Improving search in scanned documents:
Looking for OCR mismatches

David Morse[1], Anton Dil[1], David King[1], Alistair Willis[1],
Dave Roberts[2], Chris Lyal[2].

[1] Department of Computing, The Open University, Walton Hall, Milton Keynes, UK
[2] The Natural History Museum, London, UK

Corresponding author: d.j.king@open.ac.uk

Background

The ABLE (Automatic Biodiversity Literature Enhancement) project aims to enhance access to collections of scanned documents from the taxonomic literature. Biological taxonomy is the discipline that manages the names of living and fossil organisms, defining the relationships within and between them. It therefore provides the central infrastructure for information management in the biological sciences [1].

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 [2]. It has also been seen as a major impediment to implementing the Convention on Biological Diversity [3]. Taxonomic names change over time [4] 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 [5] and the National Center for Biotechnology Information [6] are different, so the collective groups that might be used in a search comprise different actual organisms.

The older literature, dating from 15th century, can inform management practices in modern concerns, especially biodiversity loss, land-use patterns, sustainability and climate change. Therefore, unlike most other sciences, taxonomic research and usage require access to the full range and history of publications on the subject.

OCR and Terminological Variation

To liberate the information and data contained in the literature of the last 500 or so years, it is first necessary for these older publications to be digitised [7], for which industrial-scale scanning projects are essential. One such project is the Biodiversity Heritage Library (BHL) [8]. However, errors are introduced at the scanning stage because current OCR (Optical Character Recognition) technology is not perfect. The errors may mean that words are not recognised by standard search techniques, but at the current rate of scanning it is not practical to engage in manual validation and error checking of documents. To enable library users to search on the terms which are difficult for OCR systems to recognise, we therefore require mechanisms to reduce the impact of OCR errors.

This also applies to the task of automatic markup of taxonomic texts. Contemporary publications exploit the benefits of markup technologies for information sharing and information searching. Automatic markup of biodiversity texts will require accurate recognition of taxonomic names and then mark-up using extensions to existing XML schemas such as DjVu XML, SciXML [9] and NLM DTD (used by BHL). Ultimately the project will work towards full mark-up in the taXMLit schema [10]. We have already developed an XSL transformer for the reverse process, to extract source text from taXMLit documents.

OCR performs poorly on scanned pages, especially of older publications. These may have old typefaces and, to the modern eye, odd layout conventions [11]. Consequently, recognition accuracy is often worse than on modern publications. 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 being under active development, these 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. The BHL have found 35% of taxon names in scanned documents contain an error, with 50% of those errors being in one or two characters. 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.

Sequence Alignment to Identify OCR Errors

In order to start identifying some of the possible errors introduced by OCR, we are comparing the output of two different OCR packages. We have assumed that those terms which are difficult for an OCR package to recognise, are those which are most likely to be interpreted differently by different packages. The outputs from the OCR packages are compared against a source document, drawn from a Biologia Centrali-Americana (BCA) volume which was used in the INOTAXA [12] project. This

volume has been manually keyed in, and so is expected to contain (as far as possible) very few incorrect interpretations of the physical page. (INOTAXA found that manual rekeying of the journal content was more financially viable than automatic analysis of page scans.)

The text files we are comparing are both derived from a common PDF of the BCA volume. The first is taken from the Natural History Museum's work as part of the BHL, created with Adobe PDF maker and the associated OCR tool. The second was obtained from the Internet Archive [13] and was created using LuraTech PDF Compressor with ABBYY FineReader for the OCR.

To identify where the two OCR systems interpret strings differently, the two text files were split into words (using either whitespace or newlines as word separators), and compared using the Needleman-Wunsch algorithm [14] to align the texts. This algorithm performs a global alignment on two sequences by identifying the common terms between them, and inserting gaps or mismatches where no identical terms can be found. In our case, the mismatches identified by the algorithm are those terms which have been interpreted differently by the two OCR packages.

In practice, many of the misaligned terms are those that we would expect an OCR system to find difficult to recognise, and in fact, are often the taxonomic names that we would hope to recognise. Some examples are:

Reference	ABBYY FineReader	PDF maker
Otiorhynchinæ	`Otiorhynchinse`	`Otiorhynchinae`
Epicærina	`Epicserina`	`Epicaerina`
Sciaphilina	`Sciaphilina`	`Sciaphiliua`

showing that features such as ligatures cause problems for the OCR systems, but are not restricted to these (all the reference terms are italicised in the original document).

The comparison also highlighted other areas where the OCR systems return different results. The term '*RHYNCHOPHOBA.*' in the reference document was returned as '`BHYNCHOPHOKA.`' by PDF maker (illustrating some of the character misinterpretations), but as the pair of terms '`KHYNCHOPHOBA`' and '`.`' by ABBYY FineReader, illustrating both a spelling variant, and a different interpretation of the punctuation in the text. We do not currently analyse the punctuation in any way.

Our ongoing work is to identify how far the differences in the OCR outputs can be used to recognise the taxonomic names, and whether it is possible to find systematic interpretations of the spelling variants in these different outputs.

References

1. Knapp, S., Lamas, G., Lughadha, E.N., Novarino, G.: Stability or stasis in the names of organisms: the evolving codes of nomenclature. Philosophical Transactions of the Royal Society. Series B, 359, 611–622 (2004)
2. Godfray, H.C.J.: Challenges for taxonomy. Nature. 417, 17–19 (2002)
3. SCBD: Guide to the Global Taxonomy Initiative. CBD Technical Series, 30, pp viii + 195 (2008).
4. D. M. Roberts.: Explaining taxonomy to kids. Society for General Microbiology Quarterly. 23(5) 7–8 (1996)
5. Catalogue of Life, http://www.catalogueoflife.org
6. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Taxonomy
7. Curry, G.B., Connor, R.J.: Automated extraction of biodiversity data from taxonomic descriptions. Systematics Association Special Volume Series. 73, 63–81 (2007)
8. Biodiversity Heritage Library, http://www.biodiversitylibrary.org
9. Lewin, I.: Using hand-crafted rules and machine learning to infer SciXML document structure. In 6th UK e-science All Hands Meeting, National e-Science Centre, Edinburgh (2007).
10. Biodiversity Information Standards (TDWG) was known as the Taxonomic Database Working Group http,://wiki.tdwg.org/twiki/bin/view/Literature/WebHome
11. Lu, X., Kahle, B., Wang, J., Giles, L.: A metadata generation system for scanned scientific volumes. In 8th ACM/IEEE joint conference on Digital libraries pp. 167–176. IEEE Press, New York (2008)
12. INOTAXA ('INtegrated Open TAXonomic Access'), http://www.inotaxa.org/jsp/index.jsp
13. The Internet Archive, http://www.archive.org/index.php
14. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 48(3), 443–453 (1970)