

Computerising Natural History Card Archives

A.C. Downton, S.M. Lucas* and G. Patoulas,
Department of Electronic Systems Engineering, University of Essex, UK
*Department of Computer Science, University of Essex, UK
{acd,sml,gpatou}@essex.ac.uk
G.W. Beccaloni, M.J. Scoble and G.S. Robinson,
Department of Entomology, Natural History Museum, London, UK
{gwb,m.scoble,gsr}@nhm.ac.uk

Abstract

This paper summarises the achievements of a multi-disciplinary Bioinformatics project which has the objective of providing a general mechanism for efficient computerisation of typewritten/hand-annotated archive card indexes, of the type found in most museums, archives and libraries. In addition to efficiently scanning, recognising and databasing the content of the cards, the original card images must be maintained as the ultimate source record, and a flexible database structure is required to allow taxonomists to reorganise and update the resulting online archive. Implementation mechanisms for each part of the overall system are described, and conversion performance for a demonstrator database of 27,578 Pyralid moth archive cards is reported. The system is currently being used to convert the full NHM archive of Lepidoptera totalling 290,886 cards.

1. Introduction

In addition to 68 million biological specimens, The Natural History Museum, London (NHM) houses global index card archives of taxonomic data for many important groups of organisms extant and extinct. One such index is that to the world species of butterflies and moths (Lepidoptera). In effect, this card archive represents a comprehensive inventory of the scientific names and associated bibliographical data for the whole of this species-rich order of insects, for which no published global catalogue currently exists. The main part of the index consists of 265 drawers containing 290,886 index cards. Each 5" x 3" card contains bibliographic data and other information for one scientific name (genus-group, species-group, and often infrasubspecific), laid out in a standardised format (Fig. 1). This information is usually type-written, but a significant minority of cards are entirely hand-written and many of them contain hand-written annotations.

Cards are ordered within the index: first, according to higher classification (superfamily, family, subfamily, tribe); second, alphabetically by genus; third, alphabetically within each genus by species; and fourth, alphabetically within each species by subspecies (hence the card sequence implies several database fields which are not explicitly included on every card). Cards with names that are no longer in use (e.g. synonyms of current species names) are arranged alphabetically by scientific name following the card for the currently valid name. This systematic index is supplemented by an alphabetic card index, containing corresponding (and hence redundant, or unnormalised) data to cards in the systematic index, arranged by superfamily. The alphabetic index allows cards to be located in the systematic index in cases where the current genus name or the higher classification is not known, highlighting one of the access limitations of non-computerised archives.

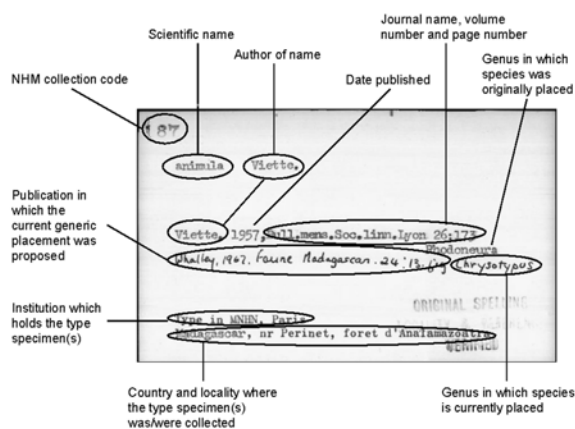


Figure 1. An index card with multiple hand print and handwriting annotations showing components to be extracted.

The importance of the NHM Lepidoptera card index is demonstrated by the fact that taxonomic catalogues for several groups of Lepidoptera have been produced largely based on data from it (e.g. for Noctuidae [i], Geometridae

[ii], and the *Butterflies & Moths of the World: Generic Names & their Type-species*, providing a full list of genus names online [iii]).

Neither standard office OCR products nor manual conversion are suitable for recognising and converting the content of cards as explained in [iv], hence our approach has been to develop a suite of document analysis and form processing tools specifically optimised to support legacy archive conversion. The overall structure of the system developed is shown in Fig 2.

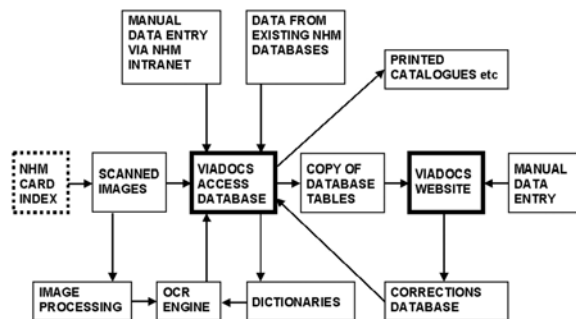


Figure 2. Overall structure of the VIADOCS archive conversion system.

2. Archive card scanning

Card archives are scanned using a modified SEAC Banche RDS-6000 bank cheque scanner. This has the capability to scan both sides of a card simultaneously in colour and/or monochrome at 200 pixels/inch resolution at a rate of about 1 card/second, and is also able to print a reference image number on the back of each card for cross-checking purposes. A customised software interface to drive the scanner was implemented in Tcl/Tk and C, using our established HUE environment [v].

At the beginning of each scanning session, the user enters the higher classification of the batch of cards to be scanned and the number of the index drawer from which the cards were taken. Using this information, the software creates a hierarchy of folders and unique filenames on the PC hard disk or DVD-RAM drive that mirrors the higher classification of the cards. It also prints an equivalent string on to the back of each card to provide a unique cross-reference. For example, if card number 38378 from the index was taken from drawer “40A”, and the name of the taxon on the card was placed in the family Apoprogonidae of the superfamily Geometroidea, then the front image of the card would be named “FC-Geometroidea-Apoprogonidae-40A-038378.jpg”. The back image would have the same name but prefixed “BC” rather than “FC”, and the code printed onto the card back would read “Geometroidea-Apoprogonidae-40A-038378”. The two images of this card would be saved in a

folder named “Apoprogonidae” placed within a folder named “Geometroidea”.

When the scanner interface programme is exited, it stores the number of the last card scanned. On restarting, this card number is retrieved and incremented, thus ensuring that all card images receive a unique number. This number can optionally be set during scanning, if, for example, a previously scanned card needs to be rescanned or cards are not scanned in sequence.

3. Archive Card Document Analysis and OCR

In addition to the scientific name, Lepidoptera cards contain author, taxonomic status, date of first description, journal reference, subsequent citations, taxonomic changes, type locality, whether the NHM holds the primary reference specimen (the 'type'), and location of the specimen within the Museum's collections. Document analysis is required to segment and parse the overall card content into word sub-images to be fed to the OCR system, which in turn maps recognised words to database fields. The cards in the archive conform to one of several basic templates, depending on whether they are genus/subgenus, or species/infraspecies. In general, each card can be divided into a header region containing the name, the author, and synonym information, a bibliographic region describing journal references and subsequent citations and changes, and a geography region describing the locality of the type specimen. Additional rules for writing species or genus names specify use of different colour (red), all upper case, all lower case or capitalised lower case typing; these are combined with spatial analysis and card sequence information to identify the position of specific components on each card, using initial layout guidance and field labels provided by a taxonomist through a Graphical User Interface to a generalised document analysis tool. For each defined word image component, specialist (though not always complete) dictionaries have been derived from existing online sources; these are applied during OCR. Details of the current Archive Card Document Analysis system and the performance it achieves are described elsewhere [vi], as is the implementation of the OCR system [vii,viii].

4. Archive card database

The card images and associated taxonomic data are managed using an Access relational database consisting of 7 linked tables, 12 lookup tables and 18 additional tables, plus 32 queries and 27 forms; it also includes more than 10,000 lines of Visual Basic code. The 7 linked tables form the main part of the database and contain a total of 135 fields (fields are included for all data that

might be present on the index cards). These tables are linked by the unique reference number (the 'card number') assigned to each card image when the images were created. The tables include one for the names of the card image files plus their paths, one for bibliographic references, one for type specimen information, one for details about the type species of genus-group names, and one for published name combinations other than the original and the currently valid combinations of the name. The structure of the database and the layout of the front-end were constructed to meet specific taxonomic requirements. It incorporates, therefore, specialist knowledge of taxonomic protocols and demanded a thorough assessment of the structure and function of existing taxonomic databases. The main purpose of the database is to enable quick visual comparison of the type- or hand-written data on the card images with data generated by OCR analysis of these images and to allow these data to be edited. The database was designed to provide an electronic substitute for the card index it replaces, and is now being made available to taxonomists in the NHM Entomology Department via the local intranet. A Web interface for the database has also been developed (see below).

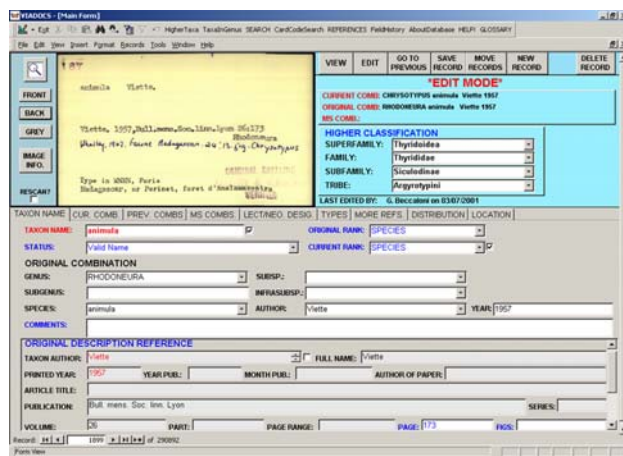


Figure 3. Access form entry for reviewing/editing archive card data after document analysis and OCR.

The main database form (Fig. 3) allows users quickly to find a card image, and the associated data, using a variety of search options (e.g. a drill-down search by higher classification and a 'simple search', with or without wildcards, for any taxon name). Authorised users are able to edit, delete and create new records. They can also 'move' records, singly or in batches, to new relative positions within the record sequence (e.g. in cases where the user wishes to transfer a species name from one genus to another). All changes made to data in the database are recorded in a set of archive tables. These tables store the old and the new field values, the name of the user, and the

date and time of the change. Deleted records are also archived, and the user name, date and time are recorded. Users can validate information in all except memo fields, by placing the cursor in the appropriate field and double clicking the left hand mouse button. The value currently stored in the field, plus the user name, date and time are recorded. If a field containing validated data is double clicked subsequently, then the validated data is displayed on a pop-up form and the user is given the option of deleting the stored validation information or overwriting it with a new validation record.

5. Web interface

A Web interface for the Access database (the "NHM online card archive") has also been developed and can be viewed at <http://www.nhm.ac.uk/entomology/lepindex/> (see Fig. 4). This interface allows users to search for records using a variety of search systems e.g. a simple search by scientific name, and an advanced search using a combination of a number of different search terms. The results page displaying a record is laid out in a similar way to the main form of the Access database (i.e. Fig. 3), except that related groups of fields (e.g. the fields which comprise a reference) are concatenated to aid readability (Fig. 4). Although users cannot add entire new records, they are able to edit existing ones. If they do so, the user's details and suggested changes are sent to the NHM server and stored in Access tables until the administrator of the system decides whether or not to include the changes in the master Access database. The Web interface operates using copies of the tables from the master Access database and these are updated periodically.

6. Evaluation

Evaluation was carried out separately on each part of the system described above.

The 290,886 cards in the full Lepidoptera index were scanned using the modified scanner in a total of 61 person days - an average of about 10 cards/minute (compared with the raw read rate of 60 cards/minute) due to the overhead of collecting, transferring and returning cards from the drawers to the small 40-card hopper of the scanner (larger hoppers are available for commercial systems, but our project budget did not stretch to buying one of these). The full archive of JPEG images requires around 30Gb of storage - well within the capacity of current PCs. Card images can be stored and downloaded about five times more efficiently in DjVu format[ix] if the DjVu plug-in to a web-browser is also installed, but the scanner can only capture images in raw or JPEG formats,

and JPEG would still be required for legacy users, so we have not implemented this option at present.

Work on improving archive card image analysis is continuing, but recent evaluation on a test set of 2000 archive cards shows that segmentation and labelling rates for individual fields of Pyralid moth cards are currently 91-95% correct (Species/Genus – 94%; Author – 91%; Reference – 92%; Locality – 95%). Overall, more than 85% of the testset archive images had all 4 fields correctly segmented and labelled. These cards are usually simple cards with few handwritten annotations, for which the OCR recognition rate (reported below) is close to 100% when full dictionary coverage is available.

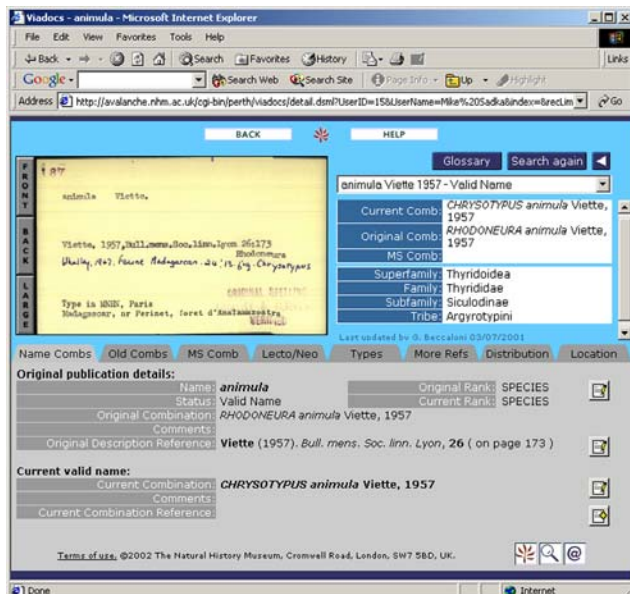


Figure 4. Web interface to the NHM Lepidoptera card archive database.

Results for the novel OCR system used to recognise the archive have been reported for two sets of test words (genus/species and author names)[7,8,x]. On a test set of 4468 typed word images of genus/species names using a genus/species dictionary with 100% coverage, recognition rates of 99.5-99.9% were achieved depending on parameterisation of the algorithm. On a later set of 1624 author name word images with 100% dictionary coverage, a similar result of 99.9% recognition was obtained (only 1 error was encountered, where the OCR result “Monroe” was obtained instead of “Munroe” – the correct word was ranked second). These excellent results were obtained by an algorithm which optimised performance at the expense of speed, typically taking around 1 minute to recognise each word on a moderate performance PC, but since the OCR process is run offline and can be parallelised as a web service, this approach is acceptable. Interestingly, our algorithm was quite resilient

in correctly recognising authors’ names with accented characters that were not encountered during training (the algorithm was not trained using any character data extracted from author names, only from genus/species names which are Latinized and contain no accents).

Initial results for the full end-to-end document analysis-OCR-database entry system are currently much lower than might be inferred from the figures above, typically around 50% overall recognition rate. Partly these results are due to using an earlier, poorer version of the document analysis system to segment word images (the system has not yet been re-tested end-to-end with the latest version reported above). More significantly, the end-to-end tests used an incomplete author dictionary and have also highlighted the presence in card images of many author name abbreviations which are absent from the author dictionaries. The recognition rate for the date field is also currently poor as no special attention has yet been given in the image analysis to extracting the date from the reference field within which it is normally embedded without spaces. Rapid improvements in the recognition rates for these different fields are expected over the next few months as the system is tuned and additional tools and rules added to the document analysis and OCR systems.

Assessment of the overall effectiveness of the new system compared with manual retyping of card data is more difficult, as direct comparators are not readily available. In an earlier project, the Geometridae archive, containing 42,503 cards, was manually converted from typed cards to electronic database over a period of 18 months. Similarly the VIADOCS project has included around 3 months spent in manually-assisted conversion of the 27,578 Pyralid moth archive cards, using early versions of some of the support tools described above during their development phase. However the process has not purely been a transcription process, as major revisions and updates to the taxonomic structure of the database (which are only possible using an electronic database) have also been undertaken as an important component of the research.

Currently, the image analysis and OCR tools are being used offline to parse and recognise the content of the remaining 200k+ Lepidoptera cards, and the results of this process will be inserted direct into the database without further editing (though the process will probably be repeated several times as improved tools evolve). On the basis of the statistics reported above we expect that more than 90% of those fields for which lexical source data is available will eventually be correctly inserted, making nearly all cards at least partially searchable online without requiring any human transcription. Further validation will take place thereafter as the database is used, so that those parts that are frequently referenced

will rapidly be verified, while user effort will not be expended on inactive parts of the archive.

7. Discussion and Conclusions

A great many institutions world-wide which house collections of objects (e.g. museums, herbaria and libraries), still use card indexes as a historical record of data about the objects. Although they are aware of the many advantages of computerisation, the cost of keyboarding all the data into database form is usually prohibitive, and standard OCR products are not designed to handle the specialist lexicons, idiosyncratic layout, poor quality typescript and database interface requirements of these indexes. The cost of computerising such archives could, however, be significantly reduced if a generalised version of the system described in this paper were available; especially if only one or a few keywords (rather than all the data from the cards) were required to index the card images. An important feature of the system is that images of both sides of the original card are included, making access to original data sources available wherever the Internet is connected, and allowing incremental online validation of the archive.

It is unlikely that systems of this type will ever represent a major commercial activity, as each application

requires specialist academic input from users and taxonomists to supply technical details of required database structure and field-specific lexicons. Furthermore, each conversion is a 'one-off' solving a specific scientific/taxonomic problem which does not then recur, discouraging the establishment of a steady-state business system. Nevertheless, our research has shown that by combining a configurable set of document image analysis and OCR tools with flexible relational database designs and web access, card archives can be made available online much more efficiently than by traditional retyping.

Our current work is aimed at extending the scope of the VIADOCS tools to enable them to be applied efficiently to a much wider range of archive conversion problems.

Acknowledgment

The VIADOCS project is sponsored by EPSRC and BBSRC as part of the UK research councils Bioinformatics research programme, under research contracts 84/BIO11933 and 40/BIO11938

References

-
- i Poole, R. W. 1989. *Lepidopterorum Catalogus* (New Series) Edited by J.B. Heppner. Fascicle 118 Noctuidae Parts 1-3: 1314 pp]. E.J. Brill / Flora & Fauna Publications. Leiden - New York - Kobenhavn - Koln.
 - ii Scoble, M. J. (Ed.) 1999. *Geometrid Moths of the World : A Catalogue*. Volumes 1 and 2: 1016 pp. + index 129 pp. CSIRO Publishing, Canberra.
 - iii Pitkin, B. R. & Jenkins, P. 2002. *Butterflies & Moths of the World: Generic Names & their Type-species*. <http://www.nhm.ac.uk/entomology/butmoth/>
 - iv Downton, A.C., A.C. Tams, G.J. Wells, A.C. Holmes, S.M. Lucas, M. Scoble, G. Robinson and G. Beccaloni, *Constructing Web-Based Legacy Index Card Archives – Architectural Design Issues and Initial Data Acquisition* Proc. ICDAR2001 6th Int. Conf. on Document Analysis and Recognition, Seattle, September 10-13, 2001, pp. 854-858.
 - v C. Cracknell and A. C. Downton, *TABS – script-based software framework for research in image processing, analysis and understanding*, IEE Proc. VISIP, v.145 No. 3, pp. 194-202 June 1998.
 - vi He, J and A.C. Downton, *User-Assisted Archive Document Image Analysis for Digital Library Construction*, to appear at ICDAR2003
 - vii E. Ishidera, S. M. Lucas and A. C. Downton, *Likelihood word image generation model for word recognition*, Proc. 16th International Conference on Pattern Recognition (ICPR2002), August 11-15 2002, Quebec City, pp. 30172-30175.
 - viii E. Ishidera, S. M. Lucas and A. C. Downton, *Top-down likelihood word image generation model for holistic word recognition*, Proc. DAS'02 Document Analysis Symposium, August 19-21, Princeton, NJ, Springer-Verlag LNCS 2423 pp. 82-94.
 - ix P. Haffner, L. Bottou, P. G. Howard, Y. LeCun, *DjVu: Analyzing and Compressing Scanned Documents for Internet Distribution*, Proc. ICDAR1999 5th Int. Conf. on Document Analysis and Recognition, Bangalore, India, Sept 20-22 1999, pp. 625-628
 - x E. Ishidera, S. M. Lucas, A. C. Downton and G. Patoulas *Likelihood word image generation model for holistic word recognition*, submitted to IEEE Trans PAMI.