Cyborg matching of taxonomic names, using nomenclatural logic

Campbell O. Webb (cowebb@alaska.edu),
Stefanie M. Ickert-Bond (smickertbond@alaska.edu),
University of Alaska Museum of the North

Abstract

Taxonomic names are the linkage keys that permit biological datasets to be joined together. However, name reconciliation is made difficult by the frequent variation in the character string for a single name. Computers can assist in the recognition of cognates between two sources using fuzzy matching, but subsequent human decision-making is usually needed. In our work on a new Flora of Alaska, we built a checklist of names from local and global name resources. For reconciling new names to our growing checklist, no existing tools (e.g, TNRS, and the GNR) were suitable. We therefore built a new taxonomic names matching application: '`matchnames`' (https://github.com/camwebb/taxon-tools). The app i) parses the elements of a raw name string, ii) applies a set of (botanical) taxonomic author string decomposition rules (e.g., omitted basionym, omitted 'ex' or 'in' author), iii) seeks exact matches on acceptable name string variants, iv) performs a fuzzy match on acceptable name string variants, and v) offers fuzzy-match choices to a human operator, who with minimal keystrokes can accept or reject candidates for the same name. In this way the app and the user work rapidly, behaving as an optimized 'cyborg' system.

The problem

Taxonomic names are the linkage keys that permit biological knowledge and datasets to be joined. However, the character string representing a taxonomic name can show large variation among different sources, making the accurate joining of datasets time-consuming. Reconciling two name lists containing name string variation is generally a three-stage problem:

1. Automatated finding of exact or acceptably close matches,
2. Automated fuzzy-matching to generate candidates,
3. Human decision-making to select among the candidates.

Increasing the speed of the reconciliation while minimizing erroneous matches can be achieved through specifying detailed rules for acceptable matches in (1), and optimizing the human-computer collaboration in (3).

Existing Solutions

Given the common need to clean and reconcile biological names, several tools have been created to match a user's list of names to the names in various online names resources (IPNI, Tropicos, NCBI Taxonomy, etc.). None of these however is appropriate for our specific needs (see Example Usage, below).

- **Taxonomic Name Resolution Service** (http://tnrs.iplantcollaborative.org/ and http://www.taxosaurus.org/)
- **TAXAMATCH** (http://www.cmar.csiro.au/datacentre/taxamatch.htm)
- **Global Names Resolver** (resolver.globalnames.org/)
- **WoRMS Webservice** (www.marinespecies.org/aphia.php?p=webservice)
- **R packages** `taxize` (https://github.com/ropensci/taxize/) and `taxonstand` (https://cran.r-project.org/web/packages/Taxonstand/)
- **OpenRefine** (https://openrefine.org/)

Our solution: rules for acceptable nomenclatural variation

Variation in names may be slight (e.g., a missing space after an author's initials, or a single character misspelling of a specific epithet), but frequently involves multi-character differences in author strings. These author variations sometimes arise via copying errors, but more often are created through the choices of how to encode the history of a name, made by taxonomists citing an earlier name: to add the basionym or not, to treat a validly publishing author as an "ex" author or as the main author, and how to abbreviate the authors' names (see Box 1 for botanical name elements). While the *International Code of Nomenclature for Algae, Fungi, and Plants* contains precise rules for name citation (Articles 46-50), there are also many recommendations, and often several "correct" ways to cite a name.

By incorporating some 'nomenclatural logic' about acceptable missing or mismatching author string elements into the automated stage of matching (stage 1, above), we can greatly decrease the number of cases that must be presented to the human operator, thus increasing the rate of processing.

(Box 1)

The elements of a botanical name

Salix alaxensis subsp. *glauca* (Andersson ex DC.) R. Coville ex Jones in Smith

<--> <------> <---> <-----> <----------> <-----> <--------> <------> <------>
gen sp irank infr basio ex_bas auth ex_auth in_auth

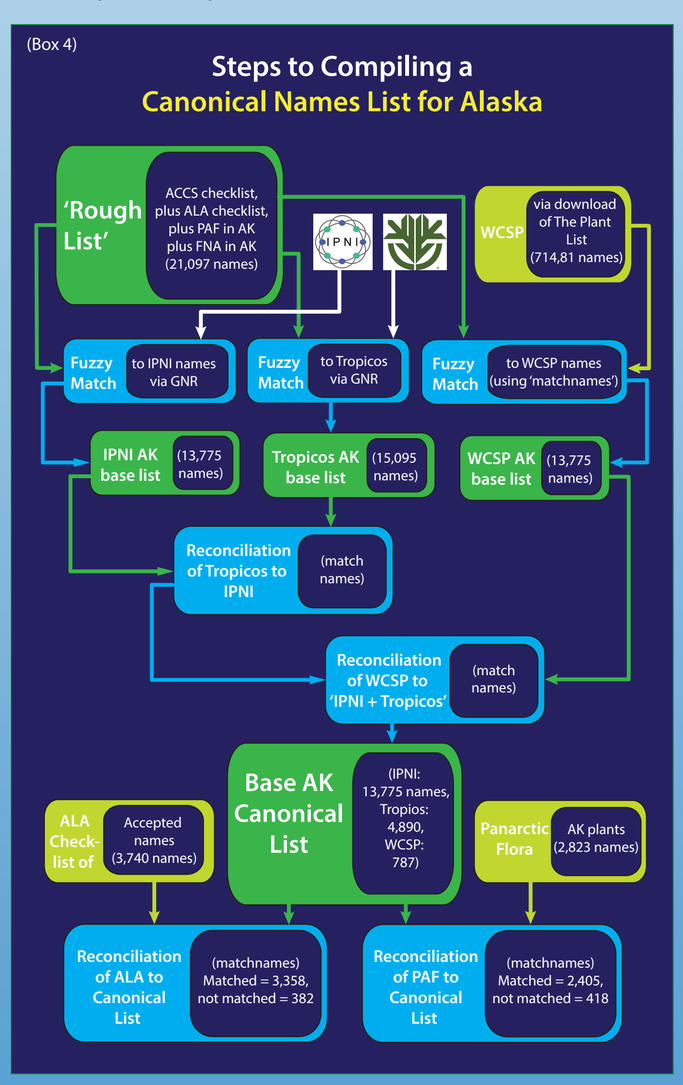
Key: gen: Genus. **sp**: Specific epithet. **irank**: Infraspecific rank. **infr**: Infraspecific epithet. **basio**: Basionym author(s): the author of the specific epithet before a change of genus or of infraspecific rank. **ex_bas**: ex Author(s) of basionym. **auth**: Primary author of name: the author responsible for first publishing the combination of gen and sp (and irank and infr if they exist). **ex_auth**: ex Author(s) for primary author: if the publication of the name by auth was invalid, the ex_auth was the author who subsequently published the combination validly. **in_auth**: in Author(s) for primary author: if auth or ex_auth were responsible for the combination but were not actually the authors of the *publication* in which that combination first appeared, the author(s) of the publication are added after '*in*'.

Example usage: Compiling a canonical names list for Alaska

As part of building the informatics backbone for a new Flora of Alaska (see http://alaskaflora.org/) we built a 'canonical' names list with these characteristics:

1) Names from "core" online names resources (e.g. IPNI, not a derived list),
2) Sourced preferentially by the highest name resource "quality" (IPNI > Tropicos > World Checklist of Selected Plant Families),
3) Internally reconciled to remove duplicates and orthographic variants,
4) Each with a Globally Unique Identifier (with a live URL), and,
5) Including almost all names applied to Alaskan plants (at least which are available online).

The process required multiple uses of `matchnames`:



(Box 2)

Screenshot of `matchnames` in use

```
   Androsace alaskana var. reedae S.L. Welsh & Goodrich
 1: Androsace alaskana var. reediae S.L.Welsh & Goodrich
  > 1
------------------------------------------------------ accs-613  --( 552/1000)
   Anemone multifida var. multifida Poir.
 1: Anemone multifida var. multifida
  > 1e
------------------------------------------------------ accs-631  --( 570/1000)
   Anemone patens subsp. multifida (Pritz.) L.
 1: Anemone patens subsp. multifida Hultén
  > e
------------------------------------------------------ accs-661  --( 596/1000)
   Anomobryum filiforme (Dicks.) Solms
 1: Anomobryum filiforme (Dicks.) Husn.
  > e
------------------------------------------------------ accs-664  --( 599/1000)
   Anotites viscosa Greene
 1: Anotites viscosus Greene
 2: Anotites pictus Greene
  > 1
------------------------------------------------------ accs-728  --( 662/1000)
   Antennaria friesiana (Trautv.) Ekman
 1: Antennaria friesiana (Trautv.) E. Ekman
  >
```

Keystrokes: `1:` accept choice 1; `e:` reject all choices; `1e:` accept choice 1 with medium probability.

The `matchnames` program

We built a new taxonomic names matching application, with a command-line interface, called '`matchnames`' (github.com/camwebb/taxon-tools).The components of a name-matching session are:

1. **Parsing** Recognizes the elements of a raw name string (into genus, specific epithet, infraspecific rank, infraspecific epithet and author string),
2. **De-punctuation** Both query name and reference names are "de-punctuated" to remove the effect of mismatching spaces, periods, non-ASCII author name characters (diacritics), etc. The depunctuation procedure is: a) converting non-ASCII characters into their appropriate ASCII character (e.g., ï to i), b) converting "and" or "et" into "&", c) removing all punctuation other than (,) and &, d) converting to lower-case.
3. **Nomenclatural rule-based decomposition** Applies a set of (botanical) taxonomic author string decomposition rules to recognize potential equivalent representations of the original name (e.g., omitted basionym, omitted 'ex' or 'in' author, etc.),
4. **Exact match?** Performs an exact match on acceptable namestring variants, storing such a match as a 'hit', and recording the taxonomic operation performed in (3),
5. **Fuzzy match** Performs a fuzzy match on acceptable name string variants, to generate candidates,
6. **Ask for human help** Offers fuzzy-match choices to a human operator, who with minimal keystrokes, without moving their hand (Box 2), can make decision on e.g., misspellings of a specific epithet, substitution of an ASCII character for a non-ASCII diacritic, different abbreviations of an author's name, etc (see Box 3). This is the 'cyborg element' ("a person whose physiological functioning is aided by or dependent upon a mechanical or electronic device").

In this way the app and the user work rapidly together. In our usage to date, between one in 20 and one in 100 names in an input list require human assistance.

Technology

`matchnames` is a single Awk script (GNU Gawk dialect), and can be run easily on Windows, Mac and Linux. Source code, installation instructions and documentation are available on Github (https://github.com/camwebb/taxon-tools).

(Box 3)

Examples of matches requiring human decisions

'SAME' refers to a decision on whether the two names strings refer to the same published name, not just the same genus and specific epithet. Note the 'probably' in some of the decisions. Human decisions are recorde with a measure of likelihood, so that unlikely similarities can be filtered at a later stage.

Spelling of name and different abbreviation for author (SAME)
Mertensia paniculata var. eastwoodae (J.F.Macbr.) Hultén
Mertensia paniculata var. eastwoodiae (Macbride) Hulten

Missing basionym (likely SAME)
Fauria crista-gallii (Menzies) Makin
Fauria crista-galli Makino

Missing author (probably SAME)
Arnica alpina (L.) Olin
Arnica alpina (L.) Olin & Ladau

Missing 'ex' author, differing orthography and abbreviation of author (SAME)
Puccinellia interior T.Sorensen
Puccinellia interior T.J. Sørensen ex Hult

Differing basionym (probably NOT SAME)
Myosotis palustris L.
Myosotis palustris (L.) Nath.

Differing author (NOT SAME)
Alnaster sinuata (Regel) Czerep.
Alnaster sinuatus (Rydb.) Czerep.

Acknowledgements

This work was funded by National Science Foundation Grant 1759964: "ABI Innovation: Taxonomically intelligent data integration for a new Flora of Alaska". We thank the developers of GNU Gawk, and the TRE library.