Liberate your historic data with

The Open University

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.

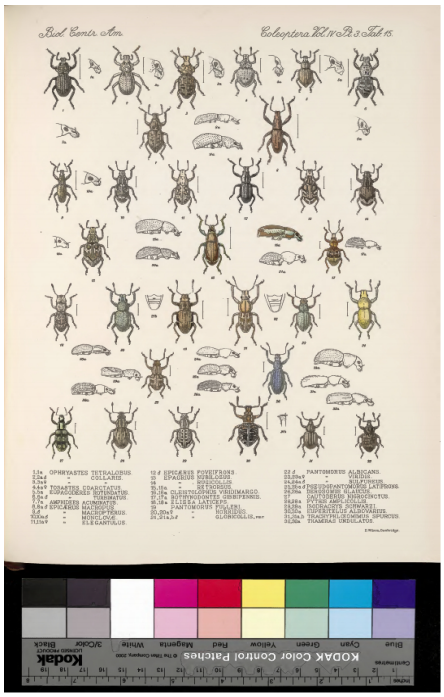
We are always interested to hear from potential collaborators who want to unlock the data in their legacy literature.

Websites describing projects we are currently involved in are:

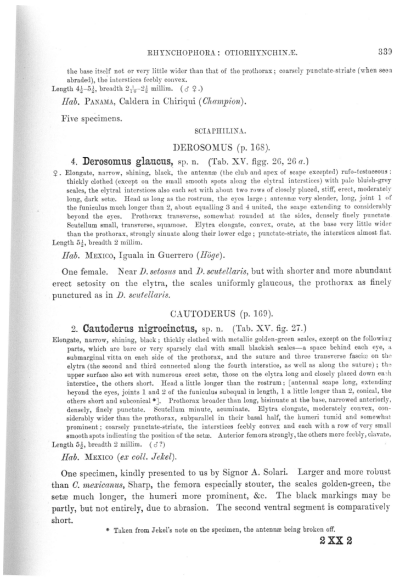
http://vbrant.eu/content/data-mining
http://taxoncuration.myspecies.info/
http://www.computing.open.ac.uk/comcur/home
http://aginfra.eu

For further information on these project ideas, please contact:
David Morse – email david.morse@open.ac.uk
David King – email 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.



RHYNCHOPHORA: OTIORHYNCHINÆ. 339

the base itself not or very little wider than that of the prothorax; coarsely punctate-striate (when seen abraded), the interstices feebly convex.
Length 4½–5½, breadth 2½–2¾ millim. (♂ ♀.)
Hab. Panama, Caldera in Chiriquí (*Champion*).
Five specimens.

SCIAPHILINA.

DEROSOMUS (p. 168).

4. **Derosomus glaucus,** sp. n. (Tab. XV. figg. 26, 26 *a*.)

♀. Elongate, narrow, shining, black, the antennæ (the club and apex of scape excepted) rufo-testaceous; thickly clothed (except on the small smooth spots along the elytral interstices) with pale bluish-grey scales, the elytral interstices also each set with about two rows of closely placed, stiff, erect, moderately long, dark setæ. Head as long as the rostrum, the eyes large; antennæ very slender, long, joint 1 of the funiculus much longer than 2, about equalling 3 and 4 united, the scape extending to considerably beyond the eyes. Prothorax transverse, somewhat rounded at the sides, densely finely punctate. Scutellum small, transverse, squamose. Elytra elongate, convex, ovate, at the base very little wider than the prothorax, strongly sinuate along their lower edge; punctate-striate, the interstices almost flat. Length 5½, breadth 2 millim.
Hab. Mexico, Iguala in Guerrero (*Höge*).
One female. Near *D. setosus* and *D. scutellaris*, but with shorter and more abundant erect setosity on the elytra, the scales uniformly glaucous, the prothorax as finely punctured as in *D. scutellaris*.

CAUTODERUS (p. 169).

2. **Cautoderus nigrocinctus,** sp. n. (Tab. XV. fig. 27.)

Elongate, narrow, shining, black; thickly clothed with metallic golden-green scales, except on the following parts, which are bare or very sparsely clad with small blackish scales—a space behind each eye, a submarginal vitta on each side of the prothorax, and the suture and three transverse fasciæ on the elytra (the second and third connected along the fourth interstice, as well as along the suture); the upper surface also set with numerous erect setæ, those on the elytra long and closely placed down each interstice, the others short. Head a little longer than the rostrum; [antennal scape long, extending beyond the eyes, joints 1 and 2 of the funiculus subequal in length, 1 a little longer than 2, conical, the others short and subcontinued *]. Prothorax broader than long, bistrinate at the base, narrowed anteriorly, densely, finely punctate. Scutellum minute, seminude. Elytra elongate, moderately convex, considerably wider than the prothorax, subparallel in their basal half, the humeri tumid and somewhat prominent; coarsely punctate-striate, the interstices feebly convex and each with a row of very small smooth spots indicating the position of the setæ. Anterior femora strongly, the others more feebly, clavate. Length 5½, breadth 2 millim. (♂?)
Hab. Mexico (*ex coll. Jekel*).
One specimen, kindly presented to us by Signor A. Solari. Larger and more robust than *C. mexicanus*, Sharp, the femora especially stouter, the scales golden-green, the setæ much longer, the humeri more prominent, &c. The black markings may be partly, but not entirely, due to abrasion. The second ventral segment is comparatively short.

* Taken from Jekel's note on the specimen, the antennæ being broken off.