NatSCA News

Title: The organisation and distribution of electronic data at the Natural History Museum

Author(s): Erica McAlister

Source: McAlister, E. (2009). The organisation and distribution of electronic data at the Natural History Museum. *NatSCA News*, *Issue 17*, 34 - 36.

URL: http://www.natsca.org/article/143

The organisation and distribution of electronic data at the Natural History Museum

Erica McAlister

Natural History Museum, Cromwell Road, London, SW7 5BD
Email: e.mcalister@nhm.ac.uk

Introduction

Museums have traditionally focused on preserving collections/treasures etc as well as allowing taxonomists, amateur as well as professionals, to gather and utilise information from the collections.

Nowadays there is a greater need to access this information, and to access it quickly due to increased awareness of global biodiversity issues. The challenge is to enable today's researchers a rapid and accurate database with as much relevant material about the specimens as possible, in an easily accessible format. . We also need to locate, manage and audit the collection effectively, both at lot level (i.e. species) and at specimen level.

History of database at NHM

Historically, at The Natural History Museum, each department, sub department, individual etc has had their own database of the specimens/species that they curate. This resulted in many databases which were incompatible with each other and more often than not did not offer online access. The only online databases in entomology were more involved in relating the research outcomes of the NHM staff or publishing a complete list of all species found within specific groups, whether or not there were specimens at the museum.

What the museum didn't have was any standard online catalogue of the species and specimens that were found in our collection. As well as many of the smaller databases, the Entomology Department held much of its data in a main collections management database – Paradox for Dos. This, along with many of the other databases were flat i.e. the information was stored in one long text file (a tab delimited file) requiring a standardised input of data. The data was entered onto one page and contained information on the taxonomy, correct status, as well as the locality within the department and what the storage method was (pinned, slide etc). As a retrieval system, it was quick and simple; you were able to ascertain whether we had any specimens, if there were any types and where they were. But this was it. We now want to garner a lot more information from the data that is attached to the specimen i.e. the specific localities for species distribution maps. There was no place on the old system to have this data.

The database was riddled with spelling errors and there was no clear way of enabling the data to be accessed online. In addition, the data had not been entered consistently so there were many different uses for the same field, and the plurality of databases limited the ease with which staff could utilise (or even access) databases from different Sections. The loan system, which is an important part of the way that the collection is managed, was stand-alone. If you traced a specimen in the main collection you had no way of knowing if this was out on loan without either coming out of the system or physically checking the collection.

The New database

A new system was clearly needed; the implementation of a universal database that could bring in line all data handling within the museum and enable it to be accessed by the wider scientific community as well as the general public.

After much deliberation the museum chose the KE-EMu software which is an object-relational database that supports text as well as multimedia objects. A relational database differs from the original flat databases in that they group the data using common attributes found within the dataset e.g. author, collecting locality and taxonomy, whilst In the Object-relational system the data are grouped around objects which are then attached to one another as appropriate rather than grouped according to common attributes i.e. specific author to the species that they have described.

One of the many advantages of this system was that KE provides a broad range of on-going services including provision of expert advice to promote best practice in museum data management.

There were many Implementation and post-implementation activities and tasks, including specification development and Software configuration and customisation – we were able to customize many of the modules to suite our specific needs as a museum. They were able to assist in the developing of reports and give training to enable these to be developed internally. And there is assistance in Web site development and web publication of collection material.

Once the initial process of deciding which system to use, specifying and developing the modules, it took over two years (and some are still going) for all of the different departments' data to be migrated into the new system, but the benefits are beginning to be felt. Each department has initially a similar list of modules including taxonomy, catalogue, parties, bibliography etc but are able to modify the modules to their own specifications.

KE-EMu Database
Entomology relies heavily on the collection index module as it does not have information on every specimen that there is in the department. We have only the specific information on certain individuals within a species, such as the type material or if there was a notable collection (Stag beetles).

The collection Index record brings together the information about the lot localities (there may be several due to slide material, spirit and pinned) and the taxonomy (including a current name if it is a junior synonym). Other modules include the catalogue module which records information about each object or species within the collection; bibliography records details about bibliographic references e.g. journals, books, citations; and parties records the details about people and organisations that can either be contemporary borrowers or authors, recorders etc.

Now that most of the migration is complete, there is a considerable amount of work to do with the data and the operation of the system;

>   1. Data clean-up – now all of the data is in the same depository a whole raft of new issues has arisen e.g. duplication of records sitting in separate datasets – this is the legacy of non-relational model previously.
>
>   2. Training – we have a large body of staff. Core Collection management staff of ca. 30. Up to 30 others to train.

KE-EMu is definitely a more complex system compared to the old Paradox system and the key to success of Emu is ensuring there is an excellent skill base across the department. There is a very good support network from KE-EMu as well as within the department

The department has been using the system for a couple of years now, and the loan system has been successfully integrated. No we are able to create a record for a specific borrower and record the actual specimens that they are borrowing (on the old system it was only the species level that was recorded), giving us a greater ability to track the movement of individual specimens. Considering that we hold one of the largest collections of Type material that is an exceptionally useful tool. We are also able to track anything that comes into the museum (recorded at object level entry) such as external loans and there is the potential for attaching documentation including permits and visas.

EMu provides fully integrated support for a wide range of multimedia resources, including images, video, audio, word processing documents, spreadsheets, presentations, in fact, any online or off-line resource. We are able to add examples of handwriting which is very useful when it comes to identifying scribbles on labels. Any images of the specimen can be added to its specimen level record which can be sent on request to other researches, thus reducing the need to send the specimen itself.

Now that there is one database the ability to serve the data online is manageable, with the data from Mineralogy and Zoology already being accessible on the web. As previously said historically, most of the information online was individual research databases generally associated with specific research staff. Now, there will be 5 specific online catalogues (for each science department), which have been designed in-house to provide specific functionality that is related to the individual departments needs.

The intention of the museum is to get all, where appropriate, records on the web which will include many dirty and redundant records. As a result of the implementation of KE-Emu and the modification and insertion by June/July all specimen records will be served to GBIF.

Just the beginning
There are many challenges that I have not mentioned including deciding what data to collect – given the vast number of fields the time possible to spend populating the database is infinite – how do we decide what to database?   There is also the question, for online data, of data sensitivity and to what extent we need to be active in this. There is much still to think about but at least now there is a system that will let us achieve this.