ComTax: *Community-driven Curation for Taxonomic Databases*

http://taxoncuration.myspecies.info/

Practising taxonomists want access to historic literature for primary taxonomic research. This research will in turn enable research into climate change, invasive species, biodiversity loss, etc.

Better access to scanned literature demands electronic search, which requires accurate curation of *named entities* (which can be used as search terms) in the scanned literature.

Converting scanned images into searchable text through Optical Character Recognition (OCR) potentially introduces errors, in addition to any synonyms and naming variations already present in the literature.

The ComTax project aims to support manual correction and verification of *named entities* by:

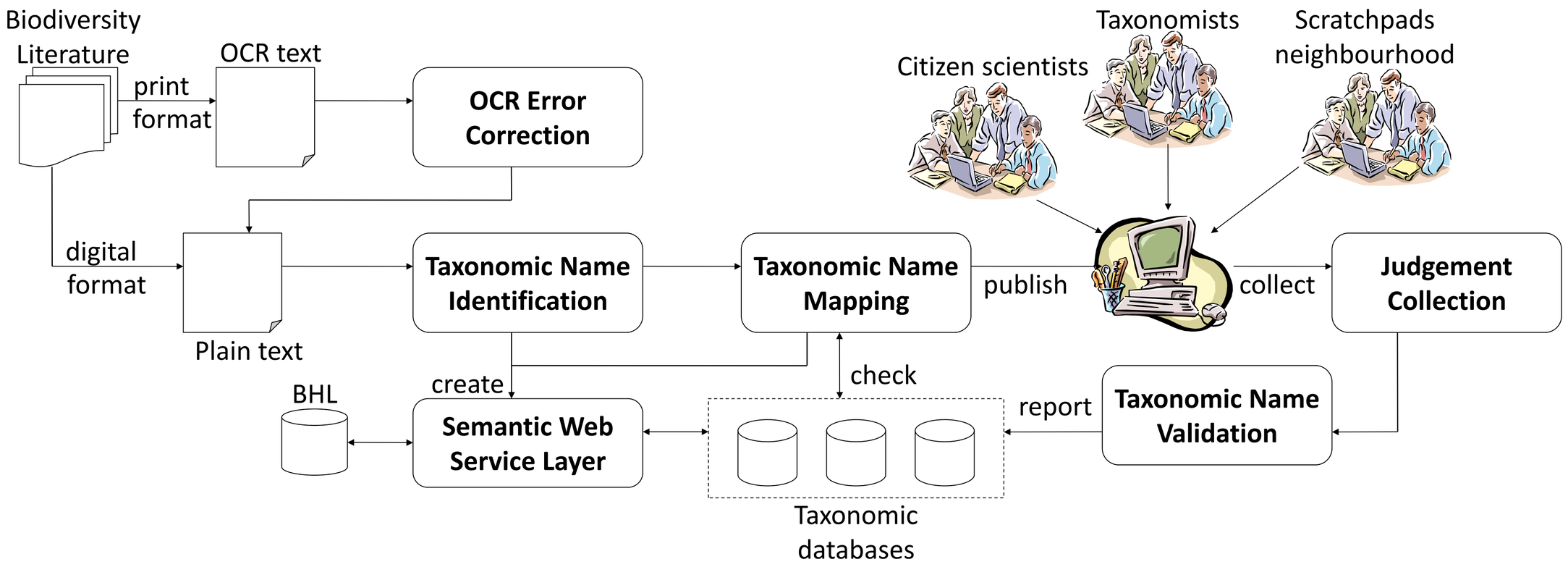
- experts applying their understanding of term variation,
- a collaborative annotation exercise including both taxonomists and citizen scientists, and
- using existing tools, especially external authority files.

The curated output of corrected and verified terms can be made available in files, databases or linked open data sets.

Example OCR error:

Rhynchites læticulus,

Is recognised as Bhynchites lasticulus



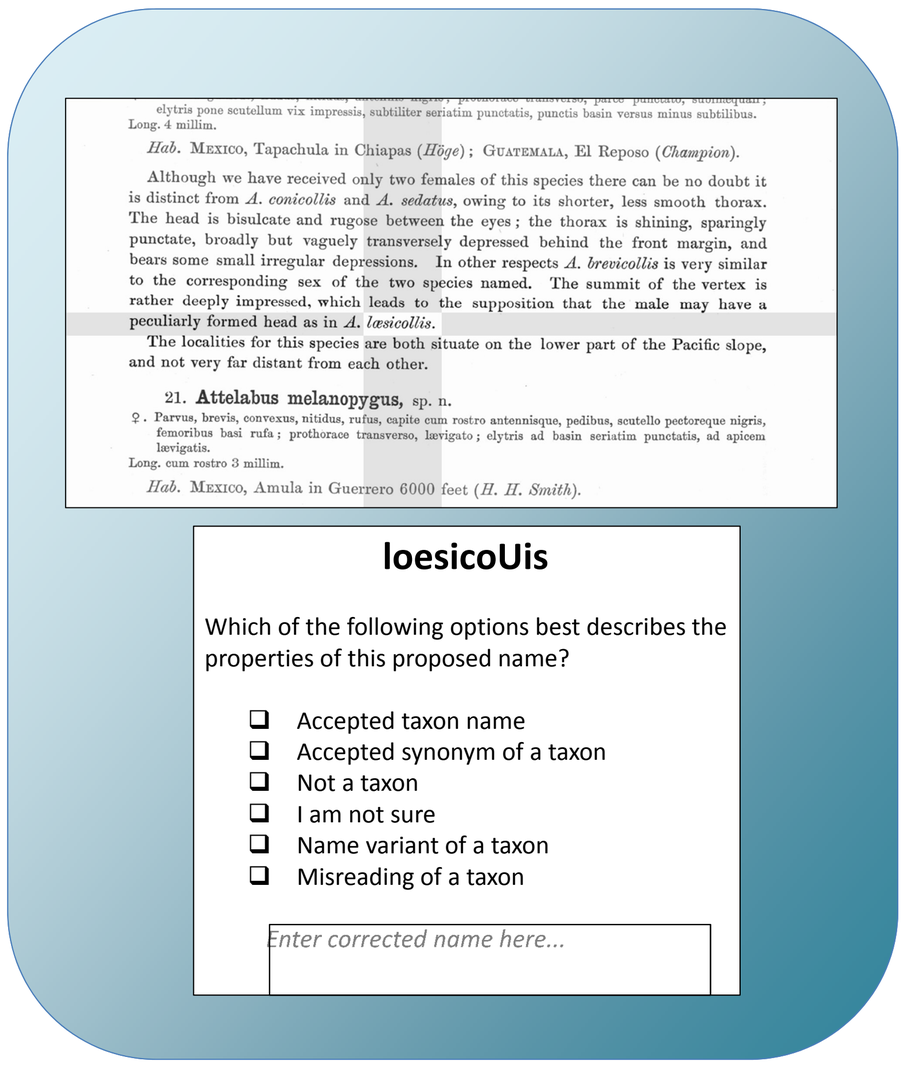
The system is intended to be used on uncorrected text after OCR on the scanned literature. The key stages in the workflow shown above are to:

- use machine learning techniques to identify possible taxonomic names
- look up the extracted names against external databases
- automatically marked-up the text if extracted name matched
- present the extracted name for validation or correction if not matched and then mark-up the scanned text with the curated name.

Unmatched names can occur for numerous reasons. The names are not currently recorded in the external databases, typically because the name in the literature has been reclassified since the text was written, or because an OCR induced error means that the name is incorrectly transcribed.

To enable users to validate or correct the OCR transcription the system presents a full page view of the OCR text with the name highlighted, as shown in the image to the right. The user is presented with a box listing the options for curating the name.

The process has been demonstrated to identify and curate taxonomic names found in biodiversity literature. However, the process will work with any *named entities* in literature drawn from other domains.


elytris pone scutellum vix impressis, subtiliter seriatim punctatis, punctis basin versus minus subtilibus. Long. 4 millim.

Hab. Mexico, Tapachula in Chiapas (*Höge*); Guatemala, El Reposo (*Champion*).

Although we have received only two females of this species there can be no doubt it is distinct from *A. conicollis* and *A. sedatus*, owing to its shorter, less smooth thorax. The head is bisulcate and rugose between the eyes; the thorax is shining, sparingly punctate, broadly but vaguely transversely depressed behind the front margin, and bears some small irregular depressions. In other respects *A. brevicollis* is very similar to the corresponding sex of the two species named. The summit of the vertex is rather deeply impressed, which leads to the supposition that the male may have a peculiarly formed head as in *A. læsicollis*.

The localities for this species are both situate on the lower part of the Pacific slope, and not very far distant from each other.

21. **Attelabus melanopygus,** sp. n.

♀. Parvus, brevis, convexus, nitidus, rufus, capite cum rostro antennisque, pedibus, scutello pectoreque nigris, femoribus basi rufa; prothorace transverso, lævigato; elytris ad basin seriatim punctatis, ad apicem lævigatis. Long. cum rostro 3 millim.

Hab. Mexico, Amula in Guerrero 6000 feet (*H. H. Smith*).

loesicoUis

Which of the following options best describes the properties of this proposed name?

- ❑ Accepted taxon name
- ❑ Accepted synonym of a taxon
- ❑ Not a taxon
- ❑ I am not sure
- ❑ Name variant of a taxon
- ❑ Misreading of a taxon

Enter corrected name here...

JISC

David Morse, Hui Yang, Alistair Willis, Anne De Roeck, David King and Jef Lay

Department of Computing and Communications, The Open University, UK.

david.morse@open.ac.uk

The Open University