From *Dendrœca blackburniæ* to *Dendrceca blackburniae*: what's in a name?

Developing a names-based architecture assumes you have good, clean names to work with. While this assumption generally holds true for modern born-digital literature, the process of digitising legacy literature can produce errors.

Therefore, when extending the names-based architecture back in time it is necessary to take into account these errors.

An example: *Dendrœca blackburniæ*.

Dendrœca blackburniæ.

Dendrœca blackburniæ,

D. blackburniæ

This text represents a common challenge for optical character recognition (OCR) because the words are not in a dictionary, so cannot be automatically verified as they are processed.

Three examples of the name from the same text with:

- a variety of formats,
- different sizes,
- stroke weights,
- directions to the face, and
- three different forms of the æ ligature.

The two ligatures, æ and œ, present an additional challenge. The OCR engine was set to expect modern English text, but the ligatures do not appear in modern English and so can **never** be recognised by the OCR process.

Fuzzy matching

Fuzzy matching can help correct the OCR rendering by finding similarities across the renderings and to the correct spelling.

Dendrœca occurs 59 times in the text, rendered as:

- Dendroeca 32 times,
- Dendreca 23 times, and
- once each for Bendrceca, Bendrwca, DendrcBca and Dendrosca.

The text also contains Dendroica, which occurs five times, correctly rendered by the OCR every time. Dendrœca and Dendroica will match fuzzily!

Therefore, we need collocation too.

Collocation

This technique examines surrounding words to provide the context of use which helps disambiguate similar words.

Collocation can help with blackburniæ, which occurs six times in the text. The OCR recognises the word as:

- blackburnice four times,
- blackburniae once, and
- blackburnw once.

Collocation shows that blackburniæ – however it is spelt – follows what looks like a genus or genus abbreviation. This additional information allows us to target our name correction to plausible binomial combinations.

Acknowledgements

This research uses the Biologia Centrali-Americana (BCA). PDFs and OCR renderings can be downloaded from the Biodiversity Heritage Library, www.biodiversitylibrary.org.

Thank you to Anna Weitzman and Chris Lyal of the INOTAXA project, www.inotaxa.org, for making their project's re-keyed texts of the BCA available for our research.

Read more about OCR post-processing

The ViBRANT project is building a corpus of marked up documents for research into OCR issues. This is freely available from git.scratchpads.eu/v/vibrantcorpus.git.

We are preparing papers from our work, covering the tools and workflows used in developing the corpus, and preliminary findings from analysis of the corpus.

One final challenge: Try looking up *Dendrœca blackburniæ* in a modern taxonomic reference. But that's another project…

David King and **David Morse,** The Open University, UK {david.king, david.morse}@open.ac.uk

ViBRANT is funded by the European Union 7th Framework Programme within the Research Infrastructures group.
Contract no. RI-261532. Period, Dec. 2010 to Nov. 2013. Coordinator: Dr Vince Smith. E-mail: enquiries@vbrant.eu