Ecological Informatics 28 (2015) 29–41

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Retrieving taxa names from large biodiversity data collections using a flexible matching workflow

Edward Vanden Berghe [a], Gianpaolo Coro [b,*], Nicolas Bailly [c,e], Fabio Fiorellato [d], Caselyn Aldemita [e], Anton Ellenbroek [d], Pasquale Pagano [b]

[a] *Vrije Universiteit Brussel (VUB), Brussels, Belgium*
[b] *Istituto di Scienza e Tecnologie dell'Informazione A. Faedo, CNR, Pisa, Italy*
[c] *WorldFish, Penang, Malaysia*
[d] *Fisheries and Aquaculture Department, Statistics and Information (FIPS), FAO, Rome, Italy*
[e] *FishBase Information and Research Group, Inc. (FIN), Los Baños, Laguna, Philippines*

ARTICLE INFO

Article history:
Received 23 November 2014
Received in revised form 18 May 2015
Accepted 26 May 2015
Available online 4 June 2015

Keywords:
Taxon name matching
Taxonomic Authority File
Taxon name parsing
Name matcher chain
Taxonomy
Taxonomic nomenclature

ABSTRACT

In the domain of biological classification there are several taxon name matching services that can search for a species scientific name in a large collection of taxonomic names. Many of these services are available online, and many others run 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. In this paper we present BiOnym, a taxon name matching system that separates reference namelists, search criteria and matching engine. The user is offered a choice of several taxonomic reference lists, including the option to upload his/her own list onto the system. Furthermore, BiOnym is a flexible workflow, which embeds and combines techniques using lexical matching algorithms as well as expert knowledge. It is also an open platform allowing developers to contribute with new techniques. In this paper we demonstrate the benefits brought by this approach in terms of the efficiency and effectiveness of the information retrieval process with respect to other solutions.

1. Introduction

"*What's in a name?*" (Shakespeare, 1599: Romeo & Juliet, Act 2, Scene 2)

Querying that question in Google Scholar[1] just in the "title of the article" field will yield more than 3400 records (as of November 2014) in a wide range of domains of human activities. It has also been used many times as the title of taxonomists' oral presentations (or as slide title) to convey the important message that the proper management of scientific names of fossil and extant living organisms is essential to the understanding and the management of biodiversity. Coining names for artefacts of the physical world and for human conceptual constructions is essential to our communication. Scientific domains are themselves named xxx-logy, the etymology of the ancient Greek suffix root being "logos" (λογος) meaning a speech/discourse/debate; those would not be possible without names.

In biological taxonomy, the meaning of that question becomes: what is to be known through the scientific name about the organism it designates? The fact is that all data, information and knowledge about species are "hooked" to a scientific name. Therefore, (i) all that we know about a species can be retrieved from the literature by looking for the species name, which can be seen as indexing metadata (Patterson, 2014); (ii) different information systems can exchange data through species names, which can be seen as identifiers.

There should be an unequivocal link between a name and an artefact or a concept. This was clearly the goal when scientific names and their codes of nomenclature were developed. With vernacular names, which usually originate unplanned from common use, this is clearly not the case. However, even with scientific names, this unequivocal relationship is not absolute. Patterson et al. (2010) summarized the main issues that name matching encounters, among them: plain simple misspellings (formally known as "lapsus calami" in the literature on nomenclature), new combinations, several name-as-string variants for one name. These issues make it difficult to use them as identifiers. But they are quite efficient as indexing metadata to retrieve information on species.

* Corresponding author at. via Moruzzi 1, 56124, Istituto di Scienza e Tecnologie dell'Informazione A. Faedo, CNR, Pisa, Italy.
E-mail addresses: evberghe@gmail.com (E. Vanden Berghe), gianpaolo.coro@isti.cnr.it (G. Coro), n.bailly@cgiar.org (N. Bailly), fabio.fiorellato@fao.org (F. Fiorellato), c.aldemita@fin.ph (C. Aldemita), anton.ellenbroek@fao.org (A. Ellenbroek), pasquale.pagano@isti.cnr.it (P. Pagano).

[1] scholar.google.com.

http://dx.doi.org/10.1016/j.ecoinf.2015.05.004
1574-9541/© 2015 Elsevier B.V. All rights reserved.

To improve the role of scientific names as key to bind information from different sources, it is necessary to standardise their spelling and use. This is achieved most often through a process of matching them with a *Taxonomic Authority File* (TAF), i.e. a list of reference terms, including indication of synonyms and variants of scientific names, their authorship and possibly their data providers. The closer the match the better the chance that both systems speak about the same taxonomic concept. The human brain is quite good at matching names while detecting errors. But electronic systems match the strings of characters that constitute names outside any context, which makes them prone to compute false negatives. One understandable source of mistakes is that colleagues of whom their mother tongue is written in a non-Roman script are more likely to make spelling mistakes. For example, based on our experience and partially supported by the statistics in Froese (1997), the number of misspellings is high in Indian, Arabic, Chinese, and Russian journals.

Matching a string of characters is not enough to understand whether the intended species concept is the same: similar names might cover different species concepts (homonyms), different names might cover identical species concepts (synonyms). Resolving these issues is a different process from the taxon name matching that we will focus on here. Taxon name matching is a necessary first step, before the *content* (or in other words, the taxonomic concept covered by the name) is considered. The second step, concept matching, generally involves expert knowledge, and is considered the role of the Taxonomic Authority File: it is through the TAF that taxonomists have made their expertise available (Lambe, 2014), and allow us to judge whether names should be considered valid or invalid, and to disambiguate homonyms. For example, this approach is evident in the knowledge building process followed by the Catalogue of Life (Bisby et al., 2004), FishBase (Froese and Pauly, 2000) and WoRMS (Costello et al., 2013).

Several taxon name matching systems are available online, and many more are no doubt living on computers of individual scientists; a brief overview of those best known to the authors is included in Section 2, based on our knowledge of the tools used by several scientific communities around taxa matching. While these systems may work very well, many suffer from the fact that the list of names used as a reference (the TAF), and the criteria to decide on a match, are hard-coded in the engine that performs the name matching. The objective of this paper is to describe the BiOnym taxonomic name matching system that separates these elements.

In constructing such a system, it is not always possible to find the one size that would satisfy all the needs; to our experience, in the area of taxon name matching it seems that this "one size" is non-existent. Our ambition was to create a flexible, highly customisable framework to facilitate taxon name matching. This flexibility is deemed important for several reasons. First of all, it is important for determining if the scope of the reference list used is as close as possible to that of the list of names to be tested. For example, if a list of names of fish is compared with a very wide reference list such as the Interim Register of Marine and Non-marine Genera (IRMNG, Rees (2008a)) or the Catalogue of Life (Bisby, 2000), chances are that a lot of near-matches will actually be false positives (or even full matches comparing zoological names against a botanical TAF, and vice-versa). Consider the case of the genus "*Tisbe* Lilljeborg, 1853", a marine harpacticoid copepod. The genus is named after Thisbe, of "Thisbe and Pyramus" fame, but actually misspells the name of the mythological character. The correctly spelled "*Thisbe* Hübner, 1814" is a genus of butterflies. If the name "Thisbe" is used for a copepod, or in any marine context, it is very likely to be a misspelling for *Tisbe*. If it is compared with a TAF of the wrong scope, it might end up as the butterfly. On the other hand, if it is compared with a TAF including exclusively marine names, or with a TAF specific for crustaceans, "Thisbe" would likely be identified as a misspelling of "*Tisbe*". Another example is reported in Table 1.

Table 1
Variations on a theme: spelling variations for *Asthenognathus inaequipes*, a crab species from the Varunidae family. All spelling variations were taken from data contributions to OBIS (Berghe et al., 2010). Here, only variations in the name proper are shown; the number of different spellings of the taxonomic authority is often much higher.

Spelling variations
Asthenognathas inaefaipes
Asthenognathus inaeqipes
Asthenognathus maefaipes
Astheognathus inaequipes
Asthenognathus inaeguipes
Astheognathus inaeqinipes
Asthenognathus inaequipes

Another reason why we need a flexible approach is that the objectives of the end-users are not always the same, and dependent on the "use case". One possible use case for taxon name matching is to suggest, to some end user, a list of alternative valid names, for a list of names (s)he wanted to test. In this case it is important that the "correct" match is in the list of potential matches returned; the fact that other, false matches are also returned is of secondary importance: the "recall" should be as high as possible. Compare this with another potential use case, where taxonomic name matching is used to automate the association of names from a new dataset with names in a reference list. In this case it would be important to have a single suggestion for the matching name – in other words, that "precision" would be as high as possible. For this second use case we can break up criteria even further, according to the weight a wrong match would carry. If, for example, the taxon name matching was performed in the framework of merging different biogeographic data sets, the number of false positives should be weighted against the number of distribution records that cannot be used because no match was found. If, on the other hand, the taxon name matching was performed in the framework of the completion of a taxonomic reference list, false positives carry a much larger penalty, and should be avoided as much as possible.

Thus, in the first use case, it will be important to have a "recall" that is as high as possible; in the second use case, the "precision" will be the most important criterion. Recall and precision, and other measures of the quality of the matching process, will be further discussed in Section 5.1.

This paper is organized as follows: Section 2 reports an overview about taxon name matching. Section 3 explains our approach step-by-step, from the general rationale to the technical details. Section 4 explains the format of the reference datasets used by our process to search for the correct transcription of a species scientific name and the test dataset we prepared to evaluate the performance of our system. Section 5 reports the evaluation of the performance of each component of our method, both in terms of efficiency and effectiveness. Finally, Section 6 draws the conclusions.

2. Overview

Lexical matching is a standard computer application that crops up in several circumstances, for example in the spell checker of a word processor. Many general-purpose algorithms have been developed to support this matching (e.g. the Damerau–Levenshtein distance, Bard (2007), based on the minimum edit distance by Levenshtein (1966)), n-grams (Owolabi and McGregor, 1988), soundex (Odell, 1956) to name just a few, and which were used in the context of BiOnym. In this section we give an overview of methods that apply such techniques to taxon name matching.

Within the domain of taxonomic names/biological nomenclature, a considerable amount of work has been invested by the international biodiversity community in the creation of the Global Names Architecture (GNA) (GNA, 2014), much of it supported by the Global Biodiversity