10 Computerising unit-level data in natural history card archives

George W. Beccaloni, Malcolm J. Scoble, Gaden S. Robinson, Andrew C. Downton and Simon M. Lucas

Introduction

Typically, unit-level data take the form of information on specimen labels. However, many organisations with natural history collections house large archives of data that have been compiled on index cards. Before the advent of computerised databases, the card index was one of the major ways by which such organisations used to catalogue information – especially information that was subject to change over time and which needed to be kept up-to-date. The data in these archives may pertain to specimens, species, or other taxa, and may include information such as nomenclature, type status, type locality, taxonomic history and distribution. Whatever natural history data is held, and usually these are of several kinds within a single archive, each individual card can be regarded as the equivalent of a specimen in a collection. Since these vulnerable archives often represent unique datasets constructed, in many cases, over decades, curators have a responsibility to ensure the protection of data within them. Curators should also aim to improve access to the information held on these cards, which at present is largely trapped within organisations. Fortunately, improving both data protection and data access are highly compatible goals and have overlapping technical requirements.

Although card index archives have served the natural history community well in the past (and continue to do so often), they have many disadvantages when compared with computer databases. Cards are all too easy to misfile, misplace or lose. They become degraded with use – characters may fade and cards get torn or bent. Handwriting on cards is often illegible, and archives are difficult to copy. Adding data to existing cards or making corrections can prove a problem (e.g. because of lack of space), and maintaining consistency of data is not straightforward (e.g. if abbreviations are used for journal names). Re-ordering a large index (e.g. a taxonomic index might need to be rearranged so that it agrees with a newly published higher classification) is a massive task, and indexes take up a large amount of physical storage space.

Computerisation of card indexes solves, or at least eases, most of these limitations, although long-term protection of electronic data remains a significant concern. However, given that the many advantages of digitisation outweigh the few possible disadvantages, there are several approaches and levels by which data in card indexes can be computerised. The most straightforward of these is to enter (keyboard) the information manually into a database. Such an undertaking might be carried out contractually, and if this is to be done off-site, then it will usually be necessary to make a copy of the archive, which can be sent off-site instead of the original. Ideally data from the cards should be entered into a database twice (or even more), so that the different versions of the same dataset can be compared and transcription errors detected.

If one of the advantages of computerising an archive were to be a saving of space, with the cards being subsequently disposed of or placed into storage, then the original archive would be either unavailable to users or difficult to access. This could be a problem if, for example, doubt arose as to whether information in the database had been transcribed correctly from the original cards. Such a limitation is overcome if digital images of the cards are incorporated into the database, together with the minimum amount of data necessary to index the images (e.g. the scientific names of the organisms). A superior system would be to have a database with images of all the cards, plus fields containing all the information contained on them.

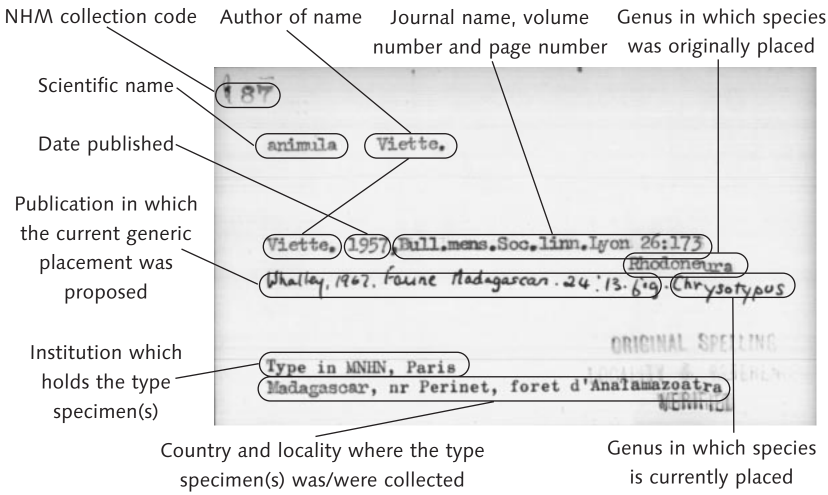
In this chapter, we describe a project to develop the latter type of system for a specific card archive housed in the Natural History Museum, London (NHM). This project is called *VIADOCS* (Versatile Interactive Archive Document Conversion System) – for further information see Downton *et al.* (2001) and the project websites at www.essex.ac.uk/ese/research/vasa/viadocs/ and avalanche.nhm.ac.uk/cgi-bin/perth/viadocs/

The NHM Lepidoptera card index

In addition to its 68 million natural history specimens, the NHM houses comprehensive card index archives of taxonomic data for many groups of organisms, extant and extinct. These archives augment the NHM's specimen-based collections, enhancing their value, and serve as an indispensable resource for biodiversity studies.

One such index is that to the world species of butterflies and moths (Lepidoptera), which is housed in the Department of Entomology. In effect, this card archive represents a comprehensive inventory of the scientific names and associated bibliographic data for the whole of this species-rich order of insects. Although taxonomic catalogues have been published for many groups of Lepidoptera, no complete global catalogue of this group exists. The main part of the index consists of 265 drawers containing 290,886 index cards. Each 5″ x 3″ (127mm x 76mm) 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 typewritten, but very many cards contain hand-written annotations, and a significant minority of them are entirely hand-written.

Figure 1. A representative index card for a species name from the NHM systematic index to Lepidoptera names, showing how text on the card relates to fields in the Access database.



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. Cards with names that are nomenclaturally unavailable (e.g. infrasubspecific names, junior homonyms, junior objective synonyms), or names that are not in use (e.g. junior subjective synonyms), are arranged alphabetically by scientific name following the card containing the name currently used for the taxon in question. This systematic index is supplemented by an alphabetic index, containing corresponding cards to all those in the systematic index. These alphabetic archives are generally arranged by superfamily. Each card contains a subset of the information recorded on the systematic index cards – the scientific name of a taxon, the author of this name, the date of publication, the current valid genus name, and some or all of the higher classification for the taxon. The main function of the alphabetic index is to enable cards to be located in the systematic index in cases where the current genus name or the higher classification is not known. Since this index duplicates information already present in the systematic index, it was not necessary to computerise it. Its existence, of course, highlights one of the access limitations of non-computerised archives.

The NHM Lepidoptera card index was compiled from a number of smaller indexes constructed by NHM Lepidopterists to assist with their work. These were combined into a single unit by staff of the Entomology Department's Indexing Section (headed initially by R.J. Collins), which operated between about 1955 and 1981. During this period of amalgamation, biological journals, taxonomic catalogues and the *Zoological Record* were scanned, new cards were added and old cards were retyped (Nye, 1975). Since 1981 certain parts of the index have been maintained and kept up to date by individual NHM Lepidopterists (e.g. pyraloid moths (i.e. Pyraloidea plus Alucitoidea, Hyblaeoidea, Pterophoroidea, Simaethistoidea, Thyridoidea and Whalleyanoidea) by M. Shaffer, Tineoidea by G. S. Robinson and Tortricoidea by K. Tuck). These workers have also incorporated much important unpublished information into the index, especially new genus/species combinations and new synonomies. The remainder of the index contains virtually all scientific names published up until 1981 when the Indexing Section was disbanded.

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 (Poole, 1989), Geometridae (Scoble, 1999), and the six volume *Generic Names of Moths of the World* (i.e. Nye (1975), Fletcher (1979), Watson, Fletcher & Nye (1980), Fletcher & Nye (1982), Fletcher & Nye (1984), and Nye & Fletcher (1991), an updated version of which is now available at www.nhm.ac.uk/entomology/butmoth/ (Pitkin & Jenkins, 2002)).

Archive card scanning

The first step in computerising the card index was to produce electronic images of the cards. We used a modified bank cheque scanner (SEAC Banche RDS-6000) attached to an IBM computer with an Intel Pentium III 1GHz processor and, importantly, a DVD writer to enable backup copies to be made of the card image files. Cards to be scanned are placed into a pocket in the scanner, about 40 at a time. When activated, the machine draws the cards, at a rate of about 1 card/second, through a scanning head and ejects them into another pocket, maintaining their original order. It produces colour jpeg images (1,000 X 600 pixels) of the front and back of each card, the storage size of which varies between about 30 KB and 50 KB, depending on the amount of text on the card. The scanner also prints a unique code (see below) onto the back of each card as it travels through the machine.

Since the software supplied with the scanner did not have all the functionality we required, a customised software interface (operating under Microsoft Windows 2000) was produced. At the beginning of each scanning session, the user enters the following information into this interface: the higher classification of the batch of cards to be scanned; the number of the index drawer from which the cards were taken; and the unique reference number of the first card in the batch (unless this has already been stored by the software – see below).

Using this information, the software creates a nested set of folders on the pc hard disk that mirrors the higher classification of the cards. It also uses the information to create names for the JPEG images and to create the unique code that the scanner prints onto the back of each card. 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 by one, 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.

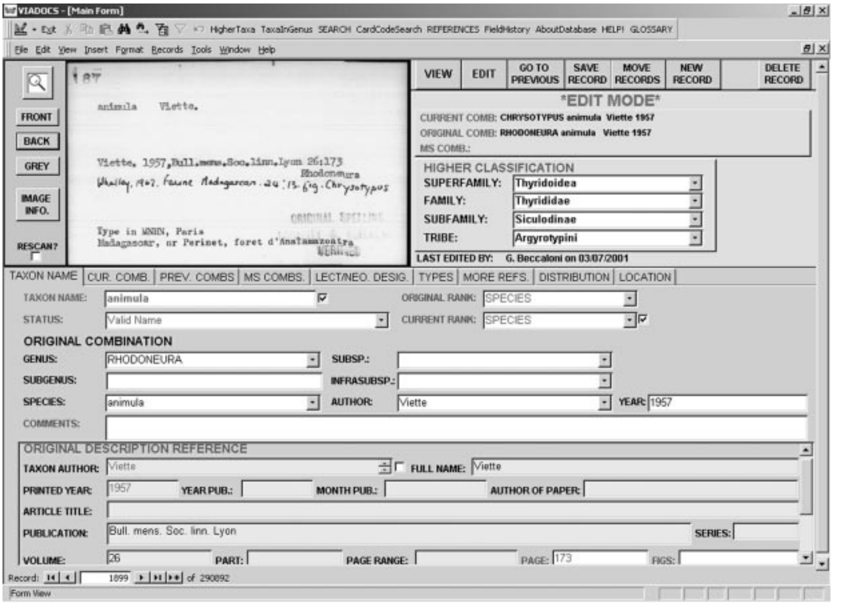
The index was scanned sequentially from beginning to end and tab cards dividing genera were ignored. Scanning all 290,886 cards in the index took a total of 61 person days – an average of 4,769 cards per person day of seven hours duration (excluding breaks).

Archive card database

A relational Microsoft Access 97 database was developed to manage the card images and associated taxonomic data (G. Beccaloni) over a period of about 18 months. It contains seven linked tables, 11 lookup tables and 18 additional tables, plus 32 queries and 27 forms, and it operates using over 10,000 lines of Visual Basic code. The seven 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 (see over) and to allow these data to be edited. The database was designed in such a way that it provides an electronic substitute for the card index it replaces, and it will be made available to scientists in the NHM Entomology Department via the local computer network (intranet). A Web interface for the database is also being developed (see below).

Once the entire index had been scanned, a 'freeware' programme ‘rjhextensions’ (www.rjhsoftware.com/rjhextensions/) was used to produce a text file containing a list of all the card image files, plus their directory paths. This list was then manipulated using Microsoft Word for Windows to produce two delimited text files – one listing card number/directory path/name of front image/name of back image for each card image, and the other listing card number/superfamily name/family name/subfamily name/tribe name for each image. These data were then imported into the appropriate tables in the Access database. Records in the database are sorted by a number that gives the current relative position of each record in the electronic index. This number is set initially to be the same as the card number, thus ensuring that the records are in the same order as the cards in the index at the time when the database is first populated with records.

Figure 2. The main data form of the VIADOCS Access database, displaying a record for a species name.



The main database form (Fig. 2) allows users to find quickly a card image, and its 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). Fields on this form are grouped onto different labelled tabs (Table 1), each of which contains a group of related fields. A slightly different set of tabs (and hence fields) is displayed according to whether the record is for a genus-group, or for a species-group/infrasubspecific name. Thus, the tab labelled 'TYPE SP. DESIG.' is only displayed for genus-group name records, while the tabs labelled 'LECT/NEO. DESIG.' and 'TYPES' are only displayed for species-group name records.

Table 1. Labels of the tabs on the main form of the VIADOCS Access database and details of the fields that are displayed on them.

Label of Tab	Fields on Tab
TAXON NAME	The availability and rank of the name, the original published combination of the name, and the reference to the publication in which the name was first proposed.
CUR. COMB.	The currently valid combination of the name and the reference in which this was proposed.
PREV. COMBS	Details of all previously published combinations of the name and the references in which these were published.
MS COMBS.	An unpublished combination of the name and the details of the person who proposed it.
TYPE SP. DESIG.	Details about the type species of a genus or subgenus plus the reference of the publication where the designation was made.
LECT/NEO. DESIG.	Information about the designation of lectotypes or neotypes of species-group taxa and the reference of the publication where the designation was made.
TYPES	Information about the type specimens of species-group taxa.
MORE REFS.	One or more references giving additional information related to the taxon (e.g. host-plant information, descriptions of early stages etc).
DISTRIBUTION	Distribution of the taxon (by zoogeographic region and country).
LOCATION	Location of specimens of the taxon in the NHM collection.

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 recordset (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.

Card image processing

Two initial pre-processing stages need to be performed on the card images before the text on them can be read by an optical character recognition (OCR) algorithm:

First, the text fields of the images must be parsed to identify which text field should be associated with each database field (scientific name, author, date of publication, etc.). Image analysis and processing techniques must also be used to extract the corresponding text image from the overall card image without distorting it by cutting off part of the required text image or including noise artefacts.

Second, once the full card image has been broken down to a set of text sub-images labelled with appropriate database fields, these sub-images must be presented to an OCR engine which produces a ranked set of candidate words matching each image. The OCR engine will normally operate using a dictionary of allowed words for each field to optimise its performance (e.g. a list of the possible scientific names which may be contained in the 'scientific name' field), though these will not always provide complete coverage of the possible text words.

Our iterative approach to implementing a practically useful system has been to focus first on the scientific names, which are found in the top left part of each card (see Fig. 1), because they represent the primary index term for searching the database. Therefore a simple image-analysis process was used to extract the scientific name as a sub-image, and this was used as the first dataset for testing the OCR component of the project (see below).

Currently a more sophisticated image-analysis tool, which attempts to extract all fields from the card images (rather than simply the scientific name) is being developed. Overall success rates achieved in an initial, fairly simplistic, implementation for correct text image extraction and labelling, range from 92% for the field giving the location of the taxon in the NHM collection, to 97% for the scientific name field, based on a sample of 2,000 card images. Further improvements are expected to be achieved with more sophisticated image analysis algorithms, which are currently being developed.

Optical character recognition

Experimental evaluation of the performance of commercial off-the-shelf OCR packages on sample card images showed that the error rate is unacceptable, due to touching characters, poor quality printing and the use of a specialist vocabulary. The following two algorithms were therefore investigated and tested as candidates for their ability to read the text from the card images:

1. An algorithm developed by S. M. Lucas (see Lucas *et al*, 2001).
2. An algorithm developed by Eiki Ishidera (with support from Gregory Patoulas) at Department of Electronic Systems Engineering, University of Essex.

A commercial algorithm designed by Parascript in the USA, generally considered to represent the state of the art for commercial offline handwriting recognition, was used for comparative evaluation against these algorithms.

Our initial system used the Lucas algorithm, and it was tested on the scientific name field of a subset of 27,578 card images (i.e. all the index cards relating to pyraloid moths). The system employed a dictionary covering about 60% of the scientific names on these cards, and it achieved a correct recognition rate of 37% of the names (compared with <10% for standard PC OCR packages). Of greatest interest at present is the algorithm developed for the project by Ishidera. A novel approach to OCR is employed, which is based upon constructing a probabilistic word image model of each possible word in a dictionary, and which incorporates image character templates, character segmentation information and linguistic knowledge. Each word image, once constructed, is used as a template against which archive word images are matched. The method works particularly well for the cards we scanned because these contain very poor quality text, but with a restricted range of typed character fonts. Exceptionally good recognition performance has been obtained with this algorithm on test sets consisting of 4,498 scientific names (over 99% recognition rate) and 1977 author names (over 97% recognition rate), compared without about 90% recognition rate for the algorithm developed by Lucas, and under 80% for the Parascript algorithm (which, however, is optimised for handwriting rather than print).

The main limitation of the Ishidera algorithm is that it is computationally extremely intensive, currently taking more than one minute to match each word image on a typical PC. However, research associated with the project has been investigating web-based deployment of OCR engines as a standardised mechanism by which more computing resource can be brought to bear on this problem.

Once OCR results have been obtained, they are imported into the appropriate fields of the Access database described above. They then are checked visually against these same data on the card images and corrections made where appropriate. This procedure saves much time compared with the alternative approach of simply typing all the values into the database, especially considering that data entered manually will usually also need to be verified. In some cases the data obtained from OCR can be validated against existing databases, leaving only a residue of data that needs to be checked visually. For example, in the case of our card archive, OCR results for the fields 'scientific name' + 'author name' + 'date of publication' for all genus-group names can be validated electronically against these field combinations in the comprehensive and accurate *Butterflies & Moths of the World: Generic Names & their Type-species* database hosted by the NHM (see www.nhm.ac.uk/entomology/butmoth/).

To date, OCR results for the scientific name field of the 27,578 pyraloid moth cards have been checked visually against the card images and corrected where necessary. Once these data had been validated, Visual Basic algorithms were written to compute the values of certain other fields in the Access database and fill them in. For example, an algorithm was written which identifies currently valid generic names and copies these to the 'current genus' field for all the records. It was possible to identify current valid generic names by the arrangement of the records in the database, plus the fact that genus-group names are in capital letters (other scientific names are in lower case).

Web interface

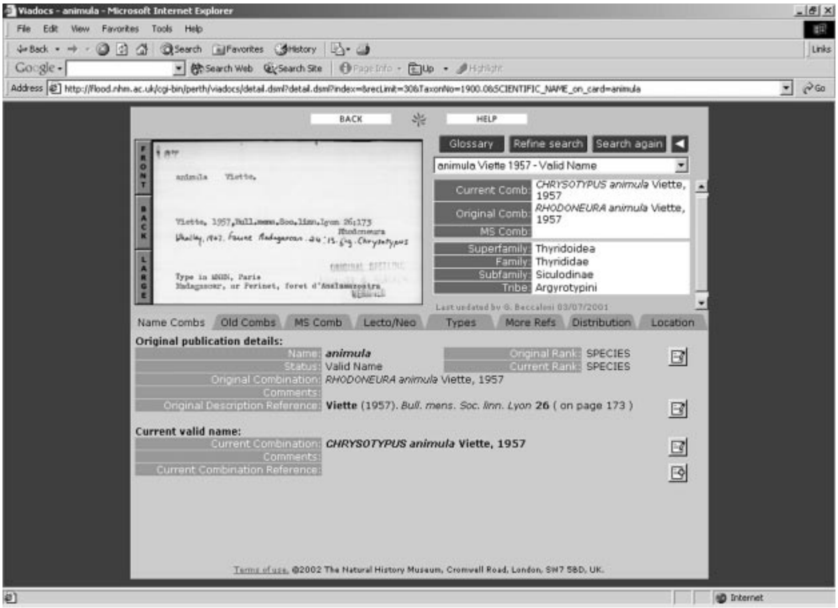


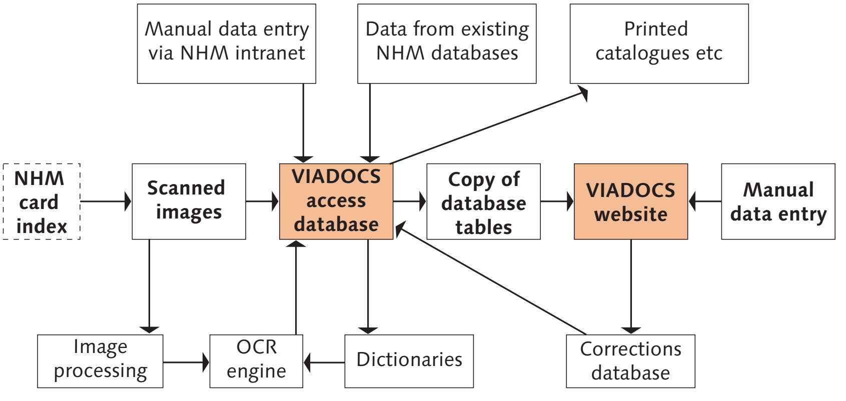
Figure 3. A record for a species name displayed on the NHM online card archive.

A web interface for the Access database (the NHM Online Card Archive) has been developed by M. Sadka and G. Beccaloni at the NHM and can be viewed at www.nhm.ac.uk/entomology/lepindex/ (also see Fig. 3). 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, which displays a record, is laid out in a similar way to the main form of the Access database (i.e. Fig. 2), except that related groups of fields (e.g. the fields that comprise a reference) are concatenated to aid readability (Fig. 3). Although users cannot add, delete or move records into new relative positions in the recordset, they are able to edit existing ones. If they do so, the user's details and the 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.

Discussion

A great many institutions with collections of objects (e.g. museums, herbaria and libraries) not only hold extensive legacy data in card indexes, but continue to record data about the objects themselves on cards and use the indexes to facilitate locating objects within their collections. 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 reduced significantly if the system described in this chapter, or one like it, were used (see Fig. 4 for a summary), 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 this 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.

Figure 4. Overall structure of the VIADOCS archive conversion system.



Acknowledgements

We thank Brian Pitkin for helping to produce the dictionaries for the OCR engine as well as assistance with many diverse IT problems; Jade Spence for voluntary databasing work; Clare Stringer for scanning cards and checking data; Michael Shaffer for help and advice on pyraloid nomenclature and information about the pyraloid card index; Mike Sadka for programming the Web interface for the card archive and help with Access; Arran Holmes for programming the scanner interface; Eiki Ishidera for developing the OCR algorithm discussed above; and Andy Tams and Grigoris Patoulas for their work on the OCR analysis and other aspects of the project. We are grateful for financial support from the BBSRC/EPSRC Bioinformatics Initiative (grant references 40/BIO11938 and 84/BIO11933).

References

Downton, A. C., Tams, A. C., Wells, G. J., Holmes, A. C., Lucas, S. M., Scoble, M., Robinson, G. & Beccaloni, G. 2001. Constructing web-based legacy index card archives – architectural design issues and initial data acquisition. *Proceedings ICDAR2001 6th International Conference on Document Analysis and Recognition, Seattle, September 10–13, 2001:* 854–858.

Fletcher, D. S. In: Nye, I. W. B. (Ed.) 1979. *Generic Names of Moths of the World.* Volume 3: xx + 243 pp. Trustees of the British Museum (Natural History), London.

Fletcher, D. S. & Nye, I. W. B. In: Nye, I. W. B. (Ed.) 1982. *Generic Names of Moths of the World.* Volume 4: xiv + 192 pp. Trustees of the British Museum (Natural History), London.

Fletcher, D. S. & Nye, I. W. B. 1984. *Generic Names of Moths of the World.* Volume 5: xv + 185 pp. Trustees of the British Museum (Natural History), London.

Lucas, S. M., Tams, A. C., Cho, S. J., Ryu, S. & Downton, A. C. 2001. Robust word recognition for museum archive card indexing. *Proceedings ICDAR2001 6th International Conference on Document Analysis and Recognition, Seattle, September 10-13, 2001:* 144–148.

Nye, I. W. B. 1975. *Generic Names of Moths of the World* Volume 1: 568 pp. Trustees of the British Museum (Natural History), London.

Nye, I. W. B. & Fletcher, D. S. 1991. *Generic Names of Moths of the World.* Volume 6: xxix + 368 pp. Trustees of the British Museum (Natural History), London.

Pitkin, B. R. & Jenkins, P. 2002. *Butterflies & Moths of the World: Generic Names & their Type-species.* (www.nhm.ac.uk/entomology/butmoth/)

Poole, R. W. In: Heppner, J. B. (Ed.) 1989. *Lepidopterorum Catalogu*s (New Series). Fascicle 118. Noctuidae. Parts 1–3: 1314 pp. E.J. Brill / Flora & Fauna Publications. Leiden – New York – Kobenhavn – Koln.

Scoble, M. J. (Ed.) 1999. *Geometrid Moths of the World : A Catalogue.* Volumes 1 and 2: 1,016 pp. + index 129 pp. CSIRO Publishing, Canberra.

Watson, A., Fletcher, D. S. & Nye, I. W. B. In: Nye, I. W. B. (Ed.) 1980. *Generic Names of Moths of the World.* Volume 2: xiii + 228 pp. Trustees of the British Museum (Natural History), London.