Journal of Natural Science Collections

Title: Exploitation of digital collection data at the Museum für Naturkunde, Berlin

Author(s): Jancke, S., Striebing, D. & Mayer, F.

Source: Jancke, S., Striebing, D. & Mayer, F. (2014). Exploitation of digital collection data at the Museum für Naturkunde, Berlin. *Journal of Natural Science Collections*, *Volume 2*, 41 - 46.

URL: http://www.natsca.org/article/2078

Exploitation of digital collection data at the Museum für Naturkunde Berlin

Saskia Jancke, Dirk Striebing
& Frieder Mayer

Museum für Naturkunde - Leibniz Institute for Evolution and Biodiversity
Science, Invalidenstrasse 43, 10115 Berlin, Germany

Corresponding author: saskia.jancke@mfn-berlin.de

Received: 14th Sept 2014
Accepted: 18th Dec 2014

Abstract

Item information for many collections in the Museum für Naturkunde Berlin (MfN Berlin) - as in many other museums - is often stored locally by each curator in different formats such as *Excel* spreadsheets. These files are often accessible exclusively by collection staff members. Within the project “Exploitation of digital collection data” funded by the Deutsche Forschungsgemeinschaft (DFG) the data from the MfN Berlin mammal collection were transferred from *Excel* spreadsheets to an SQL server using scripts developed by a database specialist at the MfN Berlin. The data were subsequently standardised (for example in terms of their taxonomic and geographical information) then transferred from the SQL server into *Specify* - a museum database software application. Further aims of this project are the development and implementation of common transfer tools to achieve data migration to open-access databases such as BioCASE and GBIF as well as information retrieval like the current distribution and protection status of the specimen from databases like the IUCN redlist. This will allow external information retrievals of collection data and thus will open new avenues for scientific exploration of the collections. We have successfully applied our data transfer pipeline to the mammal collection of the MfN Berlin which is the 4th largest of its kind worldwide. These methods and tools can be used for the data migration in other collections at the MfN Berlin with its approximately 30 million collection objects, and also by other museums.

Keywords: Collection Management System; Data Improvement; Data Retrieval; Data Transfer; Natural History Museum; SQL; Transfer Script; Mammalia; Open Source; Biodiversity Informatics

Introduction

The mammal collection of the Museum für Naturkunde Berlin (MfN Berlin) was founded about 200 years ago with the aim to collect specimen from all over the world. The collection originated with just arond 40 specimens from the cabinet of curiosities of the *Akademie der Wissenschaften* in the 18th century (Jahn, 1985). From 1816 onwards accessions were recorded in catalogues and while at first the collection grew very slowly (with rarely more than 100 specimen per year), soon this growth increased, and between 1906 and 1916 the annual accession rate comprised on average 3,500 new specimens (Angermann, 1989).

The collection grew so fast that it was not possible to keep record of all accessions. In 1926 the general collection catalogue contained 35,693 entries which meant that considerably fewer specimens were recorded in the catalogue than were present in the collection (Angermann, 1989). Specimens came from all over the world and were collected during expeditions or sent by Germans living abroad, bought from traders or professional collectors, exchanged with other Museums, or were given by the *Zoologischer Garten* in Berlin. Many new species were described during these early days and the collections hold many type specimens. It is estimated that the mammal collection of the MfN Berlin currently has around 150,000 specimens. Most of the information was kept on the specimen labels and partly in accession catalogues.

Jancke, S., Striebing, D., & Mayer, F. 2015. Exploitation of digital collection data at the Museum für Naturkunde Berlin. *Journal of Natural Science Collections*. **2.** pp. 41-46.

After the Second World War, a large amount of skins were in very bad condition due to penetration of rain water and high humidity in the collection rooms. Labels became illegible or got lost during cleaning processes; the information loss often meant academic devaluation of the specimen. Furthermore original expedition and collection lists as well as the taxonomic catalogue were lost during this time. Another problem originated from the practice to send only skins to the Zoological Museum while the skeleton, skulls and alcohol material of the individuals went to the anatomical collection of the medical faculty. However, in the anatomical collection no information on the collector, collection date, nor the locality was recorded which meant a great loss of information. Both collections were recombined later (Angermann, 1989) and only with great effort, all collected body parts could subsequently be associated to one individual and inventoried with one definite catalogue number.

About 50 years ago specimen information sorted by the inventory number from the accession catalogues started to be transcribed on file cards to facilitate research on the specimens for requests. File cards were sorted by taxonomy. 10 years ago this information started to be transferred to excel spreadsheets, one for every mammal order. These spreadsheets contain the information on the inventory number, taxonomy, preparation, locality, determination, collector, collection date as well as accession. These excel lists are continuously updated and amended. To date about 79,000 specimens are inventoried which comprises about half of the specimens in the mammal collection and the improvement of collection data and inventories are ongoing.

The Natural History Museums collections hold very valuable specimens and specimen information for scientists of different disciplines such as taxonomists, ecologists, evolutionary biologists, geneticists, paleontologists, archaeologists, as well as historians, etc. Information is frequently requested for various types of studies such as recreating the historic distribution of species or the genetic analysis of rare or even extinct species. On one hand we have important data which are often locally based with the files accessible exclusively by collection staff members. On the other hand we have global online biodiversity databases such as GBIF or BioCASE which provide an important open-access research infrastructure. However, the data transfer from locally hosted museum databases or spreadsheets to these open-source biodiversity databases is seldom accomplished. Here we developed a framework to allow the data transfer from museum's collections to open access databases. This will allow external information retrievals of collection data and thus will open new avenues for scientific exploration of the collections.

Methods

To allow the data transfer from museum collections to open access databases, the data needed to be stored in consistent data formats in SQL databases. At the MfN Berlin we use the collection management system *Specify* developed by the University of Kansas, which is an open source database system with a MySQL database backend and a Java application frontend.

A further aim was to develop transfer tools so that the data could also be stored in open-access databases such as BioCASE & GBIF allowing external information retrieval of collection data such as the species distribution or protection status e.g. from the IUCN webpage.

We developed the following methods to accomplish these goals (Figure 1, below we discuss these steps in more detail):

1. Pre-Importation of collection data from excel spreadsheets into a Microsoft (MS) SQL-database.
2. Standardisation of data and improving data quality using MS Access as the front end. Eradicating/removing double entries and duplicate inventory numbers so that every specimen is recorded explicitly and completely.
3. Transfer of the improved data to a MySQL database and final error checking.
4. Develop the transfer tools to transfer data from the SQL database to *Specify*. Transfer of the collection data into *Specify* 6.
5. Develop the transfer tools for the data transfer between *Specify* and open access biodiversity databases such as GBIF and BioCASE.
6. Develop the transfer tools to retrieve data such as the protection status and distribution from the IUCN webpage.

As a first step, the mammal collection data stored in excel files were transferred to a MS SQL database to bundle all conflicts. Using an MS Access frontend, the data were reassessed in terms of taxonomy (correct and valid species name based on the taxonomy of Wilson & Reeder, 2005), locality (update the locality information which was a challenge especially for all the old colony names and changing frontiers since the collection date), eradicating spelling mistakes and number duplicates.

The geographic tree from *Specify* was used as provided by the developers to allow for continuous and automatic updates. The geography was then automatically related to the locality data where possible, while conflicts were solved manually in MS SQL. Where possible the localities were related to modern countries, while border regions were defined as new countries, e.g. the region Abessinia reaches from Ethiopia to Eritrea, so the newly defined country would be Ethiopia/Eritrea. Historic locations were researched using the collectors' itinerary information where possible, researching place names online and specifically on getamap.org. Occasionally the locality and collector's information previously researched by a colleague from the ornithological collection was used as well.

This pre-import was an important step for the data validation but it was independent from the import solution. For the *Specify* import an import database was used which is on the same server, a MySQL server, as the *Specify* database. The data were transferred from the pre-import MS SQL database to the MySQL import-database using simple SQL-insert-commands which can be used on all platforms. The import-database consists of tables with fields from *Specify* in a 1 to 1 relationship. There are different tables for different information with fields such as determination, collector, or location. Some fields have several alternatives with different precisions or formats, e.g. the information on a collection date can be stored in a field for a complete date as well as two further fields for month and year. Subsequently, all the field values were distributed into their respective target fields.

The taxonomy was imported first to *Specify*, so that a taxonomic tree was already available in the database before the specimen data were imported. The taxonomy included a list of all valid species names from the Catalogue of Life including the synonyms, and was completed with the information from the Wilson Reeder taxonomy (Wilson & Reeder, 2005) and stored in a CSV file.

Results

Originally 86,478 specimens were recorded in 30 excel spreadsheets (one record per row) for the mammal collection in the MfN Berlin. During the data transfer process of the specimens recorded in the excel spreadsheets, duplicate or indefinite entries were eradicated, so that subsequently 78,775 valid specimens were recorded in the new *Specify* database. Indefinite entries were mostly duplicates or different body parts which were recorded separately in the spreadsheets but identified as one individual during the data improvement process in the MS SQL database.

The excel spreadsheets consisted of 26 columns with information on taxonomy, locality, collecting information, remarks and an identifier. Information stored in these fields was transferred into 49 definite fields in *Specify*.

The data transfer was completed within one year. The limiting step was the improvement of data quality. However, this was simply done based on the data already entered. If data were transcribed incorrectly from the specimens labels into the excel spreadsheets, i.e. localities or collectors were misread on the label when entered into the excel spreadsheets or if the transcription was faulty, it was not possible to correct this during the project year. Furthermore, if specimens were determined incorrectly and the information on the determination in the spreadsheet therefore incorrectly, a correction during the course of the project was impossible

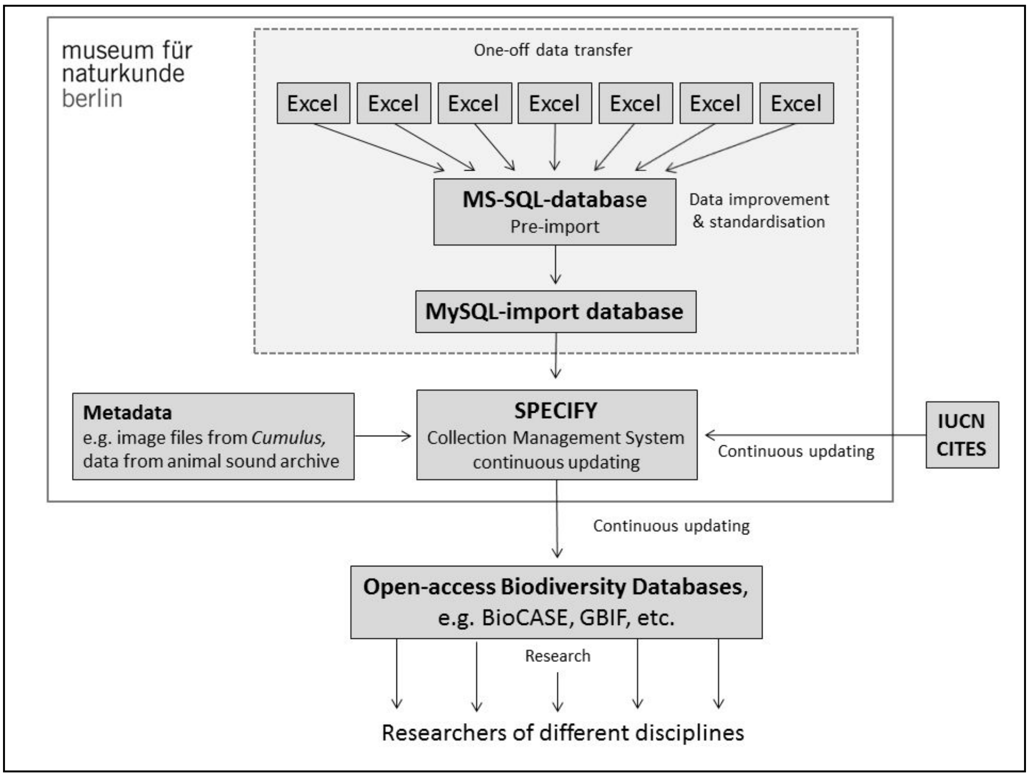
museum für naturkunde berlin
One-off data transfer
Excel
Excel
Excel
Excel
Excel
Excel
Excel
MS-SQL-database
Pre-import
Data improvement & standardisation
MySQL-import database
Metadata
e.g. image files from *Cumulus*, data from animal sound archive
SPECIFY
Collection Management System
continuous updating
Continuous updating
IUCN
CITES
Continuous updating
Open-access Biodiversity Databases, e.g. BioCASE, GBIF, etc.
Research
Researchers of different disciplines

Fig. 1. Workflow of the data transfer process.

as this would have required accessing every specimen to check them as well as the information on the labels.

However, it was possible to improve a great amount of data e.g. in terms of locality or taxonomic information. Locality information in the excel spreadsheets was often transcribed from the specimen label and entered into two fields in excel: "locality name" and "present locality name – country". If locality names were misspelled or not up-to-date (like e.g. old colonial names) the information was updated during the data improvement step in the MS SQL database. Some specimens came from the same locality and the advantage of using an SQL database is that this information can be cumulated. Accordingly the 78,775 specimens were bundled in terms of the geographical information and resulted in 20,950 different localities. It was then updated during the data improvement process after the pre-import to the MS SQL database. The localities described in the excel spreadsheets were researched and if the research was successful, the modern locality name was noted together with the affiliation to the modern country, continent etc. For 9,026 localities (43%) the information was explicit and could be transferred automatically into the *Specify* schema, while the remaining locality information had to be updated, improved and standardized before the data could be transferred.

For example, the information:
Locality name: "Amani, Usambara, D.O.A." - had to be updated and standardized before a data transfer was possible.
An example for an explicit data entry which could be transferred automatically is:
Locality name: "Potsdam" and *Present locality name - country*: "Germany".

The mostly historic locations were related to 290 countries from all 7 continents. Oceans and seas were defined as new continents which were not already present in the geographic tree existing in *Specify*.

Another aim was to validate the taxonomic information. Of altogether 3,980 different taxa, 962 taxa were updated manually in the MS SQL database following the systematics of Wilson & Reeder (2005). 2,117 type specimens were entered in the database, 383 of which were holotypes.

In terms of data standardization e.g. 11,650 collection dates were standardized. Dates can be described in different formats in excel when the field is not defined as a date field as it was the case for the mammal collection data. Information can be written inconsistently (e.g. already the month of a date can be recorded in different formats such as "3", "03", March or German März). These dates were standardized and if information was missing (e.g. only March 1906 was recorded), the information was put into fields describing incomplete dates (month: March; year: 1906).

Also the preparation (skull, skeleton, skin, and alcohol material), determination, as well as the collector/accession were standardized for the import so that all information was spelled correctly, consistently, and subsequently put into definite fields.

Since the transfer of the mammal collection data further transfers have already been completed such as data of the embryological collection stored in CSV files as well the data of the collection of Orthoptera which were kept in a FileMaker database. In the historical department additional information on portraits were added from excel spreadsheets to the existing information stored in the archiving SQL database system LARS (Leistungsstarkes Archivierungs- und Recherchesystem). For these imports the scripts from the mammal collection data transfer were used. These imports profited directly from the experience made by our data transfer to Specify and no additional time for developing these scripts was needed.

Discussion

Scientific collections are of great value for biodiversity and collection data are an important research infrastructure (e.g. Türkay, 2011; Lister *et al.,* 2011). It is good research practice to keep all the primary information on the specimens in a database and this database increases exponentially in utility when it is globally accessible (Türkay, 2011). There is a great interest in opening up natural history collection data to the wider community as show by European projects like *Open up!* (Berendsohn & Güntsch, 2012).

Excel spreadsheets were the first way of digital data capturing and storage in the mammal collection in the MfN Berlin. The spreadsheets contained all information on taxonomy, locality, preparation, accession as well as the inventory number, and were continuously added and updated. The problems that come with this excel spreadsheets are 1) updates have to be done per record and 2) inconsistencies in data entry (e.g. sex can be female, Female, F, f, fem. etc.) which makes it more difficult to search for definite terms. If the taxonomy changes or a historic locality has been researched, updates can't be done cumulatively. Inconsistencies in spelling and misspelling are also likely sources of error. A database like *Specify* offers a solution as updates can be done cumulatively, sources of errors are reduced by accessing information from the taxonomic or geographical tree, and inconsistencies in spelling are reduced if updating is done cumulatively or by using a predefined selection (dropdown list). This reduces the workload for the collection staff considerably.

Specify offers an import tool in the Workbench where excel files can be imported directly into the database. However, the data transfer into *Specify* is eventually planned for all collections in the MfN Berlin. The scripts which were developed in this project to transfer data between different Microsoft applications such as excel but also from other database systems to MS SQL can be used for various collections. Another important aspect of the data import was to create a taxonomic tree based on the Catalogue of Life and the Wilson & Reeder (2005) taxonomy including the valid taxonomic names as well as the synonyms. Importing synonyms into the database using the *Specify* Workbench was not technically feasible and the pre-import to the My SQL database therefore necessary.

In the MS SQL database the conflicts were bundled as well as their solutions. For example, in our spreadsheets the information on the locality where the specimen was found was stored in just one field, including e.g. the country, town or location, sometimes in rare occasions also the georeferences. This information had to be transferred to definite fields like one for the country, one for the continent, etc. Data improvement and standardization was the most time-consuming task due to the number of specimens, therefore to bundle the conflicts was essential to allow the data transfer from excel into *Specify* with reasonable time effort. After the data improvement on the MS SQL server, data were transferred to a MySQL server. The import scripts were written for MySQL and can also be used by other institutions and users who want to import data to *Specify*.

The data standardization and improvement was done in MS SQL, however, it could have been equally done in MySQL. There are several reasons why this intermediate step was used at the MfN Berlin: 1. MS SQL server have been used for many years at the MfN Berlin and the database staff has expert knowledge in writing scripts for MS SQL; 2. MS Access used as a frontend is an user friendly and unproblematic tool to access SQL data; 3. Microsoft extensions such as Transakt-SQL provide further applications, e.g. Table-Valued functions and it allows recursive function requests.

Subsequently, important applications were translated for MySQL users and provided for download via the GitHub link under references. In January 2015, the MS SQL and MySQL import scripts can be downloaded from the following link: https://github.com/mfn-berlin/Sp6ImportDB/tree/master.

Tools such as the BioCASe Provider Software to connect databases such as *Specify* to open-access databases like GBIF and BioCASE were already developed (Glöckler, *et al.,* 2013) and once the data are unlocked automatic updates will allow external users to screen and retrieve the data of the mammal collection. It is planned to transfer the first mammal collection data of the MfN Berlin in year 2015 when an ongoing locality-georeferencing project has been completed.

One important part of providing digital access to natural history collection data are the quantitative geospatial references of biological collection data because they provide a quantitative basis for biodiversity analyses (Beaman, *et al.,* 2004). Retrospective georeferencing makes collection material more valuable because this allows spatial analysis (Murphey, *et al.,* 2004). Subsequently, our next step is to georeference the collection material of the mammal collection following best practice (Chapman & Wieczorek, 2006; Wieczorek, *et al.,* 2012).

However, historic locations are often difficult to put into a modern context. Researching the expedition routes as well as georeferencing old maps and locality names will subsequently provide important information on former distributions of species. In the past collectors were not just interested in one taxonomic group but a wide range of collectables, and during one field trip they would collect birds as well as mammals or even ethnological examples. Exchanging the already researched information for example on an institutional or even national or international level using a collectors and historic localities database saves time and efforts for museums staff. This is another potential use of the data stored in *Specify* as the locality information in relation to the collectors information can be retrieved and provided potentially in a specific collector's database.

First experiences of using *Specify* as the collection management system in the mammal collection show the important advantage of such a tool particularly for queries concerning e.g. localities and collectors, and especially where more than one taxonomic order is involved. In the past answering these queries was more time consuming using excel spreadsheets firstly because several spreadsheets had to be searched and secondly due to the inconsistencies in spelling and in defining localities and collectors. However, the very detailed structure of *Specify* with varies information stored in numerous but definite fields can cause problems when searching for unstandardized information. Information unspecific or unspecifiable for one field can be stored in different remarks fields. This information is then difficult to localise when creating a query. However once standards for the data entry for this kind of unspecific information has been developed and queries have been refined, the advantage of using a SQL database system will prevail. Another great advantage of using a collection management system such as *Specify* especially for the mammal collection in the MfN Berlin is that different body parts such as skull, skeleton, skin or alcohol material which was by mistake inventoried using different inventory numbers can now more easily be identified as one specimen using versatile filter options and due to standardised information of localities and collectors.

In summary using the transfer tools and data standardisation processes specifically developed during this project, it was possible to complete the data transfer of the comparatively large amount of mammal data successfully within a reasonable timespan of one year including a considerable improvement of data quality.

Acknowledgements

The collection managers Nora Lange and Christiane Funk of the mammal collection MfN Berlin as well as Pascal Eckhoff of the ornithological collection for providing data; Falko Glöckler and Dieter Korb for technical advice and Emma Berdan for linguistic help. This project was funded by the Deutsche Forschungsgesellschaft (DFG).

References

Angermann, R. 1989. Die Säugetierkollektion des Museums für Naturkunde der Humboldt-Universität zu Berlin, *Säugetierkundliche Informationen*, **3 (13).** pp.47-68.

Beaman, R., Wieczorek, J. & Blum, S. (2004) *Determining Space from Place for Natural History Collections*. D-Lib Magazine May 2004. **10 (5).** [Online], http://www.dlib.org/dlib/may04/beaman/05beaman.html [accessed 23 July 2014].

Berendsohn, W. G., & Güntsch, A. 2012. OpenUp! Creating a cross-domain pipeline for natural history data. *ZooKeys.* **209.** pp.47-54.

Biological Collection Access Services (BioCASE) http://www.biocase.org/ [accessed 02.12.2014]

Chapman, A. D. & Wieczorek, J. (eds.) (2006) *Guide to Best Practices for Georeferencing,* Copenhagen: Global Biodiversity Information Facility.

Global Biodiversity Information Facility (GBIF) http://www.gbif.org/ [accessed 02.12.2014]

Glöckler, F., Hoffmann, J., & Theeten, F. (2013) The BioCASe Monitor Service - A tool for monitoring progress and quality of data provision through distributed data networks, *Biodiversity Data Journal* 1: e968.

Jahn, I. 1985. Zur Vertretung der Zoologie und zur Entwicklung ihrer institutionellen Grundlagen an der Berliner Universität von ihrer Gründung bis 1920, *Wissenschaftliche Zeitschrift der Humboldt-Universität Berlin, Reihe Mathematik/ Naturwissenschaften*. **34(3/4).** pp.260-280.

Lister, A. M. & Climate Change Research Group, 2011 Natural history collections as sources of long-term datasets. *Trends in Ecology & Evolution*. **26(4).** pp.153-154.

Murphey, P. C., *et al.* 2004. Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi), *PhyloInformatics*, **3.** pp.1-29.

Türkay, M. 2011. Wissenschaftliche Sammlungen: Unersetzbare Ressource der Biodiversitätsforschung, *Senckenberg: Natur,Forschung,Museum*, **14 (3/4).** pp.66-73.

Wieczorek, J., *et al.* (2012) *Georefer- encing – Quick Reference Guide* [Online], https://www.idigbio.org/wiki/images/1/1e/GeoreferencingQuickReferenceGuide.pdf [accessed 17 July 2014].

Wilson, D. E. & Reeder, D. M. (eds.) 2005. *Mammal Species of the World. A Taxonomic and Geographic Reference*, (3rd edition), Baltimore: Johns Hopkins University Press.