

Fuzzy matching of taxon names for biodiversity informatics applications

Acropaginula <> Arcopaginula
 Meosarmatium <> Neosarmatium
 Peneus <> Penaeus
 faveolata <> flaveolata
 capricornicus <> capricornensis
 abrohlensis <> abrohlensis



Tony Rees, CSIRO Marine and Atmospheric Research, Australia



Taxon scientific names are key identifiers in the world of biodiversity, yet for informatics applications they often fail to provide the required cross linkages on account of minor (or not so minor) differences in spelling arising from keying or phonetic errors, OCR (optical character recognition) and transcription errors, emendations, gender endings of species epithets, differences in diacritical marks, and more.

For example, data on the fish genus *Coelorinchus* (present "correct" spelling) might be stored under variant spellings *Caelorinchus* (previously considered correct), *Coelorhinchus*, *Coelorhynchus*, *Caelorhynchus*, and so on, while the potential for random or semi-random keystroke, OCR or transcription errors is almost limitless. If such potential variant spellings cannot be reconciled, some or even all of the desired data may not be retrieved.

This poster introduces TAXAMATCH, a "fuzzy" or near match algorithm developed at CSIRO Marine and Atmospheric Research (Australia), with the specific purpose of providing optimal fuzzy matching for genus and species scientific names in real world situations, and capable of deployment over a remote reference database of spellings deemed correct, or incorporation into any local system to suit a user's particular needs.

TAXAMATCH reference implementation

The reference installation of TAXAMATCH is currently installed over the IRMNG (*Interim Register of Marine and Nonmarine Genera*) database hosted at CSIRO Marine and Atmospheric Research, available via the access point www.cmar.csiro.au/datacentre/irmng/, which (at mid 2009) contains over 1.4 million species names from the Catalogue of Life and other sources, together with over 400,000 genus names. TAXAMATCH is automatically invoked when single genus + species, or genus queries are made so as to display not only exact, but also any near matches in the IRMNG database, to any user-supplied input name. Figs. 2 and 3 illustrate how TAXAMATCH will return a match of the correct spelled name "Homo sapiens" in response to an incorrectly spelled input name "Hombo sapient". Note that in this instance, operation of the genus and species pre-filters means that only 325 of the 445,004 genera, and 31 of the 1,459,171 species presently in the reference database are actually required to be tested, which contributes significantly to the relatively short execution time for the query (around 1 to a few seconds per input name, or less when conducted without the web interface and ancillary information presented).

TAXAMATCH use cases

A range of use cases can be envisaged for TAXAMATCH, including the following:

- Matching a (web or other) user's entered text against stored biodiversity information, where either the input or stored name may be misspelled or a variant spelling
- Checking of names on a "List A" that do not match entries on an equivalent "List B" (but may potentially include the same entities under variant spellings)
- Query expansion – for distributed data searches (where all name variants can be indexed in advance), as would be applicable to (e.g.) OBIS, GBIF, etc.
- Deduplication of stored lists – especially those constructed by aggregation of names from multiple sources
- "As you type" spell correction
- Application in taxonomic name recognition software, e.g. via OCR of scanned specimen labels, or detection of taxonomic names in mixed text streams (biological publications, etc.)

The web accessible IRMNG / TAXAMATCH search entry point also currently supports the input of batches of up to approximately 2,500 genus names or 1,200 genus + species names for automated checking, as shown in Fig. 4, and mechanisms for checking larger batches of names can be implemented via alternative mechanisms as desired.

TAXAMATCH operating principles

TAXAMATCH comprises a suite of custom filters and tests used in succession on genus, species epithet, plus authority where supplied, to return candidate near or "fuzzy" matches in a reference set of taxon names to any supplied input name. The actual tests employed include the following:

- An exact match test, both before and after minor normalisation
- A phonetic match test, using a custom algorithm "tuned" to the characteristics of taxon scientific names
- A custom "Modified Damerau-Levenshtein Distance" (MDLD) algorithm which looks for possible omitted, inserted, substituted and transposed characters and character blocks
- A modified *n*-gram comparison of author names and dates where supplied, including expansion of selected known abbreviations of author names as appropriate.

The custom filtering that has been developed for TAXAMATCH at both genus and species epithet levels comprises:

- Genus and species **pre-filters**, which serve to speed up the algorithm execution by excluding names deemed to be almost certain not to match from being tested
- Genus and species **post-filters**, which apply a set of rules to assist in the discrimination of likely "true" from "false" near matches
- A genus **cosmetic filter**, which presents only a subset of "genus near match" search results to the human web interface, while passing a wide range of genera through to the species stage for further testing
- A final **result shaping** stage (which can be switched out if desired), which masks more distant near matches in the presence of closer ones, but opens automatically to show them when the latter are absent.

A schematic of overall TAXAMATCH operation is shown in Fig. 1, below.

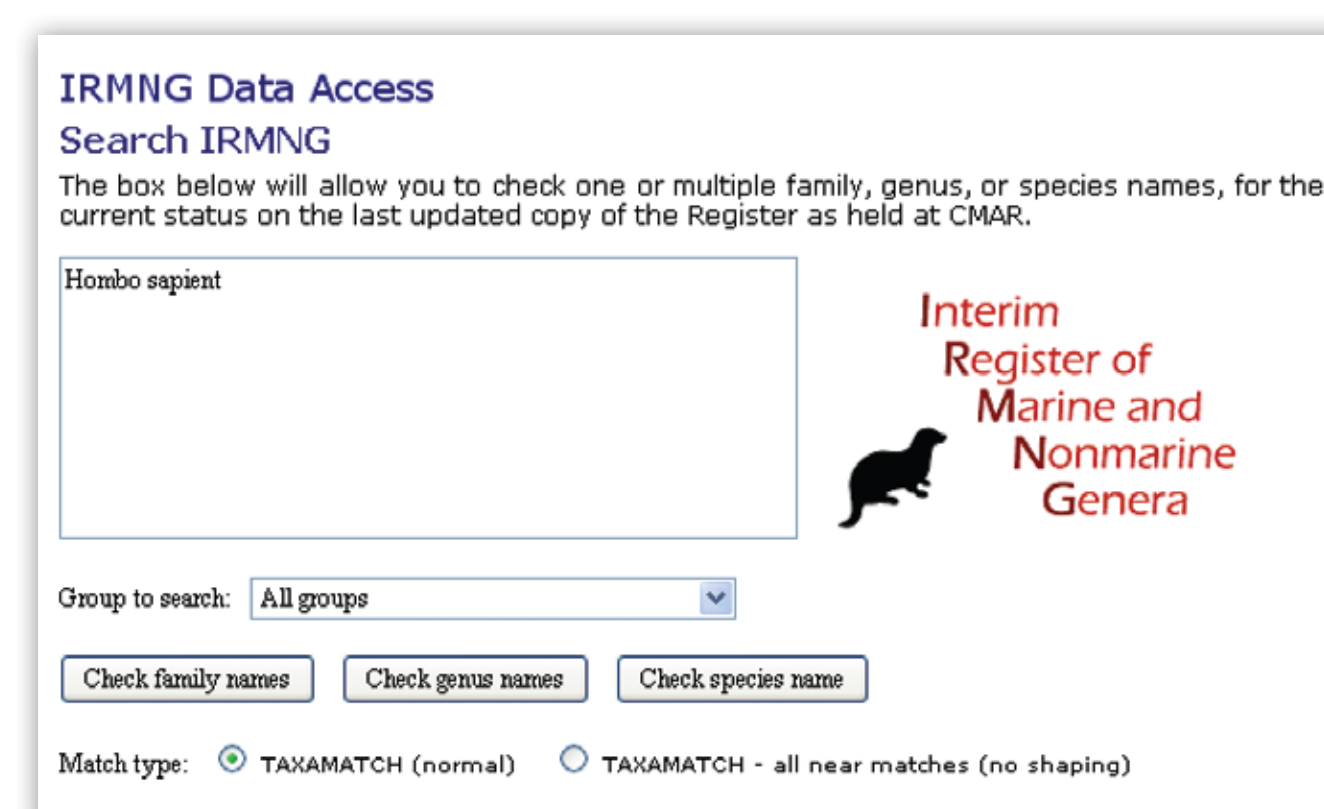
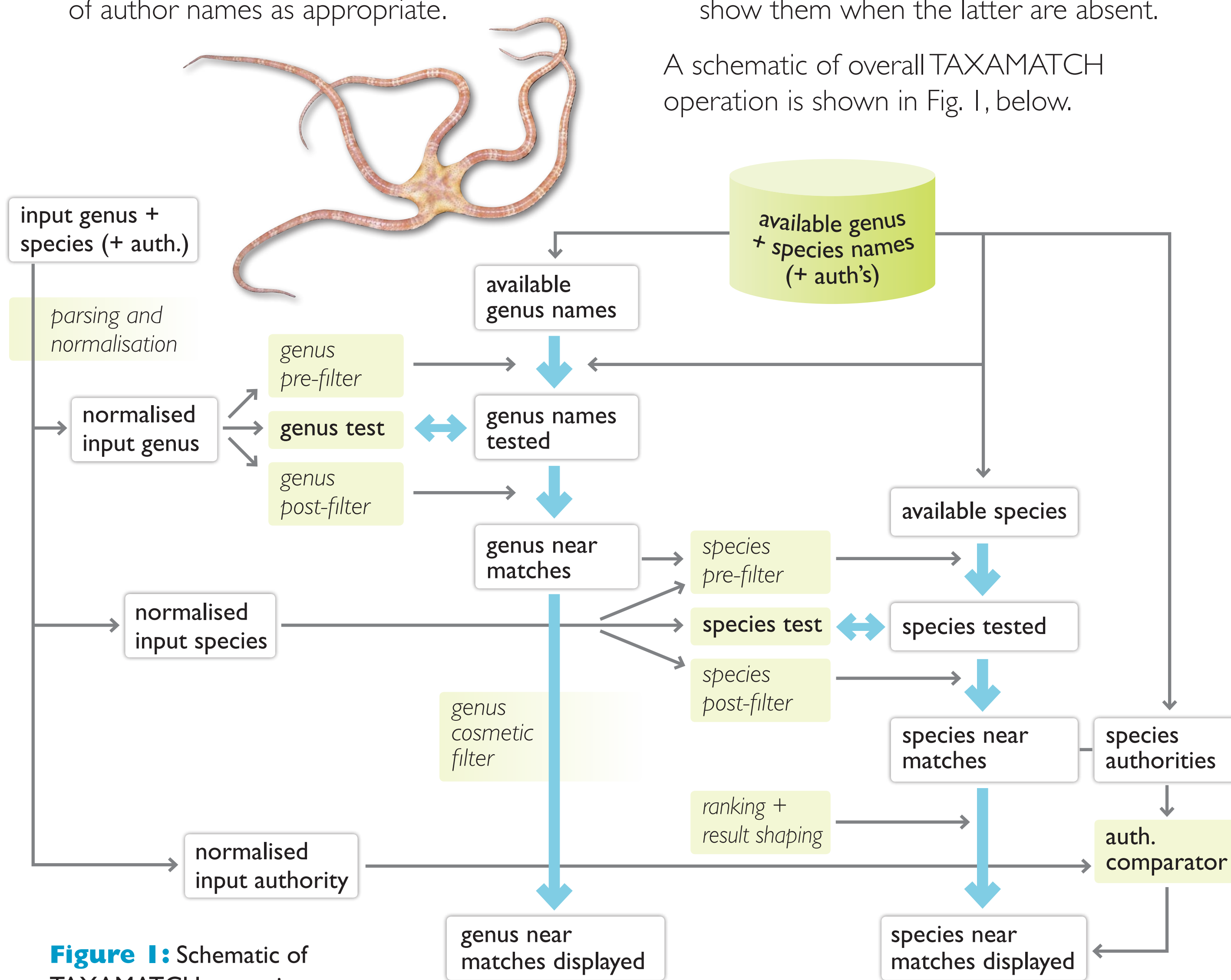


Figure 2: Web accessible IRMNG / TAXAMATCH search entry point www.cmar.csiro.au/datacentre/irmng/

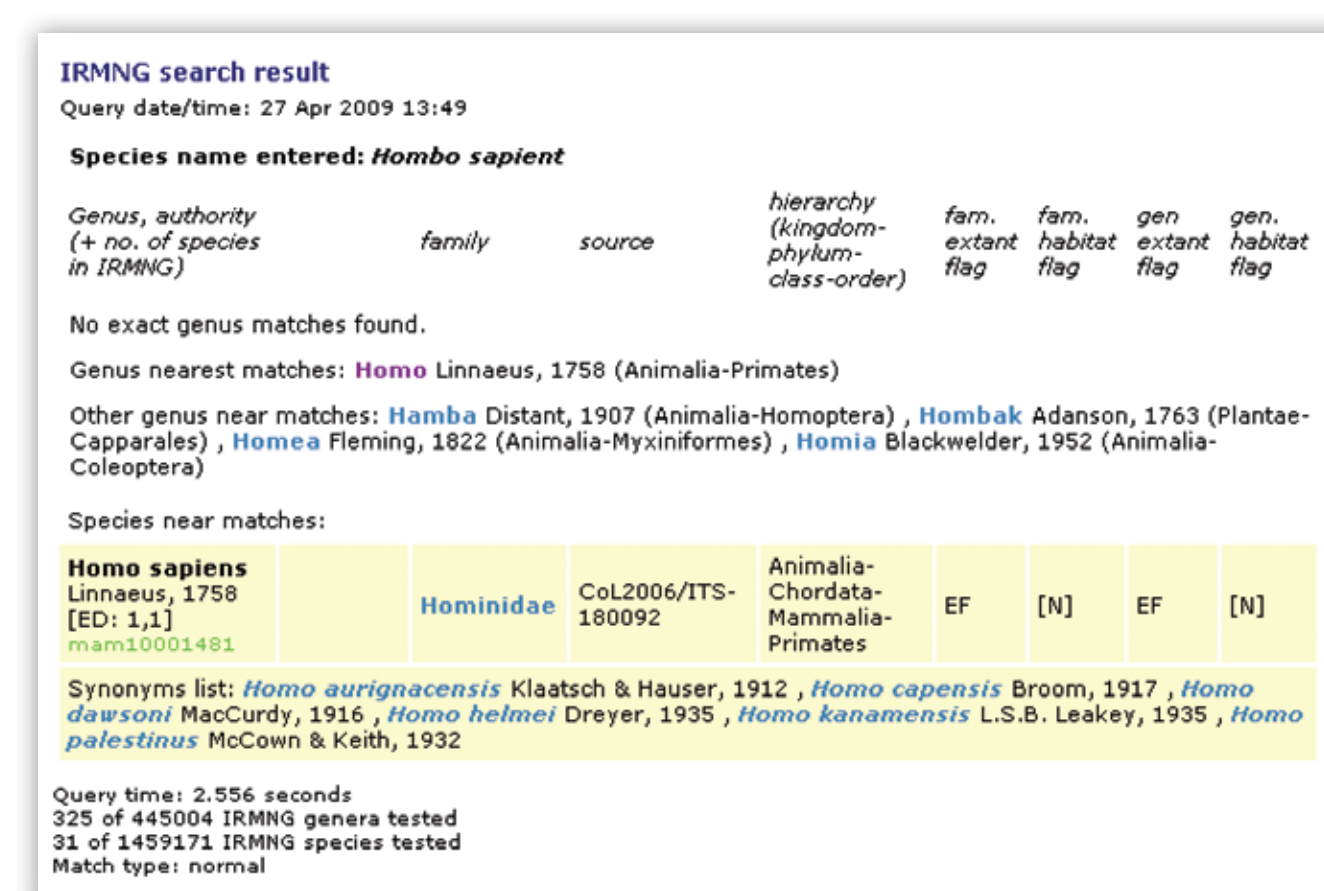


Figure 3: Result of above search for the entered term "Hombo sapient" against the IRMNG database

Species, authority	authority match	family	source	hierarchy (kingdom- phylum-class-order)
Species exact matches:				
<i>Callyspongia tozifera</i> Wiedenmayer, 1989		Callyspongiidae	Museum Victoria iKEMu database (Oct 2006)/Magna	Animalia-Porifera-Demospongiae-Haplospindelia
<i>Calocaris verribee</i> Poore & Griffin, 1979		Aviidae	Museum Victoria iKEMu database (Oct 2006); SP2000 NZ	Animalia-Arthropoda-Malacostraca-Decapoda
<i>Calocaris macandreae</i> Bell, 1853		Aviidae	Col2006/ITS-97708	Animalia-Arthropoda-Malacostraca-Decapoda
<i>Calochromis dentipes</i> Lea, 1909		Lycidae	Australian Faunal Directory (August 2007)	Animalia-Arthropoda-Insecta-Coleoptera
<i>Calochromis guerinii</i> W.J. Meadey, 1872		Lycidae	Australian Faunal Directory (August 2007)	Animalia-Arthropoda-Insecta-Coleoptera
Synonyms list: <i>Calochromis discicollis</i> Fairmaire, 1877				
<i>Calodema regale</i> (Gory & Laporte, 1838)		Buprestidae	Australian Faunal Directory (August 2007)	Animalia-Arthropoda-Insecta-Coleoptera
Synonyms list: <i>Stigmodera kirbil</i> Hope, 1836, <i>Stigmodera regale</i> Gory & Laporte, 1838				
<i>Calodera myrmeciae</i> Oke		Staphylinidae	Museum Victoria iKEMu database (Oct 2006)	Animalia-Arthropoda-Insecta-Coleoptera
Names not found:				
<i>Calocyba extranea</i>	fuzzy search / genus check			web links
<i>Calodera giachinali</i>	fuzzy search / genus check			web links
<i>Caloclea junonialis</i>	fuzzy search / genus check			web links
<i>Calogramma picta</i>	fuzzy search / genus check			web links

Figure 4: Sample IRMNG search result for a batch of multiple species names to be checked, showing option presented for "fuzzy search" on names that do not have an exact match to any current target name in the IRMNG database at this time.

Conclusion

TAXAMATCH appears to offer a good solution to the problems of near matching genus and / or species scientific names, whether for matching users' misspelled query terms to correctly stored target data (or vice versa), list cross-matching or internal deduplication, or as a prototype web accessible taxonomic spell checking service. Several development areas for TAXAMATCH are currently under active consideration, and interested potential users or developers are encouraged to contact the author at the address shown below or to visit the TAXAMATCH web page www.cmar.csiro.au/datacentre/taxamatch.htm.



References

- Rees, T. (2008). TAXAMATCH, a "fuzzy" matching algorithm for taxon names, and potential applications in taxonomic databases. *TDWG 2008 Annual Conference, Perth, Australia*, abstract and presentation available via www.tdwg.org/conference2008/program/.
- Rees, T. (2009 in press). TAXAMATCH, an algorithm for near ("fuzzy") matching of species scientific names in taxonomic databases. *Biodiversity Informatics* (submitted).

Acknowledgements

I thank Miroslaw Ryba, CSIRO Marine and Atmospheric Research, for programming and database assistance, and Barbara Boehmer, USA for assistance with modifying her original Oracle@Levenshtein Distance implementation for TAXAMATCH use. Photographs courtesy of Karen Gowlett-Holmes.

Further information

contact: Tony Rees
 phone: +61 3 6232 5318
 email: tony.rees@csiro.au
 web: www.cmar.csiro.au/datacentre/
www.csiro.au