

User-Assisted Archive Document Image Analysis for Digital Library Construction

J.He and A.C. Downton,
Multimedia Architectures Laboratory,
Department of Electronic Systems Engineering, University of Essex,
Colchester, CO4 3SQ, UK
{jhe,acd}@essex.ac.uk

Abstract

A configurable archive document image analysis system for digital library construction has been designed using rapid prototyping and top-down iterative development methods. This approach has been found to be essential in order to capture the curators' expertise about existing card archive structures, content and databases. The design currently achieves about 93% correct segmentation of the required archive card fields overall, with 81.3% of all archive cards in a testset of 2000 images having all fields correctly segmented and labelled. Analysis of errors in the testset indicates that heavily-annotated cards and non-standard card formats comprise 5-10% of the overall archive, and a significant proportion of these are unlikely to be resolvable without curatorial intervention.

1. Introduction

Archive document image conversion for online digital libraries is a major application area of interest both for cultural and scientific purposes, particularly where current and historical data need to be compared (e.g. for biodiversity or climate change monitoring). Archive documents are often stored in well-structured taxonomies (e.g. libraries, scientific specimen indexes and censuses) where the structure extends across the index as well as being present within the layout of each document. Often however, individual documents are recorded using methods which continue to challenge OCR, because poor quality and/or decayed typescript or handwriting have been used to record data, over decades or even hundreds of years.

In such cases it is imperative to utilise all available contextual information to maximise the potential quality of the OCR. Although some context can be inferred automatically using established document image analysis techniques, much is known only to the curators of the archive, and indeed may not initially be available in electronic form (e.g. field-specific dictionaries to be applied during the OCR process). Furthermore, the available context can only be associated with appropriate

document image fields by a user-assisted field labelling process, even if the fields themselves can be extracted automatically. Therefore we concluded that, in order to automate archive conversion, a user-assisted image analysis process is required, which must be carried out by curators rather than computer scientists, and which makes no initial assumptions about the document archive format. Thus, general document image analysis processes need to be set up by the curators during an initial configuration phase using a sample of the scanned archive, and once satisfactory performance has been achieved, the tool is run to completion on the full archive, yielding a set of segmented image sub-fields for each archive document comprising words or phrases to be applied to the OCR system, with identified database attributes for subsequent input to the final online database.

1.1 NHM biological archives

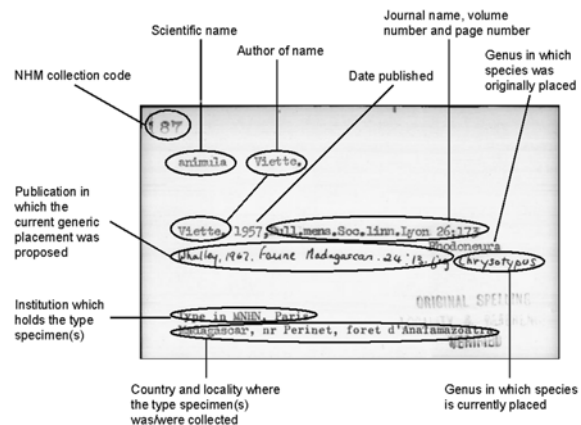


Figure 1. An index card with multiple hand print and handwriting annotations showing components to be extracted.

Archives at the UK Natural History Museum (NHM) are recorded in card indexes, which contain bibliographical data and other information for one scientific name on each card (genus-group, species-group, and often infrasubspecific), laid out in a standardised format (Fig. 1) for each archive. However, different

archives may be subject to very different recording conventions. Information is usually type-written, but a significant minority of cards are entirely hand-written and hand-written annotations are common. Some archives are entirely handwritten (e.g. Fig. 2, which includes diagrams and handwritten notes on the reverse of the index card).

Cards are ordered within each 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 card sequence implies database fields which are not explicitly included on the cards). Cards with names that are no longer in use (e.g. synonyms of current species names) are arranged alphabetically by scientific name following the current taxon card. 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, and potentially providing an independent means of validating OCR scanned data.

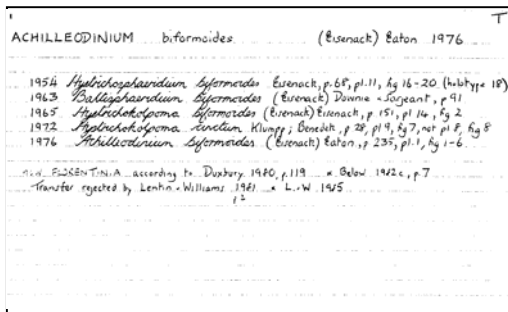


Figure 2. An index card from an entirely handwritten archive.

1.2 System framework

Our interactive document image analysis system has been built around an existing tool, HUE [1,2], which was developed during earlier research into toolkits for document image analysis. The HUE components deal with pre-processing (e.g. threshold binarization and noise filtering).

1.3 Paper structure

In this paper we describe the structure of our interactive document image analysis tool by describing its document image analysis capabilities, and explaining its Graphical User Interface (GUI), and the way in which this is used in combination with the document image analysis capabilities to configure the overall system to process the

particular type of archive documents present in the NHM archives. We then present an evaluation of the current performance of the system, and discuss how it can be further improved and extended to handle other archives, before drawing conclusions as to the potential of this type of tool.

2. Archive Card Document Analysis

2.1 System overview

The Archive Card Document Analysis System is designed for archive curators to carry out document analysis (content extraction) on archive images. The system has an integrated user interface (see Fig. 3), which provides a simplified conceptual model of the image analysis currently configured, and allows users to configure processing operations and set up parameters according to the characteristics of the archive to be processed.

The key requirement of the GUI is to present an understandable model of the document analysis process to curators who use the system for setting up archive conversions on their archives. Curators are usually extremely knowledgeable about the content and organisation of card archives, but relatively unfamiliar with the capabilities of document analysis technology. Thus much of our investigation is concerned with finding ways of presenting conceptual models of document image analysis processes which match curators' cognitive model of the taxonomic structure of scientific archives.

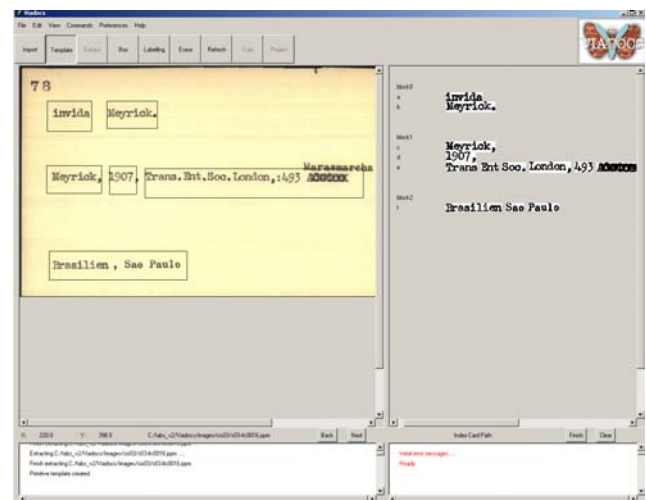


Figure 3. Archive Card Document Analysis System user interface

The core of the system consists of two processes, Template Creation and Matching. Fig. 4 shows the

system flow chart of how these mechanisms are applied.

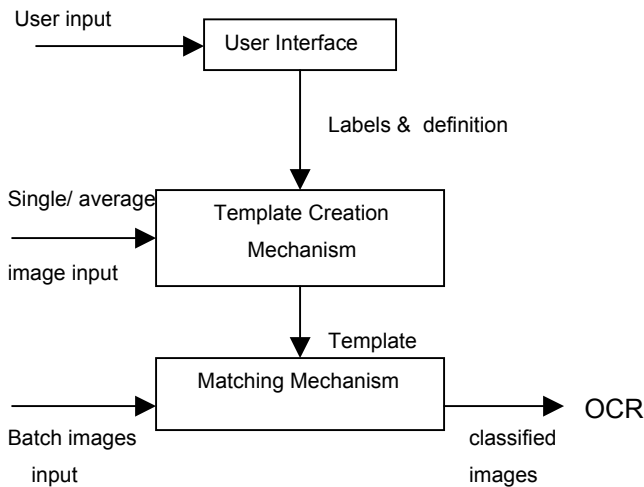


Figure 4. System flow chart

2.2 Template creation

Since each document in a specific archive will have a similar format (or sometimes one of several formats), the objective is to create one (or several) standard layouts which can be used as a template(s) for processing each document in the archive. This template creation mechanism is a process of mapping a physical image hierarchy into its relevant logical structure. The idea of this mechanism is to get logical knowledge from human users and apply it to the physical hierarchy obtained from mathematical segmentation algorithms. The mechanism has two steps, *primitive template creation* and *template intensive definition*. Both steps require user interaction through the interface.

The template creation process attempts to reconcile the curator's cognitive model of the text fields in a particular physical archive document with a more sophisticated logical hierarchical document structure derived from adaptive automated document segmentation. Initially, the curator uses a rubber band box to identify and label each text field of interest on a sample card (Fig. 3) or an average image of a group of sample cards (Fig. 5). Text fields may be single words (such as a species name) or a group of words (such as a reference) and thus do not necessarily map directly to the components extracted by the automated segmentation process.

An X-Y cuts algorithm [3,4] separately extracts and stores the contents of each target card into a hierarchical tree structure (the so-called X-Y tree). The format of archive cards consists of several independent blocks of text, and each block consists of one or more logically-related fixed text field as shown in Fig. 1. Blocks retain a fairly consistent mutual layout over a complete archive (e.g. see Fig.5), but the layout of text fields within each

block is not strictly fixed, nor are there any tabular guidelines defining fixed block boundaries. The X-Y cuts algorithm therefore seems an appropriate segmentation algorithm for this kind of adaptive top-down document image analysis.



Figure 5. Example 'average' image from a group of sample cards

Pixel smearing is used to help define the X-Y cuts, which have three levels, block, line and word. The level of the cut is dependent on the white space: conventionally, the space between blocks is greater than between lines and words. The first level cut thus separates blocks (which may contain several lines of text) based upon their spacing, the second level segments lines vertically, and the last level cut separates words horizontally within each line, using binary connected component bounding boxes [5]. The extracted contents stored in the X-Y tree thus follow a sequence where the top level of the tree stores blocks while the bottom level stores individual words.

By comparing the location of each labelled text field with that of the corresponding extracted components in the X-Y tree, all bottom-level nodes (words) of the X-Y tree that are of interest can be labelled. We can then easily find the corresponding block for each labelled word, hence decide how many fields each block contains, and hence the primitive mapped physical hierarchy, which is the so-called primitive template.

2.3 Template finalization and card matching

Once the primitive template is created, the image representations, which are word images and their labels and the name of the block in which each labelled image is contained, are shown on the interface (see Fig. 3). The curator can further edit properties of the representations, e.g. names of labels, sequence and number of text fields in a block and so on. Once all relationships between the logical and generic physical card structures are defined, the final template is complete. The template contains information including the relationship between text fields

and blocks, lines and individual extracted word images, labels for all text fields, and the fuzzy location of each component. The fuzzy location for blocks is global to the target image, and is represented by using a range from 0 to 1. For example, if a block is located on the top-left of the target image, it could be expressed as FP(0.1,0.1). The fuzzy location for low level components like words is local to the block with which the word is associated. For instance, if a word is located on the bottom-left of the block, it could be expressed as FP(0.1,0.9).

The matching mechanism applied to the card archive once the template is finalised is as follows. Firstly, the block, line and word content of each card is extracted by using the same X-Y cut method as for template creation outlined above. Secondly, the minimum mean square error is calculated by comparing each extracted high level (block or line) component's fuzzy position with that of the template's text fields. Finally, based on the located high level components, low level words are matched using both mean square error calculation to text fields and component relation constraints from the template.

2.4 Cross-card archive taxonomy

Not all aspects of the extracted database are derived directly from archive card document analysis. The highest levels of definition of the species taxonomy are defined by the order of cards within the archive and are specified by periodic divider cards. As this data is quite limited, it is most conveniently entered during the scanning process itself, by taking advantage of the fact that cards are scanned in strict index sequence [6]. During database construction, each card database record then retains the same higher-level data (superfamily, family, sub-family and tribe) as its predecessors until a new name is encountered.

3. Evaluation

The current system has been constructed in two stages. Initially a 'hard-programmed' archive document analysis system was developed specifically for NHM archive card conversion, similar to many other document analysis systems reported in the literature. However it quickly became apparent that a much more flexible tool was required for archive document analysis, capable of being reconfigured to analyse a wide range of different archive formats by someone other than the programmer who designed it. While the programmer has the technical capability to set up the document analysis software to suit a specified archive, they seldom have the taxonomic knowledge or domain expertise required. Conversely, the archive curator is usually very knowledgeable about the archive structure and content, but is unlikely to be able to configure the analysis tools at a programmer level.

Thus the second phase of design of this archive document image analysis tool was to attempt to reconcile the underlying algorithmic structure of the document analysis process with a graphical user interface which allows the curator to visualise the analysis in terms of his/her own cognitive model of the archive structure. This is an iterative process where ideas are implemented and then demonstrated to staff at the Natural History Museum before being finalised as part of the system design.

The current version of the system has been evaluated on a sample of 2000 cards randomly chosen from the Pyraloidea dataset of 27,578 archive cards. At present, the text fields extracted from each archive card are: genus/species name and author name (top text block); reference (middle block) and locality (bottom block). The reference line is then further sub-divided into a second instance of the author name, the date published and the reference itself consisting of journal name or abbreviation, volume number and page number. Evaluation was carried out using a semi-automated graphical tool written using Tcl tk which presented each archive card image and its corresponding segmentation and label to the researcher, who then confirmed whether these were correct, and flagged the image for further analysis if not. Current results for the 2000 image sample testset are:

- Genus/species name – 94% correctly segmented and labelled;
- Author – 91% correct segmentation/labelling
- Reference line as a whole – 92% correct segmentation/labelling
- Locality – 95% correct segmentation/labelling
- Overall – 81.3% of cards currently have all the above fields correctly segmented and labelled.

The reference line is also sub-divided to extract the date published and reference, but tools for separating continuous text have not yet been added to the system, and as a result only about half the dates and references are correctly segmented.

Analysis of segmentation errors in the evaluation dataset shows that by far the majority (60%) occur in cards with complex layouts, caused by combinations of machine typed text with handwritten annotations and touching lines, and additional content such as corrections and stamps that change the card format (Fig. 6 shows an example). In these cases, the corrections and annotations usually update the scientific content of the card, so that recognition of the original typed content is insufficient to generate a correct database entry.

A further 25% of errors result from poor contrast on the index card, resulting from faded typing or annotations, resulting in incorrect segmentation. Improvements in this area could be achieved by better preprocessing of the image to enhance contrast and improve text thresholding.

A few errors (5%) result from non-standard card formats which can probably only be handled manually; the remainder (20%) require further debugging to clarify their cause.

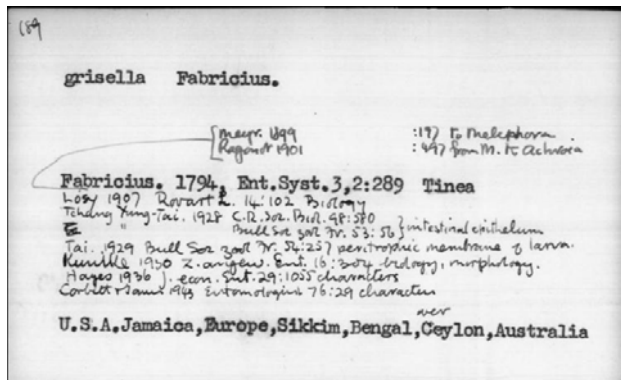


Figure 6. An index card with multiple handwriting annotations.

4. Further work

The system currently reported is an early prototype interactive document analysis tool which is being built in a top-down fashion to address a specific scientific need. At present, image analysis tools are being incrementally added specifically to address the requirements for converting archive card indexes found in the UK Natural History Museum and other similar environments. By evaluating performance as each component is upgraded the overall system performance will be optimised. There are many relatively straightforward enhancements remaining to be added to the current system, but a law of diminishing returns operates, since some cards have such extensive annotation that they will inevitably be incorrectly segmented. At present the most obvious segmentation weakness is in separating logically independent fields within continuous horizontal text using cues such as punctuation symbols.

In the longer term, the aim is to extend the system by the application of fuzzy logic to handle a much wider range of archive data formats. For example the University of Essex houses some of the most extensive handwritten and typed social science archives in the UK, many of which provide critical core data for chronological analyses of social change over the last two centuries.

5. Conclusions

Rapid prototyping and top-down iterative development have been used as a method for designing an interactive and configurable archive document image analysis system for digital library construction. This approach has been found to be essential in order to capture the curators' expertise about existing card archive

structures, content and databases. Although the design is still at an early stage, with many obvious refinements yet to be added, it already achieves about 93% correct segmentation of the required fields overall, with 81.3% of all archive cards in a testset of 2000 images having all fields correctly segmented and labelled. We expect to improve on these figures significantly over the next six months, before generalising the system to handle a much wider range of archives in the longer term. Analysis of the segmentation and labelling errors in the testset indicates that heavily-annotated cards and non-standard card formats comprise 5-10% of the overall archive, and a significant proportion of these are unlikely to be resolvable using the existing image analysis system without curatorial intervention.

6. Acknowledgment

This work 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

- 1 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.
- 2 C. Cracknell and A. C. Downton, 'A Handwriting Understanding Environment (HUE) for rapid prototyping in handwriting and document analysis research', Proc. ICDAR'99 5th Int. Conf. on Document Analysis and Recognition, Bangalore, India, September 1999, pp.362-365.
- 3 Nagy, G., and S.Seth, 'Hierarchical Representation of Optically Scanned Documents,' Proc. 10th Int'l Conf. Pattern Recognition(ICPR), IEEE CS Press, Los Alamitos, Calif., 1984, pp. 347-349.
- 4 Jaekyu Ha, Robert M.Haralick and Ihsin T. Phillips, 'Recursive X-Y Cut using Bounding Boxes of Connected Components', Proceeding of 3rd Int Conf on Document Analysis and Recognition (ICDAR), Vol. II, pp 952-955 August 1995.
- 5 Baird, H.S., S.E. Jones, and S.J.Fortune, 'Image Segmentation Using Shape-Directed Covers,' Proc.10th Int'l Conf. Pattern Recognition (ICPR), IEEE CS Press, Los Alamitos Calif., 1990, pp.820-825.
- 6 Downton, A. C., S. M. Lucas, G. Patoulas, G. W. Beccaloni, M. J. Scoble and G. S. Robinson, *Computerising Natural History Card Archives*, to appear at ICDAR2003