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.
Curation Tools for Taxonomic Databases

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

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

Bioinformatics and 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.)

Practicing taxonomists need:
- Curation of names from scans
- Better access to scanned literature
  - via electronic search

Rhynchites lasticus

Rhynchites laticulus

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...

Biodiversity literature

Extracted information

Crowd-sourcing

Citizen scientists

Taxonomists

Taxonomic name databases

Biodiversity text

Taxonomic name identification

Taxonomic name mapping

Scratchpads neighbourhood

Citizen scientists

Taxonomists

Taxonomic name validation

Collect

Check

Publish

Report

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

Biological taxonomy:
- Knowledge of species and their relationships
- Underpins life sciences
- Conservation of biodiversity

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

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