



# Open Research Online

---

The Open University's repository of research publications and other research outputs

## Liberate your historic data with The Open University

### Other

How to cite:

Morse, David and King, David (2013). Liberate your historic data with The Open University. Open University.

For guidance on citations see [FAQs](#).

© 2013 Morse and King

Version: Version of Record

---

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's [data policy](#) on reuse of materials please consult the policies page.

---

[oro.open.ac.uk](http://oro.open.ac.uk)

# Liberate your historic data with



In the past five years a team from the Open University's Department of Computing has won four grants, worth approximately £590,000, from both UK and European sources, to extract information from the legacy scientific literature. We specialise in addressing the problems inherent in print, not 'born-digital', literature. We have been working on the biodiversity legacy literature with collaborators from the Natural History Museum, London, and elsewhere.

Our intention is to make it easier for researchers and citizen scientists to find information of interest to them in the legacy literature by improving search. We also want to extract information from the literature automatically so that we can build links between different data sources.

The research challenges that we are investigating include:

**Big data** – The legacy biodiversity literature is estimated to be 300 million pages. The breadth and depth of biodiversity research requires processing the body text itself to answer meaningful research questions. Processing abstracts and table of contents alone are not sufficient.

**Noisy data** – Optical Character Recognition (OCR) errors introduced during scanning means that up to two thirds of named entities such as scientific names might be spelt incorrectly. Simple spell checking or look up against an authority is not sufficient to address this problem.

For example *Homo*, the genus name for humans, can be mis-interpreted by an OCR program as the butterfly genus *Homa*, so the context of use is very important.

**Disambiguation** – taxonomic nomenclature calls for unique names only within Kingdoms, hence there is a bacteria genus *Bacillus* and an insect genus *Bacillus*. We work with both technical solutions to this problem such as collocation of terms within texts and social solutions such as engaging citizen scientists to analyse extracted terms.

**Dissemination** – having extracted data from legacy literature, how do we share it? The inherent flexibility of Linked Open Data is permitting all manner of new possibilities for researchers to interrogate the liberated data.

We are always interested to hear from potential collaborators who want to unlock the data in their legacy literature.

Websites describing projects we are currently involved in are:

- <http://vbrant.eu/content/data-mining>
- <http://taxoncurator.myspecies.info/>
- <http://www.computing.open.ac.uk/comcur/home>
- <http://aginfra.eu>

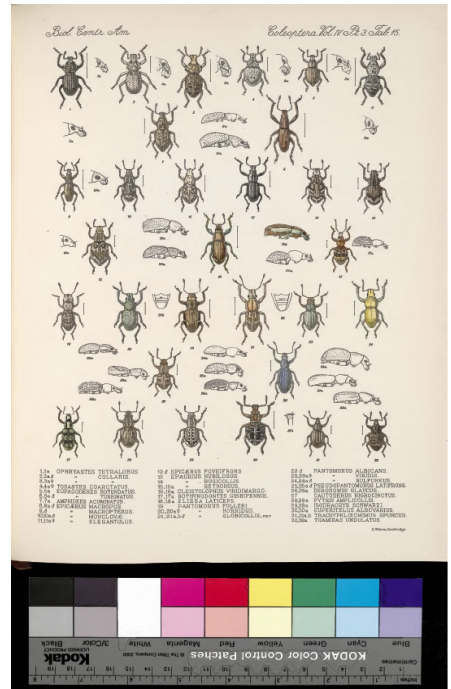
For further information on these project ideas, please contact:

David Morse – email [david.morse@open.ac.uk](mailto:david.morse@open.ac.uk)

David King – email [david.king@open.ac.uk](mailto:david.king@open.ac.uk)

Department of Computing, The Open University, Milton Keynes, MK7 6AA, United Kingdom

Scan of a plate from *Biologia Centrali Americana* showing beetles described in the volume. Taxonomists want to be able to search for descriptions of the beetles, in the volumes, and extract illustrations and information on the plates themselves.



This page contains descriptions of some of the beetles shown above. Note the variety of text styles, sizes and alignments used. Special symbols, fractions and a footnote appear.

