The Open University

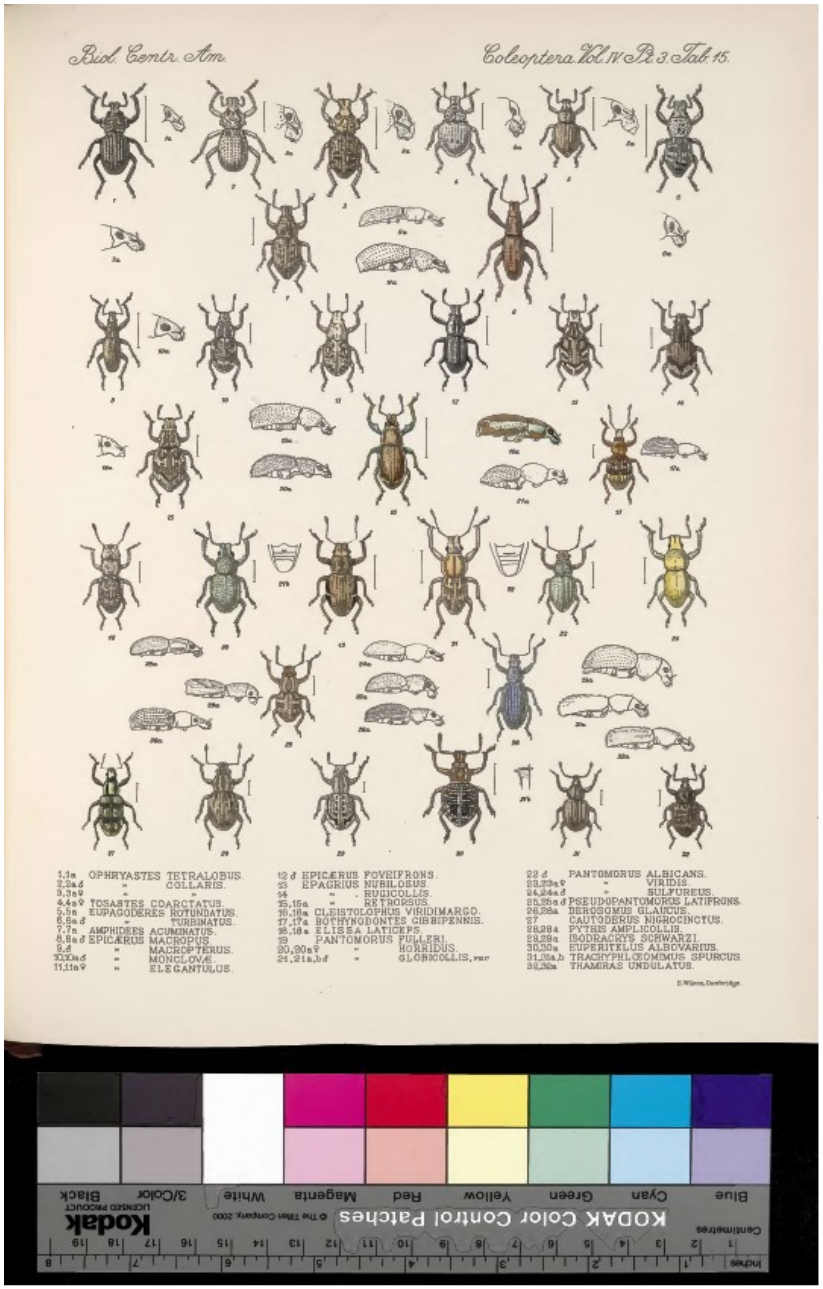
NATURAL HISTORY MUSEUM

Improving search in scanned documents

Looking for OCR mismatches

Our project problem

We are looking to extract taxonomic names from scanned documents, but the authoritative definitions of taxonomic names are in those same documents.



Sample plate:

This is a scan of the final plate in our sample volume of the Biologia Centrali Americana from 1911. It illustrates the beetles described in the volume. Taxonomists want to search for descriptions of the beetles in the volume.

Sample text:

This page contains descriptions of some of the beetles shown in the sample plate. Note the variety of text styles, sizes, weights and alignments used. In addition, special symbols, fractions and a footnote appear.

the base itself not or very little wider than that of the prothorax; coarsely punctate-striate (when seen abraded), the interstices feebly convex.

Length $4\frac{1}{2}$–$5\frac{1}{2}$, breadth $2\frac{1}{10}$–$2\frac{1}{2}$ millim. (♂ ♀.)

Hab. PANAMA, Caldera in Chiriqui (*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 $5\frac{1}{2}$, 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 subconical *]. Prothorax broader than long, bisinuate at the base, narrowed anteriorly, densely, finely punctate. Scutellum minute, acuminate. 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 $5\frac{1}{2}$, 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.

Needleman-Wunsch Alignment

This is a well established global sequence alignment algorithm. We use it to align terms in our sample scanned texts and hand corrected text of our sample volume.

We can vary the weightings for match, mismatch and gap insertion between texts. We are looking for differences between the scanned texts to help us identify characters that are frequently misidentified.

Sample results

Our program aligns and annotates the texts.

schwarzi	schwarzi	MATCH
34	34	MATCH
1	1	MATCH
	Â—	GAP
	,	GAP
	-	GAP
obsoletus	obsoletus	MATCH
273,	273,	MATCH

Future work

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

This understanding can be to used to clean up the scanned text should we be allowed to revise the published material, and if not then to enhance fuzzy searching of the text so that plausible variants are identified.

Contact details

Alistair Willis[1], David Morse[1], Anton Dil[1], David King[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

This project was funded by JISC