**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 taxonomic 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 ‘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 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 on if desired), which makes 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.

---

**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 name 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](http://www.cmar.csiro.au/datacentre/taxamatch.htm).

---

**References**

- Rees, T. (2009) TAXAMATCH, a ‘fuzzy’ matching algorithm for taxonomic names and potential applications in taxonomic databases. *TAXON 58* (2009) pp. 22 (CSIRO Publishing and the Australian Biological Resources Study, Canberra, Australia). For assistance with using this software, contact (?) or for assistance with developing TAXAMATCH, including the following:

- Matching a web or user 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 potentially may include the same entities under variant spellings)
- Query expansion – for distributed data searches (where all variant names 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.