Curation tools for taxonomic databases

Conference or Workshop Item

How to cite:
Morse, David; De Roeck, Anne; Willis, Alistair and Yang, Hui (2013). Curation tools for taxonomic databases. In: BioCuration 2013, Su ay 7 Apr to Wednesday 10 April 2013, Churchill College, Cambridge, UK.

For guidance on citations see FAQs.

© 2013 Open University

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
Curation Tools for Taxonomic Databases

http://taxoncuration.myspecies.info/

Biological taxonomy
- knowledge of species and their relationships
- underpins life sciences
- conservation of biodiversity

Much knowledge only held in historical literature
- often in specialist libraries
- difficult to access

Have identified and named ~1.78 million species of animals, plants, and microorganisms
- but 5-30 million species on Earth (est.)

Practising taxonomists need
- curation of names from scans
- better access to scanned literature
  - via electronic search

Curation from scans introduces errors/ambiguities
- synonyms and naming variations
- orthography (hyphens, spacing, ...)
- OCR errors

**Rhynchites laticulus, ⇒ Bhynychites lasticulus**

**Comtax project**
Project aims to support manual correction/verification of taxons
- by experts
- understanding of term variation
- a collaborative annotation exercise
- taxonomists and citizen scientists
- using existing tools
- taxonomic Scratchpads

Curation of corrected and verified terms
- into databases or linked open data sets

Using ongoing digitisation projects
- Biodiversity Heritage Library
- to create open access resources

**Term validation**
- Potential taxonomic names are identified using Named Entity Recognition
- Potential names presented to users with relevant context
- Judgements collected through taxonomic Scratchpads community tools
- Users can validate, or correct the OCR interpretation of the scan

**Database curation**
- taxonomic names matched against existing large-scale taxonomic databases
  - Encyclopedia of Life
  - Catalogue of Life
  - uBio Name Database
- Unmatched taxon names
  - collected for validation by taxonomists
  - used for further database curation
  - used to build indices to support search of the literature

David Morse, Anne De Roeck, Alistair Willis and Hui Yang
Department of Computing
The Open University, UK.
(d.r.morse, a.deroeck, a.g.willis, h.yang) @open.ac.uk