen solution match your goals, limitations, and resources?

- *Yes* Implementing the project should be straightforward.
- *No* Refine your solution.

Will your solution handle your future requirements?

- *Yes* Maintaining and extending your data will be easier.
- *No* Not an issue for the current project but may be a problem for future projects.

Is your solution a good return on your investment?

- *Yes* Begin to consider ways to implement your plan.
- *No* Revise your plans until they are practical.

How long will it take to database your collection?