BiOnym – a flexible workflow approach to taxon name matching

Authors: Edward Vanden Berghe[1], Nicolas Bailly[2], Caselyn Aldemita[2], Fabio Fiorellato[3], Gianpaolo Coro[4], Anton Ellenbroek[3], Pasquale Pagano[4]

1: Vrije Universiteit Brussel (VUB), Brussels, Belgium. Corresponding author (evberghe@gmail.com)
2: Fishbase Information and Research Group (FIN), Los Baños, Philippines
3: Statistics and Information (FIPS), FAO, Rome, Italy
4: Istituto di Scienza e Tecnologie dell'Informazione A. Faedo, CNR, Pisa, Italy

Abstract

Several taxon name matching services are available on line, and many more are no doubt living on computers of individual scientists. While these systems may work very well, most suffer from the fact that the list of names used as a reference, and the criteria to decide on a match, are hard-coded in the engine that performs the name matching. One of the objectives of the EU FP7 project 'iMarine' (http://www.i-marine.eu) is to create a taxonomic name matching system, BiOnym, that would separate these elements.

The user will be offered a choice of several taxonomic reference lists, including the option to upload his/her own list to the iMarine infrastructure. Where possible, internationally recognized references are dynamically linked to the iMarine infrastructure; this avoids issues with intellectual property rights, and eliminates the inconvenience of keeping the reference lists up to date. The following lists are available in the infrastructure: the Catalogue of Life, World Register of Marine Species, Interim Register of Marine and Non-marine Genera, National Center for Biotechnology Information, and the Integrated Taxonomic Information System.

The matching process follows a workflow approach, starting with a pre-processing step, followed by series of operators to do the actual matching, concluding with a post-processing step. The pre-processing includes a parser, to split a taxonomic name in its atomized components (e.g. splitting the string in the name proper and the authority field), and a resolver to settle common spelling variations (e.g. replacing all occurrences of 'var.' to 'v.'). In the post-processing step, the modalities governing how the results of the matching process are presented to the user is defined. The matching is performed through a series of operators acting as switches. Each switch decides, on the basis of customizable criteria, whether a pair of names should be considered as 'matches', and splits the input list in 'matched' and 'non-matching' names. The matches go, with the criteria that were used to establish the match, to post-processing; the non-matching names are sent to the next switch. Two broad categories of switches are considered. A first type uses some kind of distance, such as the Levenshtein or soundex distance. Another type of switch applies a

transformation to both test- and reference names (e.g. strip off gender-specific suffix of specific epitheton), and then look for matches.

The switches are configurable and it is possible to upload customized character/string substitutions to configure the pre-processing step and transformations used by switches. Preliminary results of a series of experiments in taxonomic name matching will be presented.

The BiOnym approach lends itself to collaborative development. The iMarine infrastructure is open for any scientist, and anyone can contribute switches (as has been done by Tony Rees and Dima Mozzherin), or explore existing ones. Contributed switches can be installed on the iMarine infrastructure, or included via web services. The facility will allow quantitative comparison of the performance of different switches and their settings. By providing BiOnym, iMarine enables researchers to concentrate on taxonomic name matching rather than to develop data access or processing facilities.