Automatic Biodiversity Literature Enhancement (ABLE): A Project Overview

D.R.Morse, A. Dil, D.J.King, A.G.Willis
Department of Computing
The Open University
Walton Hall
Milton Keynes, UK.

D.M.Roberts, C. Lyal
The Natural History Museum
London, UK.

Abstract

We introduce the ABLE project, which aims to enhance access to collections of scanned documents from the biological taxonomy literature. Searching this literature needs to be robust to errors introduced by Optical Character Recognition and other sources. Biological knowledge, especially taxonomic knowledge, is often presented in a stylised form, generally using typographical clues to its meaning. This project aims to use typographical information and other contextual clues to identify and tag document content by its type. We describe some of the difficulties encountered in interpreting these scanned texts, and briefly discuss some methods of dealing with these issues.

The work in this document is wholly funded by JISC, the UK's Joint Information Systems Committee.

1 Introduction

The science of natural history began in the Renaissance and from it the various modern life-science disciplines have developed. Publications from the 15th century onwards provide a wealth of information, rich in observation, as natural science has moved from descriptive to the hypothesis-driven science that dominates today's publication landscape. The older literature can inform management practices in modern concerns, especially biodiversity loss, land-use patterns, sustainability and climate change.

Biological taxonomy is the discipline that manages the names for living and fossil organisms, defining the relationships within and between them. It therefore provides the central infrastructure for information management in the biological sciences (Knapp et al., 2004). Unlike most other sciences, taxonomic research and usage require access to the full range and history of publications on the subject. Publication through peer-reviewed journals is a relatively recent phenomenon. Until the 1930s, scientific observations appeared in a wide variety of publications, including learned Societies (e.g. *Proceedings of the Royal Society*), Institutional annual reports (e.g. *Abhandlungen der Akademie der Wissenschaften der DDR Berlin*) and encyclopaedias (e.g Bronn's *Thier-riechs*). Many of these publications are only held in a few libraries and are difficult to access. The difficulty of accessing taxonomic information is a severe impediment to research and delivery of the subject's benefits (Godfray, 2002). It has also been seen as a major impediment to implementing the Convention on Biological Diversity (SCBD, 2008). Taxonomic names change over time (Roberts, 2008) and while this is both inevitable and desirable as knowledge advances, it makes information management more challenging. For example, the taxonomic hierarchies used by Catalogue of Life[1] and the NCBI[2] are different, so the collective groups that might be used in a search comprise different actual organisms.

To liberate the information and data contained in the literature of the last 500 or so years, it is necessary to be able to search the documents electronically. This requires that the collections be digitised (Curry and Connor, 2007), for which industrial-scale scanning projects are essential. However, current OCR

[1] `http://www.catalogueoflife.org`
[2] `http://www.ncbi.nlm.nih.gov/Taxonomy`

(Optical Character Recognition) technology is not perfect. Errors are introduced at the scanning stage so that key words may be unrecognised by standard search techniques. To maintain, or better increase, the rate of scanning, it is not practical to engage in manual validation and error checking of documents. Therefore a mechanism to reduce the impact of OCR errors and to flag such errors for human correction is necessary.

2 OCR and Terminological Variation

Terminological variation is known to be a significant problem for the management of terms in biomedical curation (Nenadić et al., 2004) where orthographic and other linguistic variations can make automated recognition of similar terms difficult (e.g. for searching document collections). Such errors are are introduced by OCR processes. OCR can have high accuracy when applied to born-digital text (i.e. modern literature, where the target image has been computer-generated) as demonstrated by the PaperBrowser project (Karamanis et al., 2008), which supports curation of the FlyBase genomic database.

However, OCR performs markedly less well on scanned pages, especially of older publications. These have old typefaces and, to the modern eye, odd layout conventions (Lu et al., 2008) so recognition accuracy is consequently worse. Errors introduced by the OCR process give potential variations in recognised taxonomic names. For example, erroneous recognition of '*o*' in place of '*c*' might propose the taxon *Pioa*, not a known name, rather than *Pica* (European magpie). External data sources, e.g. Catalogue of Life and NameBank associate known latinised names with common names and synonyms, but these are under active development and are incomplete, and so cannot form the only basis for term recognition. In addition, mistaking '*o*' for '*a*' can change the genus *Homa* (a hemipteran insect) into *Homo* (mankind), so that non-appearance in an existing database cannot be used to identify errors. BHL observe 35% of taxon names in scanned documents contain an error and 50% of those errors are in one or two characters[3].

Further, the genus name *Pieris* is a valid name for both a plant (*Ericaceae*) and a butterfly (including the cabbage white), so a single name can represent two quite separate concepts. Abbreviation within text is also common, so we would seek to associate *E. coli*, for instance, with *Escherichia coli*, if it is a bacterium, or *Entamoeba coli*, if it is a protozoan.

3 Layout

PaperBrowser has demonstrated the value of representing layout information in a suitable markup language (SciXML). Such layout is normally self-consistent, but varies between publications.

The Biodiversity Heritage Library (BHL)[4] is pursuing a digitisation programme to improve the accessibility of taxonomy documents. The industrial scale of the project means that scanning takes place by volume rather than by article, so in BHL, the original scanned material must be identified by its volume without being able to identify individual articles within that volume. Although scientific tradition uses the article as the basic unit of reference, BHL cannot currently deliver that level of resolution. Typographical layout is an integral part of the information structure (Bringhurst, 2005), but often obeys conventions that have developed within a particular field of study (Hollingsworth et al., 2005). This structural information is independent of the language in which the text is written, so someone familiar with the principles of layout within the field of study can readily identify the section of a work that needs to be translated (Figures 1 and 2).

In our experience OCR from scanned pages recovers certain typographical features, such as paragraphs and headings, but it does not reliably determine other features, especially indent position and the distinction between normal, bold and italic text (Bapst and .Ingold, 1998). The very best modern OCR systems available, such as JSTOR, are more accurate than the desktop versions but such software is expensive and even the JSTOR system does not accurately capture typographi-

[3] Chris Freeland, personal communication

[4] `http://www.biodiversitylibrary.org`

MYCETES. 1

1861, in which he showed, principally from the observations of Herr Deppe and M. Sallé, that five genera are represented in Central America, and that *Mycetes* extends its range up to the 23rd parallel of north latitude. Eleven years later the same zoologist, in a paper "On the Quadrumana found north of Panama" (P. Z. S. 1872, pp. 2–9), enumerated six genera and ten species, tracing their distribution as far as was then known. This memoir has continued to be our chief authority on the Monkeys of the subregion; and further observation has only rendered it needful to modify it in some details, as will be seen by comparing the table of the distribution of the genera in the Introduction to the present volume with the similar one given by Mr. Sclater in 1872.

Fam. I. CEBIDÆ.

1. MYCETES.

Alouatta, Lacépède, Mém. de l'Inst. Nat. iii. p. 490 (1801)*.
Mycetes, Illiger, Prod. Syst. Mamm. et Av. p. 70 (1811).
Stentor, Ét. Geoffroy, Ann. du Mus. xix. p. 107 (1812).

Among the American Monkeys which have a partially naked prehensile tail, this genus is distinguished by the high pyramidical skull, deep massive mandible, small vertical incisors, well-developed thumbs, and above all by the extraordinary development of the larynx. In this organ the basihyal is much enlarged, and is expanded into a bulla or bony capsule, which is lined by a continuation of the thyroid sac, and forms the instrument by which the Howling Monkey produces the deep-toned and far-reaching shouts from which it derives its name †.

Eight or nine closely allied species of Howlers inhabit South and Central America, from Paraguay in the south to Guatemala in the north, most of them being confined to limited geographical areas. They agree closely in osteological and structural characters, but appear to be constant in their differences in colour. The two species found in our subregion, to which both are peculiar, may be thus characterized :—

1. *M. palliatus*. Hair of flanks elongated, forming a mantle; colour brown-black or black, the long hairs more or less fawn-coloured.
2. *M. villosus*. Hair very long and soft; colour uniform black.

* The new generic names here proposed by Lacépède, in his "Tableau des Divisions &c. des Mammifères" (often quoted as if it were a separate work), have almost all been entirely disused by subsequent writers, although an attempt to revive some of them has been made by Dr. J. H. Slack (Proc. Acad. Philad. 1862). Fortunately they are so imperfectly characterized as to come fairly under Rule XII. of the Strickland Code; had it been otherwise, many well-known generic names would have been ousted by such barbarous titles as *Sapajou*, *Ago-ára*, *Kinkajou*, *Stentor*, &c. &c.

† For a description of this structure see Owen, Anat. of Vert. vol. iii. pp. 599, 600, fig. 471.

B 2

Figure 1: A sample page from the *Biologia Centrali America*. This layout includes a page heading (centred capitals) on the same level as the page number; a continuation of body text from the previous page; two centred headings, one in bold and the other in capitals; a set of synonyms (not indented); body text (first line indented); two identification key questions (to differentiate species), strongly indented with outdented first lines; and two footnotes in smaller font.

System. 1765

Zoothamnium (Zoothamnia Bory 1824) **Ehrb. 1838, emend. Stein 1854 (261 u. 428)**; Brightwell (304); Clap. und L. (301); Wright (in 177, cf. IV); Uhlem (382); Quennerst. (405 b); Greeff (467); Engelmann (526); Wrzesniowski (546); Mereschkowsky (584); Forrest (522); Kent (661); Gruber (693); Entz (694); Stokes (713, 855); Andrussow (766); Perejasl. (788); Gourret und R. (774); Möbius (876).

Synon. Polype à bulbe Trembley (21); Clustering polypes Baker (29); plant ind. with sensibility Brady (32); Iris sessilis p. p. Linné (36, ed. X); Vorticella p. p. Linné (36, ed. XII); Müller p. p. polypinum (76); Modeer p. p. (82), Schrank (94), Dujardin p. p. (178); Der Baum p. p. Eichhorn (34); Dendrella Bory (115); Zoocladium Ehrb. (322, 129); ?Vorticellida M. Edwards (126); ?Antechlor Joseph (485).

Taf. 34, Fig. 2—4.

Einzelthiere mittelgross (gewöhnliche bis 0,08). Farblos und recht contractil. Bau der Individuen nicht wesentlich verschieden von Vorticella und Carchesium. Kolonienbildend ähnlich Carchesium. Der Aufbau der Kolonien ziemlich verschieden. Selten mehr reguläre dichotomische Verzweigung, wobei jedoch ein basaler Anfangs- oder Grundstamm gewöhnlich deutlich hervortritt. Weiterhin findet sich federförmiger Aufbau mit geradem oder zickzackförmig gebogenem Hauptstamm, der eine Fortsetzung des Grundstammes ist, und mit Seitenästen, welche dann gewöhnlich keine Aeste 2. Ordnung tragen. Eine Modification dieses Baues tritt durch sehr starke Verkürzung des Hauptstammes ein, so dass alle Aeste 1. Ordnung fast gleichzeitig aus dem dicken Grundstamm entspringen; sie bilden dann gewöhnlich Aeste 2. Ordnung. Hauptcharakter ist, dass sich bei jeder Zweitheilung auch der Stielfaden theilt, jeder Sprössling also die Hälfte des alten Stielfadens fortsetzt. Die Stielfäden sämmtlicher Individuen stehen daher in continuirlichem Zusammenhang, weshalb auch die Contraction stets die gesammte Kolonie ergreift. — Bei einigen Arten finden sich grössere Individuen (bis 0,12), sog. Makrogonidien, zwischen den gewöhnlichen; meist an den Ursprungsstellen der Aeste 1. Ordnung, doch auch nicht selten der 2. Ordnung. — Süsswasser (Europa, N.-Amerika) und Meer (zuweilen in ziemlicher Tiefe; N. Meer, Mittelmeer, Schw. Meer, Rothes Meer). Selten auf Wasserthieren befestigt. Artenzahl bis jetzt recht unsicher; ich glaube nicht, dass sich mehr wie 5—6 sichere Species auf Grund des Vorliegenden unterscheiden lassen.

Kent beschreibt unter dem Namen Zooth. pictum eine Süsswasserform, welche sich durch die Art ihrer Kolonienbildung von den übrigen Zoothamnien, doch auch den anderen kolonialen Vorticellinen so wesentlich unterscheidet, dass die Aufstellung einer besonderen Gattung wohl angezeigt wäre, wenn die Schilderung richtig ist. Das Charakteristische ist, dass sich auf dem Ende eines contractilen Stiels eine Gruppe von gleichen Individuen findet, welche selbst gar keine Stiele bilden. Das Verhalten der Stielfäden zu den Individuen wurde nicht bekannt.

c) **Tribus Acontractilia.**

Gestielt ohne Stielfaden. Einfach lebend oder kolonienbildend. Zuweilen gesellt sich Gallertausscheidung zu den Stielen.

Figure 2: A sample page from Bütschli's (1887 1889) *Protozoa*. Note that this has been scanned on a standard flat-bed scanner (darkening background towards the spine, on the left) and has not been de-skewed.

cal elements. The INOTAXA project found that scanned images of the *Biologia Centrali-Americana* to be intractable and the cheaper option was to have the content re-keyed[5].

The detection of text blocks on a page is normally achieved by pre-processing in the OCR package (for instance, the detection of left and right margins and columns), and these image features can be quickly determined (Lebourgeois and Emptoz, 1999). Lu *et al* (2008) have recently made substantial headway using rule-based pattern matching to recognise and analyse volume- and issue-title pages and a machine-learning approach to detect article title blocks and thus to generate article metadata. BHL scanning uses Abbyy FineReader and produces a light XML output (no styles, only words and paragraphs coordinates). However, certain terms are restricted or to particular types of narrative block; typographical cues such as paragraphs or columns are generally not a sufficiently accurate discriminator (Caracciolo and de Rijke, 2006). The efficient TextTiling algorithm (Hearst, 1997) can be used to provide a decomposition of a document into its argumentation components rather than its physical components. The argumentation passages identified by TextTiling have been shown to be more appropriate for such linguistic analyses than the typographical structural information.

Additional features such as the synonymy block in Fig. 1 are significant, and indicate that the synonymy text is not bodytext. We believe that Image analysis to recognise such structural information can be achieved using the open source application NIH Image[6] and extensions of the work previously carried out

[5]C. Lyal, personal communication

[6]`http://rsbweb.nih.gov/nih-image`

by Lu *et al*.

Figure 2 demonstrates taxonomic information that can be obtained from the typographical structure of a document. The taxon heading (*Zoothamnium ...*) is presented in a typographical structure very similar to the body text, except that it includes a list in smaller font. The synonymy statement is also in list-form but further-indented with an aligned first line. The single centred line below the synonymy statement is a direction to the illustrations which, typically for publications of this age, are gathered into a set of plates rather than presented near the referencing text. The single paragraph of body text is followed by a comment, logically equivalent to a footnote, with the same typography as the body text except in a smaller font. This comment is at the end of the section and is followed by a heading and finally more body text.

Techniques based on incremental parsing and markup are being increasingly used to manage the huge volume and variation of terminology across scientific literature (for example, (Cohen and Hersh, 2005) and in GoldenGATE[7]), in particular for the (difficult) task of Named Entity Recognition. Availability of the abstract collection Medline[8] has meant that research has generally focussed on the identification of biomedical terminology (typically gene and protein names) within plain text records; the preliminary stage of obtaining the documents through OCR and the subsequent possibility of incorrectly scanned terminology has received relatively little attention. Additional layout markup can be incorporated through enstensions to existing XML schema such as DjVu XML, SciXML (Lewin, 2007) and NLM DTD (used by BHL). Ultimately we are working towards full mark-up in the taXMLit schema[9].

4 Distributional Similarity and Term Search

Existing taxonomic hierarchies such as the Catalogue of Life[10] are incomplete and, as discussed, subject to continual change as the discipline advances. Such taxonomies and ontologies cannot therefore be used as the sole basis of search, although it must be possible to augnment them as additional taxa are obtained through analysis of the scanned documents.

To make a large quantity of scanned literature accessible, processing to extract the index terms must be automatic, and robust in the face of the OCR and other terminological variations discussed in section 2.

Initial similarity can be achieved with string matching techniques such as the Levenshtein edit distance (Sahinalp et al., 2003), but Weeds *et al* (2007) have also demonstrated the value of distributional similarity in managing biomedical terminology, where a high distributional similarity means that both words are surrounded by other similar terms. For example, consider the earlier example in which the taxon *Pica* has been incorrectly interpreted as *Pioa*. If the surrounding terms have contextual link with birds (or *Aves*, *Passeriformes*, *Corvidae*) or magpies, then the name is likely to be *Pica* (European magpie) and the term can be sensibly returned against a search for *Pica*. Similarly, the context should allow a distinction to be drawn between *Pieris* as used for a plant or for a butterfly. In this latter case there is no error in the OCR or the original typography but a single name representing quite separate concepts. Again, the context of the name usage should be able to resolve these instances. Weeds *et al* discuss possible distributional similarity measures that could form the basis for the current project. While both authors consider deep grammatical analyses as well as shallow measures, grammatical analysis is computationally expensive, and so in the first instance this project would use only a measure of co-occurrence of neighbouring terms to estimate term similarity.

There are four main categories of interest to modern research which are significant for contextual analysis: the scientific name (taxon), geographical location and personal names (e.g. authors, collectors or expedition members) and observation date. The first three categories are outside standard language, in that they are unlikely to be found in dictionaries available to OCR software, so are the most likely areas in which OCR errors

[7] `http://idaho.ipd.uni-karlsruhe.de/GoldenGATE`
[8] `http://www.nlm.nih.gov`
[9] `http://research.amnh.org/informatics/taxlit/schemas/taXMLit-v1-3.xsd`
[10] `http://www.catalogueoflife.org`

will occur (Tong and Evans,). The routines in GoldenGATE can be augmented with knowledge about additional clues, such as that personal names are often associated with an in-text citation, and that taxon names are generally italicised (ABLE is also investigating how the typographic features of a term can assist matching).

As given strings could match against more than one potential meaning, the local context is used to determine which concept is added to the XML mark-up. As strings potentially contain OCR errors, like the *Pioa* example given above, it would be imprudent to try to guess the correct form in all cases. It is better to return potential matches against a user query, so *Pioa* should be returned against a search for *Pica*, but it is also a plausible match for *Rea*, also a passerine bird but not a magpie. Linkage information should enable association tables to be built so that a search for 'magpies' also recovers *Pica pica*, for instance. Such information is being used to augment existing external data sources, particularly Catalogue of Life, NameBank and Global Names Architecture (GNA), which can be used for preliminary associations of latinised names with common names and synonyms.

5 Conclusion

There is particular urgency for this work in the fields of climate change and biodiversity loss, where biodiversity literature can provide base-line occurrence data and reveal historical patterns of change that can inform current management practices. The work pioneered by Lu *et al* needs to be extended to make searching the scanned literature more straightforward for the non-specialist, both within the HE sector and in the broader scientific community.

BHL currently scans material in units of a volume without being able to identify individual articles within a volume. Scientific tradition uses the article as the basic unit of reference and, at present, BHL is not able to deliver that level of resolution. It is this barrier to access that the ABLE project is attempting to lower.

6 Acknowledgements

The authors would like to thank Joint Information Systems Committee (JISC) for funding the work described in this document.

References

F. Bapst and R .Ingold. 1998. Using typography in document image analysis. In *Electronic Publishing, Artistic Imaging, and Digital Typography*, Berlin/Heidelberg. Springer.

R. Bringhurst. 2005. *The Elements of Typographic Style*. Hartley and Marks, 3 edition.

O. Bütschli. 1887-1889. Protozoa. abt. iii. infusoria und system der radiolaria. In H. G. Bronn, editor, *Klassen und Ordnung des Thiersreichs*, pages 1098–2035. Leipzig.

C. Caracciolo and M. de Rijke, 2006. *Generating and Retrieving Text Segments for Focused Access to Scientific Documents*. Lecture Notes in Computer Science. Springer-Verlag.

A. M. Cohen and W. R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71.

G. B. Curry and R. J. Connor. 2007. Automated extraction of biodiversity data from taxonomic descriptions. *Systematics Association Special Volume 73*, pages 63–81.

H. C. J. Godfray. 2002. Challenges for taxonomy. *Nature*, 417:17–19.

M. A. Hearst. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1).

B. Hollingsworth, I. Lewin, and D. Tidhar. 2005. Retrieving hierarchical text structure from typeset scientific articles a prerequisite for e-science text mining. In *Proceedings of the 4th UK e-Science All Hands Meeting*, pages 267–273, Nottingham, UK.

N. Karamanis, R. Seal, I. Lewin, P. McQuilton, A. Vlachos, C. Gasperin, Drysdale R., and E. Briscoe. 2008. Natural language processing in aid of flybase curation. *BMC Bioinformatics*, 9.

S. Knapp, G. Lamas, E. N. Lughadha, and G. Novarino. 2004. Stability or stasis in the names of organisms: the evolving codes of nomenclature. *Phil. Trans. Roy. Soc. Series B*, 359:611–622.

F. Lebourgeois and H. Emptoz. 1999. Document analysis in gray level and typography extraction using character pattern redundancies. In *Fifth International Conference on Document Analysis and Recognition (ICDAR'99)*, page 177.

I. Lewin. 2007. Using hand-crafted rules and machine learning to infer SciXML document structure. In *Proceedings of the 6th UK e-science All Hands Meeting*.

X. Lu, B. Kahle, J. Wang, and L. Giles. 2008. A metadata generation system for scanned scientific volumes. In *Proceedings of the 8th ACM/IEEE joint conference on Digital libraries*, pages 167–176.

G. Nenadić, S. Ananiadou, and J. McNaught. 2004. Enhancing automatic term recognition through recognition of variation. In *Proc. 20th International Conference on Computational Linguistics*.

D. M. Roberts. 2008. Explaining taxonomy to kids. *Society for General Microbiology Quarterly*.

S.C. Sahinalp, M. Tasan, J. Macker, and Z.M. Ozsoyoglu. 2003. Distance based indexing for string proximity search. In *Proceedings of the 19th International Conference on Data Engineering*, pages 125–136. IEEE, March.

X. Tong and D. A. Evans. A statistical approach to automatic OCR error correction in context. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 88–100, Copenhagen, Denmark.

J. Weeds, J. Dowdall, G. Schneider, W. Keller, and D. Weir. 2007. Using distributional similarity to organise biomedical terminology. In F. Ibekwe-SanJuan, A. Condamines, and M. T. Cabre Castellvi, editors, *Application-Driven Terminology Engineering*.