ComTax: community-driven curation for taxonomic databases

Conference or Workshop Item

How to cite:


For guidance on citations see FAQs.

© 2013 Open University

Version: Version of Record

Link(s) to article on publisher’s website:

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.
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.

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.

Example OCR error:

*Bhynchites lasticus*,

Is recognised as *Bhynchites lasticus*

---

**Biodiversity Literature**

- **print format**
  - OCR text
  - OCR Error Correction

- **digital format**
  - Plain text
  - BHL
  - Semantic Web Service Layer
  - Taxonomic Name Identification
  - Taxonomic Name Mapping
  - Taxonomic Name Validation

**Judgement Collection**

- **collect**
  - publish
  - report
- **Scratchpads neighbourhood**
- **Taxonomists**
- **Citizen scientists**

---

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.

---

**JISC**

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

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

david.morse@open.ac.uk