Toward a taxonomic name resolution service: Summary and prioritization of end-user requirements

April 25, 2010

CONTENTS

INTRODUCTION

The past decade has seen an explosive growth of large biological databases aggregated from multiple sources. Regional, continental and global-scale data warehouses and networks such as SEINET, SpeciesLink, REMIB, the Australian Virtual Herbarium, and GBIF provide access to millions of records from biological collections worldwide. Thousands of ecological inventories and species trait measurement are now available through portals such as VegBank, SALVIAS, and TraitNet. Electronic archives such as GenBank and TreeBase house millions of records of sequence data and phylogenies for hundreds of thousands of organisms. Such "mega datasets" represent a major new tool for the study of biodiversity, and have made possible analyses at spatial and temporal scales unimaginable even a decade ago (e.g., Loarie et al. 2008, Weiser et al. 2007, García 2006, Peterson et al. 2002).

Unfortunately, the increasing use of mega datasets has highlighted a major obstacle within the biological sciences: the taxonomic impediment. Do two different names represent two species or one? Does the same name used in different data sets at different times refer to the same species? How to extract the intended meaning of misspelled names, abbreviations and variant spellings?

We find that *even in the most reliable sources*, when taxonomic data are reported in the literature, about 15% or more of Latin binomials are either misspelled or are ambiguous, and many more are out of date. Unfortunately, for plant phylogenetic, ecological, and trait databases, error rates approach 25 to 35%. Resolving such errors for large biological databases—with hundreds of even thousands of taxon name strings—is currently a time-consuming, error-prone chore. This taxonomic impediment is perhaps the largest barrier remaining to conducting comparative science.

As a contribution toward resolving the taxonomic impediment, BIEN (the Botanical Information and Ecology Network) is coordinating the development of a Taxonomic Name Resolution Service (TNRS), to be developed by iPlant in collaboration with the Missouri Botanical Garden (MBG). The TNRS will be a suite of applications for automated and computer-assisted correction and standardization of taxonomic names, with the primary goal of facilitating taxonomic standardization of large biological databases through machine-to-machine transfer and manipulation of taxonomic information. Drawing upon the extensive taxonomic resources of MBG's TROPICOS database (www.tropicos.org), the initial implementation of the TNRS will provide essentially complete coverage of plants of the New World. However, incorporation of additional source of names and synonymy are incorporated (e.g., the International Plant Names Index and World Checklist of Monocots) will enable essentially global coverage of vascular plant taxon names and concepts.

This document provides a summary and prioritization of end-user needs, as identified during the meeting "Toward a Taxonomic Name Resolution Service" (Missouri Botanical Garden, St. Louis, MO, March 31 - April 2, 2010; for a summary see meeting website) and during subsequent discussions among meeting participants. User needs are presented as prioritized components performing specific functions, as documented by use cases compiled during and after the meeting (see Appendix and Use Cases on the TNRS Meeting website).

SUMMARY AND PRIORITIZATION OF TNRS COMPONENTS

1. Name standardization service

Software needed (S1): An advanced taxon scrubber tool that maps name strings onto a comprehensive list of validly published names with only one variant for each name. To be truly useful and an advance over existing tool such as Taxonscrubber and TaxaMatch (also see http://www.silverbiology.com/products/taxamatch/) this application should include both parsing capability (Appendix 2b) and fuzzy matching (Appendix 2c), with the option of automatically selecting the single best match when multiple options are available. The initial implementation could simply dump results as text, returning multiple records for ambiguous names. Next, implement as a web service. The final implementation should include web-based tools to assist user in selecting among multiple options and resolving ambiguities.

Data needed (D1): List of valid names and authors in standardized format with just one standard variant per name. This could eventually comprise part of GNUB. Such a list is currently available for most New World taxa within the TROPICOS database. Global coverage could be provided by merging this list with IPNI. However, IPNI currently presents multiple variants of the same name, a legacy of merging names from three sources: the Index Kewensis (IK), the Gray Card Index (GCI) and the Australian Plant Names Index (APNI). Merging of IPNI with TROPICOS will require that the former de-dupe their records, but we are not sure how long this will take. We suggest that initial development proceed immediately using TROPICOS names.

2. Hierarchical checklist name resolution tool

Software needed (S2): A tool that will allow the user to batch-correct names according to one or more synonymized lists. The user selects one or more lists and orders them according to priority of application. A name corrected in one list will be ignore in lower-priority lists. The tool should also distinguish simple and complex synonyms, and, in the case of complex synonyms, present the user with a list of possible options. Initial output could be a simple text dump, with subsequent development as a web service. Final implementation should include web tools for alerting the user to ambiguities and assisting him or her in selecting the preferred single best result.

While not optimum, this would still be extremely useful, quick to implement, and has the advantage of providing a simple and transparent "paper trail" of the decisions used to standardize a particular set of names.

Data needed (D2): Access to synonymized digitized project checklists and literature-based synonymies within TROPICOS database. Access to other sources of databased synonymy (Kew World Checklist of Monocotyledons, efloras, etc.). Ability to freely incorporate new sources of synonymy as they become available (e.g., digitized versions of Henderson's Palms of the Americas, Palms of Southern Asia).

3. Dynamic checklist generation tool

Software needed (S3): A tool for combining regional and monographic synonymized checklists into a single master (regional or global) checklist (see Appendix 3c). The application should be capable of selecting a single "best" opinion regarding the taxonomic status of a particular name,

according to a transparent set of rules. Ideally, it should be capable of deducing and presenting to the user possible relationships between nominal concepts associated with a name and the concepts represented in the checklist. This application is required for S4.

Data needed (D3): As for D2, with emphasis on the master set of checklists & synonyms being used by TROPICOS and Kew to create their world checklist, and perhaps other key New World and global checklists that they are not using.

Progress and plans: We learned at the TNRS meeting that MO and IPNI are working on such an application, with a goal of having a preliminary world checklist available within a year. However, it is not clear how far they have actually progressed. Furthermore, we do not know if they are generating software that will incorporate new lists as they become available; we need that functionality. Finally, the algorithms and decision rules they are using are not apparent. It is important that these rules be made apparent to both developers and users; indeed, in the interest of verifiability and repeatability, users should be able to document the sources consulted and decisions used to synonymize any name submitted. It is important that very soon we have an open and comprehensive conversation with developers from MO and IPNI to clearly identify the attributes of any software being developed for this purpose and the decision rules being employed within that software.

4. Dynamic checklist name resolution tool

Software need (S4): A tool that will allow the user to batch-correct taxonomic names, returning only one accepted name for each name submitted (Appendix 3c). Reference list of names and relationships is compiled using the Dynamic Checklist Tool (S3). Application should return the accepted name, sources (literature or database) upon which the name is based, and flag any names with possible ambiguities resulting from complex synonymy, homonyms, etc. Initial output could be a simple text dump, with later implementation as a web service. Final application should include web-based tools for documenting the sources and decision rules behind each taxonomic decision, with links to information needed to resolve ambiguities.

Data needed (D4): As for D3.

5. Taxon observation resolution tool

Software needed (S5): A tool that takes a taxon name approved by the name standardization tool (S1), maps it onto a master (regional or global) checklist (S3) and further resolves or verifies the name by using additional information pertaining to locality, date of observation, etc. (use case 3d). All data associated with taxon occurrence records could potentially be used. For example, geocoordinates, date of determination, and identification reference used would help determine which taxon a name associated with an observation should map to. Ideally this application would operate as a web service. Final implementation should include web-based tools to assist user in resolving ambiguities for cases requiring user inspection.

Data needed (D5): Preliminary mapping of nominal concepts of names in the standard name list (D1) onto the master checklist, as inferred from synonym information in various databases, monographs, and regional checklists (D3), as generated using S3. Information on:

1. Taxon occurences within regions. Much literature-base distribution information is currently available within the TROPICOS database (e.g.,

http://www.tropicos.org/NameDistributions.aspx?nameid=3900422), albeit with a New World bias. More global coverage could be provided directly via specimens (e.g., http://www.tropicos.org/NameSpecimens.aspx?nameid=3900422); however, this approach would require extensive additional taxonomic and geographic verification.

2. Taxonomic literature, authors, dates, and the synonymies they represent (some of this already available in TROPICOS (e.g., http://www.tropicos.org/NameReferences.aspx?nameid=3900422), but much additional data capture would be required).
3. People (collectors and authors), taxonomic fields of expertise, and links to taxonomic literature (e.g., TROPICOS Collectors Database, Harvard Names Database).

6. Concept Mapper tool

Software need (S6): Tool for experts to manage and create taxon concept relationships. This might well build on the ConceptMapper tool created by the SEEK project. Ultimately, taxon concepts are the only means of accurately dismbiguating the biological meanings (in terms or traits of distribution) of multiple usages attached to particular taxon name (see R. Peet - Integration of diverse taxon observation records).

Data needed (D6): Extensive data capture needed to extract concepts from taxonomic literature. A start could be made by automated and user-assisted extraction from currently databased taxonomies.

LITERATURE CITED

García, A., 2006. Using ecological niche modelling to identify diversity hotspots for the herpetofauna of Pacific lowlands and adjacent interior valleys of Mexico. Biological Conservation, 130(1): 25-46.

Loarie, S.R. et al., 2008. Climate Change and the Future of California's Endemic Flora. PLoS ONE, 3(6): e2502.

Peterson, A.T. et al., 2002. Future projections for Mexican faunas under global climate change scenarios. Nature, 416: 626-629.

Weiser, M.D. et al., 2007. Latitudinal patterns of range size and species richness of New World woody plants. Global Ecology and Biogeography, 16: 679-688.

APPENDICES. USE CASES AND EXAMPLES

Appendix 1. Why a TNRS?

Why would an ecologist or comparative biologists use at TNRS? I think we envisioned two generic but likely the typical use cases.

(from Brian Enquist's summary notes of breakout group discussion)

Use case #1 – User obtains a list of taxa or a taxon (usually just genus and species but may contain family information). The source of this list or name may vary – it may come from the old or new literature or from the users own recent studies. User wants to take their list or name and ensure that the names are "good names". User generally has minimal knowledge about taxonomy and only wants to use "the right name". User only wants good names to put in their publication and/or to learn about what names they should be using for their work.

Use case #2 – User has two lists of names. User wants to compare these lists of names. For example, the user wants to compare the lists say something about change or similarity of diversity (compare spatial or temporal changes in diversity). Lists may come from different time periods or different geographic regions. In this case the user just wants to "standardize" each list of names so that he or she can then analyze the data using "accepted" names.

Use case #3 – - A more knowledgeable user for use case #1 or #2 but will want the most accepted name informed by the geographic region where these lists are coming from

Use case #4 - A more knowledgeable user for use case #1 or #2 who will want not only an output of accepted names but also (a) a list of all of the synonyms and taxonomic references associated with those cases

Use case #5 - The taxonomically savvy ecologist or the taxonomist who is looking for the most up to date information on taxonomy and synonymy and will want as much of this information as possible.

Use case #6 – Any of the above but with batch processing capabilities.

**Step-by-step applications of increasing complexity.

(1) A website with GUI where user can paste in his or her list of names.
a. Data input is highly regulated where at minimum user must return genus and species separated by spaces, or family, genus, and species (if they don't do this then submitted request barfs, i.e. the user input is highly constrained and unforgiving).

(2) TNRS will then take submitted list and then output to the user
a. a list of matched names and flagged unmatched names.
b. These matched names are the "taxonomically accepted names" with the authority.
The user will then use these accepted names for their analyses.

c. Accepted names would then have a 'time stamp' of when request was approved (such as TROPICOS_3_23_2010).

(3) TNRS will output references associated with the "accepted name" such as the Flora of Ecuador etc.
a. Should this include just the oldest reference, the most recent reference? Both?
b. This will allow user to have information to gain access to relevant taxonomic literature without being overwhelmed.

(4) Option to have TNRS provide a returned list of accepted names plus synonymies.
Plus additional names associated with the "accepted" name
a. This will allow the taxonomically savvy user to access all associated names

(5) Option to have TNRS add Family names to returned synonymies.
a. This will help the user names to help interpret applicability of the returned synonymies.
b. Note – would it be possible to add a geographic scope to returned synonomies?
Levels of priority decisions – The next sort of steps will involve taxonomic intelligence on the returned taxonomic strings that do not match

(6) The TNRS will return, for each submitted string a more detailed return if submitted Family, genus, and species is accepted or not.

For example, the submitted string may have an acceptable family and genus name but not species. The output would
then indicate that the Family and genus names are acceptable but species is not.

(7) The same as #6 but the user will submit a taxonomic string followed by an author.
The output would then parse the string and the author and TNRS would return if the submitted author is accepted or not.
Increasingly complicated. If the names don't match then user will want the TNRS to provide recommendations.

(8) If the family and/or genus and/or species do not fit then:
a. TNRS will assess if the 'bad name' is due to simple spelling mistake (switching i and e, etc.). If such a switch results in an accepted name then the TNRS will return the corrected name
b. If simple spelling correction does not work then TNRS will perform increasingly complex spelling variants

Appendix 2. Name standardization

The purpose of name standardization is to purge misspellings from taxonomic names and provide the single canonical spelling. That is, attempt to match one or more of a list of user-submitted names to a list of published taxonomic names. The goal is not to find the "accepted" name, but rather *any* taxonomic name at all. Name standardization is done with reference to a comprehensive names database such as IPNI or TROPICOS .

Name standardization procedures can range from simple matching , to parsing (and variously classifying additional information extracted), to fuzzy matching.The ideal application would include all three.

2a. Simple batch name matching

Goal: Match list of user-supplied names against merged authority file of all names in TROPICOS and IPNI.

Requirements:

- Reference list of taxonomic names and authors

Caveats

- Does not correct synonyms, simply detects whether a given name in user's list exists in authority file
- The goal is not to update synonymy, but to purge errors and standardize spelling. Only once these errors have been corrected can synonymy be adjusted (as a separate step, not covered by this use case)

General work-flow. User-submits list of raw taxon names at any rank, with or without author. Results returned indicate whether taxon and taxon+author match name in authority file. If taxon matches, rank of taxon is returned as well.

Example:

Names submitted:

taxon	author
Pinus ayacahuite	C. Ehrenb. ex Schltdl.
Pinus pondersa	
Poa annua var. eriolepis	Desv.
Quercus	L.
Fagaceae	

Result returned:

taxon	author	taxon_match	taxon_rank	taxon+author_match
Pinus ayacahuite	C. Ehrenb. ex Schltdl.	1	species	1
Pinus pondersa		0		
ua var. eriolepis	Desv.	1	variety	0
Quercus	L.	1	genus	1
Fagaceae		1	family	

Comments. Spelling errors are so frequent that the utility of simple name matching is limited. Additional manipulations such as parsing and fuzzy matching , followed by user inspection, are required to recover a reasonable fraction of valid names.

2b. Parsing

A more productive approach than simple name matching is to parse the name first by (1) extracting certain common "contaminants" such as indications of uncertainty, and (b) atomizing into genus, specific epithet, rank indicator, infraspecific epithet, etc. After parsing, the atomized name is then matched to a reference list which includes not only species and subspecies but also higher taxa such as genera and families. Even when the full name is misspelled, partial matching to higher taxa can speed discovery of the correct name. An additional advantage of this approach is recovers additional useful or even critical information, such as indications of uncertainty or morphospecies strings.

For an example of an application which uses this approach, see SALVIAS Taxonscrubber.

Goal: Parse list of user-supplied taxon names into name components, extract contaminating information (such as botanical annotations of uncertainty), and match resulting name components against authority file of published names. Interface assists user inspection and correction of remaining unmatched name components.

Requirements:

- Reference list of published taxonomic names
- Library of standard botanical annotations, used to recognize and extract annotations from user-supplied names

Caveats

- Does not correct synonyms, simply detects published names, whether accepted or not.

Comments. This examples below illustrate in the value of recognizing and extracting standard botanical latin annotations such as "cf.". These strings are common in ecological data, are easily recognized, and should be retained as part of the original data. Removing them greatly increases the yield of recognizable names.

General work-flow. User-submits list of raw names, receives list of names atomized into components (Genus, specific_epithet, infraspecific_rank, infraspecific_epithet, author), along with any standard botanical annotations included with names, and remaining unmatched text. Each name component is flagged to indicate if it matches to a standard reference list of published names (e.g., IPNI). Interface assists using in inspecting and correcting any remaining unmatched name components. See detailed examples below.

Example 1: Noel Kempff Savannah Plots, Bolivia.
These data from the SALVIAS database are clean taxonomically but contaminated with annotations and morphospecies.

Table 1.1. Original names. Note (1) the inclusion of family (common in ecological data), and (2) the use of numbers for morphospecies (records 5, 6, and 10), and (3) the single annotation (aff., =latin "affinis": "related to") in record 7.

ID	Family	Species
1	Annonaceae	Annona dioica A. St. Hil.
2	Annonaceae	Duguetia furfuracea (A. St. Hil.) Benth & Hook. F.
3	Guttiferae	Caraipa aff. densifolia Mart.
4	Asteraceae	Riencourtia oblongifolia Gard.
5	Asteraceae	Indet. 5
6	Asteraceae	Indet. 6
7	Bignoniaceae	Tabebuia aff. roseo-alba (Ridley) Sandwith
8	Bignoniaceae	Tabebuia aurea (Manso) Benth & Hook. F. ex Moore
9	Bombacaceae	Eriotheca gracilipes (K. Schum.) Robyns
10	Bombacaceae	Pachira sp.2

Table 1.2. Final output, after match-parse-match. In this case, no taxon names were misspelled. Names 3 and 10 did not match during original round of matching due to contamination with annotations ("aff.") and morphospecies strings ("sp.2"). Removal enabled matching. Note that morphospecies names can be re-formed by concatenating `lowest_taxon_matched` and `unmatched`.

ID	family	genus	specific_ epithet	author	anno-tation	family_ match	genus_ match	species_ match	author_ match	lowest_taxon_matched	unmatched
1	Annonaceae	Annona	dioica	A. St. Hil.		1	1	1	1	Annona dioica	
2	Annonaceae	Duguetia	furfuracea	(A. St. Hil.) Benth & Hook.		1	1	1	1	Duguetia furfuracea	
3	Clusiaceae	Caraipa	densifolia	Mart.	aff.	1	1	1	1	Caraipa densifolia	
4	Asteraceae	Riencourtia	oblongifolia	Gard.		1	1	1	1	Riencourtia oblongifolia	
5	Asteraceae	Indet.	5			1	0	0		Asteraceae	Indet 5
6	Asteraceae	Indet.	6			1	0	0		Asteraceae	Indet 6
7	Bignoniaceae	Tabebuia	roseo-alba	(Ridley) Sandwith	aff.	1	1	1	1	Tabebuia roseo-alba	
8	Bignoniaceae	Tabebuia	aurea	(Manso) Benth & Hook.		1	1	1	1	Tabebuia aurea	
9	Bombacaceae	Eriotheca	gracilipes	(K. Schum.) Robyns		1	1	1	1	Eriotheca gracilipes	
10	Bombacaceae	Pachira	sp.2			1	1	0		Pachira	sp.2

Example 2: Match-parse match of sample of names from the Alwyn Gentry Forest Transect Data Set

This is a diablolical, real-life example of names containing numerous taxonomic problems and other peculiarities. It also illustrates many attributes typical of ecological data, including morphospecies and incomplete determinations. For simplicity I have omitted authors and match flags for infraspecific taxa (for original data, see http://www.mobot.org/MOBOT/research/gentry/transect.shtml).

Table 2.1. Original names from spreadsheet. Note that genus and specific epithet are already in separate fields; subspecific ranks and epithet are concatenated within the specific epithet field (e.g., name #9). This is common in ecological data. Note misspellings (underlined).

ID	Family	Genus	SpecificEpithet
1	INDET	M1	M1
2	LAURACEAE	M1	M1
3	MELASTOMATACEAE	MICONIA CF	M6
4	MELASTOMATACEAE	MICONNIA?	M1
5	POLYGALACEAE	MONINA??	M2
6	MORACEAE	FICUS	CITRIFOLIA CF
7	MORACEAE	PSEUDOLMEDIA CF	M1
8	MORACEAE	SOROCEA	OPIMA
9	SAPOTACEAE	CHRYSOPHYLLUM	ARGENTEUM SSP AURATUM
10	PROTEACEAE	Panopsis?	M1
11	BORAGINACEAE	Cordia	cf. alliodora
12	BORAGINACEAE	Cordia	curasavica
13	BORAGINACEAE	Tournefortia	aff. ternifolia
14	FABACEAE	Erythrina	amazonicum
15	STERCULIACEAE	Dombeya	ankarafantiskae
16	FABACEAE	Leucaena	trichodes

Note the following problems with the above names:

- Spelling errors, including:
  - Incorrect doubling of consonants (e.g. "Miconnia" should be "Miconia", "Monina" should be "Monnina", "curasavica" should be "curassavica"; records 4, 5 & 12)
  - Incorrect gender of specific epithet. Must agree with gender of genus (e.g., "amazonicum" should be "amazonica"; record 14)
  - Interchange of adjacent letters ("ankarafantiskae" should be "ankarafantsikae"; record 15)
- Contamination with:

  - Annotations of uncertainty ("?", "cf.", "aff.").
  - Morphospecies names. Strings which identify a species as locally unique (i.e., within one or a series of plots or collections) when full latin name is unknown. Commonly morphospecies are distinguished by a alphanumeric code or number. In the case of the Gentry data, morphospecies are indicated by an "M" (="Morphospecies") followed by a number (e.g., " M1", M6"); however, the most common usage is "sp" or "sp." followed by a number (e.g., "Cyperus sp.1", "Quercus sp6"). Yet another convention is to reference a particular specimen, using a collector's name plus collection number as the morphospecies string (e.g., Miconia Gentry 56321).
- Incomplete identification. Some plants are determined to genus only (records 3, 4, 5, 7, 9), others to family only (record 2). One record in the above example is unidentified to family (record 1, "INDET" means undeterrmined or unknown).
- Use of family. Family is a higher classification, strictly speaking not necessary when genus or genus plus specific epithet are given. Furthermore, familial classification can vary according to different concepts (e.g., *Dombeya*, above, is now usually placed in the Malvaceae *sensu lato*). However, in the above example that family is essential for specimens only determined to family (.e.,g name #2). Also, family can aid discovery of the correct lower taxon when genus is misspelled.

Table 2.2. Step 1: Match. Sample output after initial round of matching to published names (from IPNI). Only 3 names match to species at this stage.

ID	Family	Genus	SpecificEpithet	family_ match	genus_ match	species_ match	lowest_taxon_matched
1	INDET	M1	M1	0	0	0	
2	LAURACEAE	M1	M1	1	0	0	Lauraceae
3	MELASTOMATACEAE	MICONIA CF	M6	1	0	0	Miconia
4	MELASTOMATACEAE	MICONNIA?	M1	1	0	0	Melastomataceae
5	POLYGALACEAE	MONINA??	M2	1	0	0	Polygalaceae
6	MORACEAE	FICUS	CITRIFOLIA CF	1	1	0	Ficus
7	MORACEAE	PSEUDOLMEDIA CF	M1	1	0	0	Moraceae
8	MORACEAE	SOROCEA	OPIMA	1	1	1	Sorocea opima
9	SAPOTACEAE	CHRYSOPHYLLUM	ARGENTEUM SSP AURATUM	1	1	0	Chrysophyllum
10	PROTEACEAE	Panopsis?	M1	1	0	0	Panopsis
11	BORAGINACEAE	Cordia	cf alliodora	1	1	0	Cordia
12	BORAGINACEAE	Cordia	curasavica	1	1	0	Cordia
13	BORAGINACEAE	Tournefortia	ternifolia	1	1	1	Tournefortia ternifolia
14	FABACEAE	Erythrina	aff amazonicum	1	1	0	Erythrina
15	STERCULIACEAE	Dombeya	ankarafantiskae	1	1	0	Dombeya
16	FABACEAE	Leucaena	trichodes	1	1	1	Leucaena trichodes

Table 2.3. Steps 2 & 3: parse & match again. Sample output after atomization and parsing of names from Table 2, followed by second round of matching. Six names now match to species and two previously unmatched genera have been recovered.

ID	family	genus	specific_ epithet	infra_ rank	infraspecific_ epithet	anno- tation	family_ match	genus_ match	species_ match	lowest_taxon_matched	unmatched
1	INDET	M1	M1				0	0	0		INDET M1 M1
2	Lauraceae	M1	M1				1	0	0	Lauraceae	M1 M1
3	Melastomataceae	Miconia	M6				1	1	0	Miconia	M6
4	Melastomataceae	Miconnia	M1				1	0	0	Melastomataceae	Miconnia M1
5	Polygalaceae	Monina	M2			cf.	1	0	0	Polygalaceae	Monina M2
6	Moraceae	Ficus	citrifolia			cf.	1	1	1	Ficus citrifolia	
7	Moraceae	Pseudolmedia	M1				1	1	0	Pseudolmedia	M1
8	Moraceae	Sorocea	opima				1	1	1	Sorocea opima	
9	Sapotaceae	Chrysophyllum	argenteum	subsp.	auratum		1	1	1	Chrysophyllum argenteum subsp. auratum	
10	Proteaceae	Panopsis	M1				1	1	0	Panopsis	M1
11	Boraginaceae	Cordia	alliodora			cf.	1	1	1	Cordia alliodora	
12	Boraginaceae	Cordia	curasavica				1	1	0	Cordia	curasavica
13	Boraginaceae	Tournefortia	ternifolia				1	1	1	Tournefortia ternifolia	
14	Fabaceae	Erythrina	amazonium			aff.	1	1	0	Erythrina	amazonicum
15	Sterculiaceae	Dombeya	ankarafantiskae				1	1	0	Dombeya	ankarafantiskae
16	Fabaceae	Leucaena	trichodes				1	1	1	Leucaena trichodes	

Note the following:

1. Annotations extracted to separate fields
2. Infraspecific taxa concatenated with specific_epithet detected and atomized to two fields, one indicating rank and a second containing the epithet
3. Unmatched content dumped to separate field
4. Question mark "?" treated as equivalent to latin "cf." ("confer": compare or consult)
5. The majority of no-matches from Table 2 were in due to contamination with annotations and morphospecies strings. Spelling errors are restricted to records 4, 5, 12, 14, 15.

Further correction of misspellings (in records 4, 5, 12, 13, and 15; see Table 2.1) requires inspection by user. See Table 4, below.

Table 2.4. User correction of remaining errors. Final result after match-parse-match followed by manual inspection and correction of remaining spelling errors from Table 3. Nine names now match to species. The remaining unmatched name components are morphospecies, as intended. There are no remaining errors in this list of names.

ID	family	genus	specific_epithet	infra_rank	infraspecific_epithet	anno-tation	family_match	genus_match	species_match	lowest_taxon_matched	unmatched
1	INDET	M1	M1				0	0	0		INDET M1 M1
2	Lauraceae	M1	M1				1	0	0	Lauraceae	M1 M1
3	Melastomataceae	Miconia	M6				1	1	0	Miconia	M6
4	Melastomataceae	Miconia	M1				1	1	0	Miconia	M1
5	Polygalaceae	Monnina	M2			cf.	1	1	0	Monnina	M2
6	Moraceae	Ficus	citrifolia			cf.	1	1	1	Ficus citrifolia	
7	Moraceae	Pseudolmedia	M1				1	1	0	Pseudolmedia	M1
8	Moraceae	Sorocea	opima				1	1	1	Sorocea opima	
9	Sapotaceae	Chrysophyllum	argenteum	subsp.	auratum		1	1	1	Chrysophyllum argenteum subsp. auratum	
10	Proteaceae	Panopsis	M1				1	1	0	Panopsis	M1
11	Boraginaceae	Cordia	alliodora			cf.	1	1	1	Cordia alliodora	
12	Boraginaceae	Cordia	curassavica				1	1	1	Cordia curassavica	
13	Boraginaceae	Tournefortia	ternifolia				1	1	1	Tournefortia ternifolia	
14	Fabaceae	Erythrina	amazonicum			aff.	1	1	1	Erythrina amazonicum	
15	Sterculiaceae	Dombeya	ankarafantsikae				1	1	1	Dombeya ankarafantsikae	
16	Fabaceae	Leucaena	trichodes				1	1	1	Leucaena trichodes	

Options for acceleratiing or automating the final step of user inspection and correction:

1. Assisted manual inspection. Gui uses membership in higher taxa to reduce the number of options. For example, a pick list of all species within a given genus (e.g., for record 14, user could choose from a list of all species in Erythrina, and would quickly discover the correct spelling, Erythrina amazonicum). See the "Hand scrub" form of Taxonscrubber for an example of assisted manual inspection.
2. Fuzzy matching of unmatched names, followed by presentation to the user of a list of possible matches ranked by match scores. In the case of record 14, *Erythrina amazonicum* (correct spelling) should be the closest match (i.e., highest rank) to the *Erythrina amazonica* (incorrect original spelling). This process could be further automated by automatically accepted the highest scoring matches, within a certain tolerance.

Note that morphospecies strings can be re-constructed by concatenating `lowest_taxon_matched` and `unmatched`.

2c. Fuzzy matching

Goal: Match list of user-supplied names against authority file of published names. For any non-matching names, provide ranked list of nearest matches, with option to accepted automatically the nearest match above a certain minimum match threshold.

Requirements:

- Reference list of published taxonomic names and authors

Caveats

- Does not correct synonyms, simply detects closest matching published name

General work-flow. User-submits list of raw names, with or without author. Receives list of closest-matching names, along with match_rank scores for each name component (genus, species, subspecific taxa if applicable, author). If one and only one name matches 100%, then only one name is returned. If no name matches 100%, or if >1 name matches 100% (would happen in case of homonym when no author provided with original name), then multiple nearest-matching names are returned. Interface provides mean for user to review best matches for each name and select the desired name, or to automatically select the best match, above a certain minimum threshold. Names with no matches above the minimum threshold would return "no match".

Example:
Names submitted (third names is a misspelling of Pinus ponderosa P. & C. Lawson]:
Pinus ayacauite C. Ehrenb. ex Schltdl.
Pinus jaliscana
Pinus pondesa Lawson

Results returned:

name_submitted	genus	genus_ match_score	species	species_ match_score	author	author_ match_score	mean_ match_score
Pinus ayacauite C. Ehrenb. ex Schltdl.	Pinus	1.00	Pinus ayacauite	1.00	C. Ehrenb. ex Schltdl.	1.00	1.00
Pinus jaliscana	Pinus	1.00	Pinus jaliscana	1.00	Pérez de la Rosa		1.00
Pinus pondesa Lawson	Pinus	1.00	Pinus ponderosa	0.95	P. & C. Lawson	0.62	0.86

Pinus pondesa Lawson	Pinus	1.00	Pinus polita	0.15	(Siebold & Zucc.) Antoine	0.04	0.39
Pinus pondesa Lawson	Pinus	1.00	Pinus orientalis	0.07	L.	0.05	0.37

In the above example, if user set acceptance criterion to "nearest match with >.70 match score" then "Pinus ponderosa P. & C. Lawson" would be accepted as the only canonical near-match for "Pinus pondesa Lawson".

Comments. Match scores in the above example are made up. mean_match_score is simply the arithmetic mean of the genus, species and author score. This is probably not an optimum algorithm. Should probably down-weight author score, considering how frequently authors are misspelled, also the existence of very different ways of indicating the same author (abbreviated and spelled in full).

For a working example of fuzzy matching, see:

Tony Rees's TaxaMatch:

http://www.cmar.csiro.au/datacentre/irmng/
http://code.google.com/p/taxon-name-processing/wiki/TaxamatchInfo

For a php/mysql implementation of TaxaMatch, see:

http://www.silverbiology.com/products/taxamatch/

Appendix 3. Synonymy

Resolving synonymy involves both (a) finding the accepted name for a non-accepted (synonym) name, and (2) discovering all synonyms associated with a particular name, whether accepted or not.

The purpose of **synonym correction** is to label each of a list of names as either accepted or a synonym, and, if the name is a synonym, provides the accepted (correct) name. Other name status values include "invalid", "illegitimate", "of uncertain status", etc.

Synonyms can be **simple** (the synonym matches to one and only one accepted name at the taxonomic level of interest) or **complex** (the synonym potentially matches to more than one name). The latter would happen in the case of a name matching to a species which has been split into more than one species. This can even be true if the species name itself is valid, if one or more of it's subspecies has been elevated to the level of full species. Such cases require additional information to be machine resolvable, or should be flagged for user inspection.

Other sources of ambiguity include conflicting taxonomic concepts (one author regards a species are valid, whereas another considers it a synonym). Such ambiguity can become quite complex in the case of multiple taxonomic splits and multiple taxonomic opinions. In extreme cases, name linkages can even be reticulate or circular (e.g., author 1 says synonym A = accepted name B, author 2 says synonym B = accepted name C, author 3 say synonym C = accepted name A).

Although name matching can be performed during the synonym correction stage, it is a sufficiently complex task on its own and should best be kept as a separate process.

3a. Single authoritative synonymy

Goal: Find the accepted name for one or more user-submitted taxon names, according to an authoritative synonymized checklist.

Requirements:

- User-submitted list of published taxon names. By published, I mean that names have already been checked against authoritative global list of published names and any errors corrected (see Name Stadardization Use Cases)
- Reference list of accepted and synonymized taxon names, with links to accepted name(s) for any synonyms
- This example assumes a single authoritative synonymized list.

Assumptions:

- User is submitting a name at rank of genus or below

Caveats:

- Only names and authorities but no references are submitted, therefore only name relationships, not concepts, are returned.

General comments. Because we are assuming user names have already been matched to an existing published name (see Name matching use cases), we don't need to consider such complications as inclusion of higher taxa, contamination by extraneous data, degree of atomization of the name submitted, etc. Here, the user is just submitting a single string, or better yet an LSID pointing to a "known" name or name+author combination. It is still possible for a name to not match if that name is not included in the particular synonymy being used if the synonymy being used is not global and exhaustive. In this case the user will know that the non-match is not due to a spelling error.

General work-flow. User submits one or more names, and either (a) chooses one of a list of possible synonymies, or (b) application provides single authoritative list only (e.g. synonymized world checklist). For each name submitted, the application returns the exact name matched (at the lowest taxonomic level), if the matched name is accepted or a synonym, and, for synonyms, the accepted name. Results are atomized to each name component, allowing user to make use of partial matches. The latter is critical as in most cases, we are interested only in the accepted species. For each name, application returns the following:

- **name_submitted:** verbation taxon name submitted by user
- **full_match**: whether the *submitted* name matches completely, partially (e.g., to genus but not to species), or not at all. Values: full, partial, nomatch.
- **match_type**: relationship of submitted name to matched name. Values: = (names exactly match), > (submitted name is a higher taxon containing lower name), < (submitted name is contained by matched name, i.e., partial match to higher taxon only), NULL
- **match_status**: indicates if matched name is accepted or a synonym. Values: acc, syn, NULL.
- **match_gen**: genus of matched name.
- **match_sp**: specific epithet of matched name.
- **match_rank**: rank of matched infraspecific taxon, if applicable. Values: subsp., var., fo.
- **match_infra_ep** epithet of matched infraspecific taxon, if applicable.
- **match_auth**: author of the lowest matched taxon.
- **acc_gen**: accepted genus.
- **acc_sp**: accepted specific epithet.
- **acc_rank**: rank of accepted infraspecific taxon, if applicable. Values: subsp., var., fo.
- **acc_infra_ep** epithet of accepted infraspecific taxon, if applicable.

- **acc_auth**: author of the lowest matched taxon.

Examples:

Example 1. The following is based on Cam Webb's Asian Plant Synonym Lookup and his own example from this wiki (thanks Cam!). I have added a neotropical tree (Ceiba pentandra) to show an example of a valid name not in the reference synonymy. For simplicity I have omitted the "forma" infraspecific taxa for Abarema clypearia. Just assume that Abarema clypearia has only two infraspecific taxa, one variety and one subspecies.
User submits the following list of names:

Hopea grisea
Shorea gibbosa
Abarema clypearia
Abarema clypearia ssp hirsutus
Abarema clypearia var angulata
Ceiba pentandra

…and selects "Asian plant synonymy" (reference list should provide links to references and sources). Application returns the following:

name_submitted	**full_ match**	**match_ type**	**match_ status**	**match_ gen**	**match_sp**	**match_ rank**	**match_ infra_ep**	**match_auth**	**acc_genus**	**acc_ species**	**acc_ rank**	**acc_ infra_ep**	**acc_auth**
Hopea grisea	full	=	syn	Hopea	grisea			Brandis	Shorea	gibbosa			Brandis
Shorea gibbosa	full	=	acc	Shorea	gibbosa			Brandis	Shorea	gibbosa			Brandis
Abarema clypearia	full	=	syn	Abarema	clypearia			(Jack) Kosterm.	Archidendron	clypearia			(Jack) Nielsen
Abarema clypearia	full	>	syn	Abarema	clypearia	subsp.	velutina	(Merr.&Perry) Verdc.	Archidendron	clypearia	subsp.	velutina	Nielsen
Abarema clypearia	full	>	syn	Abarema	clypearia	var.	angulata	(Benth.) Kosterm.	Archidendron	clypearia	var.	angulata	(Jack) Nielsen
Abarema clypearia ssp hirsutus	partial	<	syn	Abarema	clypearia			(Jack) Kosterm.	Archidendron	clypearia			(Jack) Nielsen
Abarema clypearia var angulata	full	=	syn	Abarema	clypearia	var.	angulata	(Benth.) Kosterm.	Archidendron	clypearia	var.	angulata	(Jack) Nielsen
Ceiba pentandra	nomatch												

Comments:

- Note that some names return more than one value (e.g., Abarema clypearia). This is because application returns all infraspecific taxa as well for a given species name. Submitted name could match to any of them, unless nominate infraspecific taxon was intended (in which case, this should be indicated in submitted name, e.g., Abarema clypearia subsp. clypearia). Nominate infraspecific taxa should never be assumed.
- In this case, although Abarema clypearia matches to one specific taxon and plus two infraspecific taxa, it still returns a single accepted species (Archidendron clypearia). This will not always be the case however. See below for a more complex example.
- Note the partial match for Abarema clypearia ssp hirsutus. This could be because name is misspelled (if names have not be scrubbed prior to submission) or name is not in reference synonymy. If user is only interested in accepted species, this partial match should be adequate.
- Ceiba pentandra is a valid name, but does not occur in reference synonymy.

Example 2. User submits the following names and chooses to have them synonymized according to USDA Plants Checklist. To save space, I have concatenated name components (Genus, specific epithet, etc) for matched and accepted names.

Acalypha virginica
Pinus arizonica

Application returns the following:

name_submitted	full_ match	match_ type	match_ status	match_ taxon	match_auth	acc_taxon	acc_auth
Acalypha virginica	full	=	acc	Acalypha virginica	L.	Acalypha virginica	L.
Acalypha virginica	full	>	acc	Acalypha virginica var. viriginica		Acalypha virginica	L.
Acalypha virginica	full	>	syn	Acalypha virginica var. deamii	Weath.	Acalypha deamii	(Weath.) H.E. Ahles
Acalypha virginica	full	>	syn	Acalypha virginica var. gracilens	(A. Gray) Müll. Arg.	Acalypha gracilens	A. Gray
Acalypha virginica	full	>	syn	Acalypha virginica var. monococca	(Engelm. ex A. Gray) Müll. Arg.	Acalypha monococca	(Engelm. ex A. Gray) Lill. W. Mill. & Gandhi
Acalypha virginica	full	>	syn	Acalypha virginica var. rhomboidea	(Raf.) Cooperr.	Acalypha rhomboidea	Raf.
Pinus arizonica	full	=	acc	Pinus arizonica	Engelm.	Pinus arizonica	Engelm.
Pinus arizonica	full	>	acc	Pinus arizonica var. arizonica		Pinus arizonica var. arizonica	
Pinus arizonica	full	>	acc	Pinus arizonica var. stormiae	Martiñez	Pinus arizonica var. stormiae	Martiñez

Comments:

- *Pinus arizonica* is unproblematic. Even though the species matches to three taxa (the species, one variety plus the nominate variety), all equate to the same accepted species, Pinus arizonica.
- *Acalypha virginica* is complicated. Various subspecies have been elevated to the rank of species. Therefore, even though the name Acalypha virginica is still accepted, it refers only to the narrowest sense indicated by the nominate variety. The user will need additional information to determined which meaning was intended. This is a good example of a complex synomym that cannot be resolved programmatically without additional information.

3b. Simple hierarchy of synonymies

Goal: Find the accepted name for one or more user-submitted taxon names, according to a user-ordered hierarchy of referemce synonymies. Reference synonymies can be either monographic or regional.

Requirements:

- User-submitted list of taxon published taxon names. Names have already been checked against authoritative global list of published names and any errors corrected (see Name matching use cases).
- Multiple reference lists of accepted and synonymized taxon names, with links to accepted name(s) for any synonyms. Lists can be monographic (e.g., World Checklist of Monocots, Revision of Lecythidaceae), regional (Flora of Peru Checklist, USDA Plants Checklist for USA & Canada) or both (Palms of the Americas).

Assumptions:

- User is submitting a name at rank of genus or below

General work-flow. User is presented with a list of checklists and monographic synonymies and selects one or more to be used to standardize his/her list of names. Chosen synonymies are then ordered by user, from top priority to lowest priority ("apply first" to "apply last"). In theory, the first list should be the highest overall quality and the last list the lowest, meaning that monographic should be preferred to regional. The user then submits a list of taxon names. Application checks each name against the list of reference synonymies, starting with the first list. If a name is found on the first list, then that name is synonymized according to that list and ignored in all subsequent lists. If the name is not in the first list, then it is checked against the second list. If it is not found in the second list, it is checked agaiin the third list. And so on, down to the last list. All submitted names are synonymized in this fashion. Some names may not be on any list. Conflicts between lists are not resolved, as a given name is only synonymized according to a single list.
chooses one of a list of possible synonymies, or (b) application provides single authoritative list only (e.g. synonymized world checklist). For each name submitted, the application returns the exact name matched (at the lowest taxonomic level), if the matched name is accepted or a synonym, and, for synonyms, the accepted name. Results are atomized to each name component, allowing user to make use of partial matches. The latter is critical as in most cases, we are interested only in the accepted species.
For each name, application returns the following:

- **name_submitted:** verbation taxon name submitted by user
- **match_source**: name or code of list used to check name
- **full_match**: whether the *submitted* name matches completely, partially (e.g., to genus but not to species), or not at all. Values: full, partial, nomatch.

- **match_type**: relationship of submitted name to matched name. Values: = (names exactly match), > (submitted name is a higher taxon containing lower name), < (submitted name is contained by matched name, i.e., partial match to higher taxon only), NULL
- **match_status**: indicates if matched name is accepted or a synonym. Values: acc, syn, NULL.
- **match_gen**: genus of matched name.
- **match_sp**: specific epithet of matched name.
- **match_rank**: rank of matched infraspecific taxon, if applicable. Values: subsp., var., fo.
- **match_infra_ep** epithet of matched infraspecific taxon, if applicable.
- **match_auth**: author of the lowest matched taxon.
- **acc_gen**: accepted genus.
- **acc_sp**: accepted specific epithet.
- **acc_rank**: rank of accepted infraspecific taxon, if applicable. Values: subsp., var., fo.
- **acc_infra_ep** epithet of accepted infraspecific taxon, if applicable.
- **acc_auth**: author of the lowest matched taxon.

General comments. The idea behind the multi-list approach is to cover as many taxa as possible----by quilting together existing region-specific (floras, checklists) and taxon-specific (mongraphic) synonymies. The hierarchy method is the simplest possible algorithm for applying multiple synonymies. Despite its shortcomings, it is very simple to use and interpret. However, it does not necessarily provide the best or most up-to-date taxonomic understanding for a given taxon, and does not alert the user to conflicts or ambiguities between lists (this information may be informative).

Example:

1. User submits the following list of names:

Macoubea witotorum
Macoubea guianensis
Geonoma sodiroi
Socratea exorrhiza
Socratea durissima
Abarema clypearia
Ceiba pentandra
Tecoma stans
Pinus arizonica

2. User chooses reference synonymies to be used, ordered as follows:

priority	checklist_ code	checklist
1	PALM	Palms of the Americas (Henderson 1995)
2	PERU	Peru Project Checklist (Brako & Zarucchi 1993)
3	USDA	USDA Plants (USA & Canada)

3. Application returns the following (authorities omitted to save space):

name_submitted	checklist_ code	full_match	match_ type	match_ status	match_ gen	match_sp	match_ rank	match_ infra_ep	acc_genus	acc_ species	acc_ rank	acc_ infra_ep
Macoubea witotorum	PERU	full	=	syn	Macoubea	witotorum			Macoubea	guianensis		
Macoubea guianensis	PERU	full	=	acc	Macoubea	guianensis			Macoubea	guianensis		
Geonoma sodiroi	PALM	full	=	syn	Geonoma	sodiroi			Geonoma	cuneata		
Socratea exorrhiza	PALM	full	=	acc	Socratea	exorrhiza			Socratea	exorrhiza		
Socratea durissima	PALM	full	=	syn	Socratea	durissima			Socratea	exorrhiza		
Abarema clypearia		nomatch										
Ceiba pentandra	PERU	full	=	acc	Ceiba	pentandra			Ceiba	pentandra		
Tecoma stans	PERU	full	=	acc	Tecoma	stans			Tecoma	stans		
Tecoma stans	PERU	full	=	syn	Tecoma	stans	var.	velutina	Tecoma	stans	var.	stans
Pinus arizonica	USDA	full	=	acc	Pinus	arizonica			Pinus	arizonica		

Notes:

- Abarema clypearia (a southeast Asian tree) is not on any of the chosen reference synonymies and therefore returns no results
- Each name is checked according to one list only: the highest-priority list on which it appears. Even though some names are found on more than one list (for example, Tecoma stans is in both USDA Plant and the Peru checklist), only the first occurrence is reported.

3c. Dynamic checklist

Goal: Taxonomic names are corrected by consulting multiple synonymized checklists and databased taxonomies. Application chooses one and only one concept for each name according to a series of transparent decision rules, thus effectivly compiling a dynamic, authoritative regional or global checklist. Application should be capable of displaying multiple opinions for a particular name, deducing relationships between these concepts, and reporting agreements, conflicts, and ambiguities.

Requirements:

- Access to names, synonymy and taxonomic literature source (including dates) in TROPICOS database
- Access to taxonomic checklists and literature citations from other sources (e.g., Kew Monocots of the world checklist, ITIS synonymy)
- Ability to ingest additional sources of synonymy and taxonomic literature as they become available

Assumptions:

- Names submitted by user and names in all taxonomic sources have already been standardized again authoritative global list of names (see Name matching).

General work-flow. Uses submits a list of names, and the application provides the single correct name for each name submitted (according to the selection algorithm used), and the source upon which that taxonomic decision was made. User can optionally see all opinions for a particular name if necessary, and choose to apply a concept different from the one provided. Application flags names with potential ambiguities caused by homonyms, splitting off of subordinate taxa, etc.

General comments. Some of this functionality is already available via the TROPICOS interface, for a single name at a time. For a particular name, TROPICOS reports any names which are regarded as synonyms of that name, and any names which are regarded as accepted replacements for that name (i.e., when the original name is a synonym). It also provides the literature source of each taxonomic opinion. However, as TROPICOS is agnostic, multiple conflicting opinions can be presented for a particular name, and it is up to the user to choose which one is correct.

For example a search on the name *Ceiba mandonii* Britten & Baker f. in TROPICOS returns one synonym and three accepted names:

Synonyms:

- *Ceiba speciosa* A. St.-Hil. ex Brako
    - Saravia, E. F. 1996. Estud. Veg. Prov. Campero Mizque Cochabamba i--92.

Accepted names:

- *Ceiba boliviana* Britten & Baker f.
  - Gibbs, P. & J. Semir 2003. A taxonomic revision of the genus *Ceiba* Mill. (Bombacaceae). Anales Jard. Bot. Madrid 60(2): 259--300.
- *Ceiba pubiflora* (A. St.-Hil.) K. Schum.
  - Macbride, J. F. 1956. Bombacaceae, Flora of Peru. Field Mus. Nat. Hist., Bot. Ser. 13(3A/2): 477-478/593-622.
- *Ceiba speciosa* A. St.-Hil. ex Brako
  - Brako, L. & J. L. Zarucchi 1993. Catalogue of the Flowering Plants and Gymnosperms of Peru. Monogr. Syst. Bot. Missouri Bot. Gard. 45: i--1286.

Which name is correct? The simple approach of "use the most recent taxonomic opinion" would select *Ceiba boliviana* Britten & Baker f. as the correct name for *Ceiba speciosa*, according to Gibbs, P. & J. Semir 2003.

Example:

User submits the following list of names, one with author, the remainder without:

Ceiba speciosa
Ceiba mandonii
Ceiba pubiflora (A. St.-Hil.) K. Schum.
Acalypha virginica

Application returns:

name_submitted	status	accepted_name	source	ambig	ambiguous_reason
Ceiba speciosa	acc	*Ceiba speciosa* (A. St.-Hil.) Ravenna	Gibbs, P. & J. Semir 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). Anales Jard. Bot. Madrid 60(2): 259--300.	1	homonym
Ceiba mandonii	syn	*Ceiba boliviana* Britten & Baker f.	Gibbs, P. & J. Semir 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). Anales Jard. Bot. Madrid 60(2): 259--300.	0	
Ceiba pubiflora (A. St.-Hil.) K. Schum.	acc	*Ceiba pubiflora* (A. St.-Hil.) K. Schum.	Gibbs, P. & J. Semir 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). Anales Jard. Bot. Madrid 60(2): 259--300.	0	
Acalypha virginica	acc	*Acalypha virginica* L.	Gleason, H. A. & A. Cronquist 1991. Man. Vasc. Pl. N.E. U.S. (ed. 2) i-lxxv, 1-910.	1	homonym; pro-parte synonymy via subordinate taxa

Notes:

1. Ceiba speciosa, while accepted according to the most recent taxonomic concept, was flagged as potentially ambiguous due to the existence of the homonym *Ceiba speciosa* A. St.-Hil. ex Brako (actually a nomen nudum, probably mis-applied to *Ceiba mandonii*). The application has assumed the nomenclaturally valid *Ceiba speciosa* (A. St.-Hil.) Ravenna, but the user have additional information indicating the homonym, not the correct name, was intended. Note that had the user intended the nomenclaturally correct name and included the authority "(A. St.-Hil.) Ravenna ", this name would not have been flagged as ambiguous.
2. Acalypha virginica, an accepted name in the strict sense, was flagged as ambiguous for two reasons: (1) the existence of two posterior homonyms, *Acalypha virginica* Wall. and *Acalypha virginica* Michx., and (2) the existence of 4 subordinate taxa which have been elevated to the rank of full species, according to recent taxonomic concepts (even though a species name is "accepted", it may not be the correct name if it could belong to a subspecies that was transferred to a different species). The user would then have to decide whether the recent strict sense of the name was intended (in which case the correct name is indeed Acalypha virginica) or if the older broader meaning,

including the now-split varieties, was intended. In the latter case, the name may not be resolvable to species without additional information (locality of collection, re-identification of original specimens).

3. If this were an interactive display, the user would click on the hyperlinked terms under "ambiguous_reason" to view additional information and links to literature, taxa, etc. on TROPICOS or other sources.

3d. Dynamic checklist + taxon observations

(Please read Synonym correction using dynamic checklist first).

Goal: Correct not just taxon names but taxon observations (a taxon observed at a particular place and time). Taxonomic names are verified and corrected as in Dynamic checklist . However, information pertaining to place and time of observation of the taxon, as well as identification details (by whom, when, with what reference) are optionally included with the name by the user, and may be used by the application to further resolve ambiguities and confirm decisions.

Requirements: As for Dynamic checklist , plus:

- Input is a list of unique taxon observations (name x region x time combinations). Thus the same name may thus occur in multiple records.
- Access to Dynamic Master Checklist based on all databased taxonomies within TROPICOS, plus additional sources of regional and monographic synonymy. Checklist algorithm choses single "accepted" name based on transparent selection criteria (usually, most recent monographic taxonomy)
- Access to additional information within TROPICOS database, including taxonomic literature (and databased synonymies from those references), collectors names and taxonomic specialties, and specimen and literature-based distributional information.

Assumptions: Names submitted by user and names in all taxonomic sources have already been standardized again authoritative global list of names (see Name matching).

General comments. See comments about current TROPICOS functionality under Dynamic checklist . In addition to names, synonymies and taxonomic literature, information on collectors, authors, fields of taxonomic specialization, and data on distributions of taxa is currently available for many taxa in TROPICOS database. All this information could in theory be used to perform programmatically the decisions demonstrated in the following example. This would be a spectacular implementation of "Taxonomic Intelligence".

General work-flow. Uses submits a list of names, and the application provides the single correct name for each name submitted, and the source upon which that taxonomic decision was made. In addition, application checks accepted name against database of name x region occurences and alerts user if taxon is know from region. Region may also be used to resolve ambiguities, for example if only one of a possible series of names is know from the region where the taxon was observed. Date of taxon observation (or date of identification, if supplied) can be further used to deduce or rule out particular taxonomic concepts (for example, if the plant was identified prior to the publication of a particular concept, that concept can be ruled out as the intended meaning. Identification references would further help to identify intended concepts.
User can optionally see all opinions for a particular name if necessary, and choose to apply a concept different from the one provided. Application flags accepted names with potential ambiguities caused by splitting off of subordinate taxa (i.e., even though a species name is "accepted", it may not be the correct name if it could belong to a subspecies that was transferred to a different species). Links are provided to supporting information used to make particular decisions, and to assist users in resolving ambiguities.

Example:

User submits the following list of taxon observations. Note that each record represents not just a taxon, but a taxon observation, consisting of a name accompanied by one or more pieces of information pertaining to locality and date of observation, determiner, determination date and reference.

ID	taxon	country	state_ prov	lower_ politcal	decimal_ latitude	decimal_ longitude	collection_ date	det_by	det_date	det_ reference
1	Socratea exorrhiza				10.1822	-83.5366				
2	Miconia tetraspermoides	Peru								
3	Miconia tetraspermoides	Bolivia								
4	Miconia tetraspermoides	Brazil								
5	Zauschneria californica ssp. angustifolia	USA	Arizona	Pima					4-Mar-08	Kearney, TH, & Peebles RH. 1960. Arizona Flora. University of California Press. ISBN 0520006372.
6	Ceiba speciosa	Peru					5-Jun-94		4-Jul-94	Brako, L. & J. L. Zarucchi 1993. Catalogue of the Flowering Plants and Gymnosperms of Peru. Monogr. Syst. Bot. Missouri Bot. Gard. 45: i--1286.
7	Ceiba speciosa	Boliva							7-Jul-07	Gibbs, P. & J. Semir 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). Anales Jard. Bot. Madrid 60(2): 259--300.
8	Ceiba speciosa	Boliva					7-Dec-10	P. Gibbs		

Application returns the following (link to previous by ID; Note hyperlinks to source information in TROPICOS).

ID	name_submitted	match_ type	matched_ name_ status	matched_name	in_ country	in_ state_ prov	in_ lower_ polit.	accepted_name	ambig	message
1	Socratea exorrhiza	full	acc	Socratea exorrhiza (Mart.) H. Wendl.	1			Socratea exorrhiza (Mart.) H. Wendl.	0	country match by coordinates
2	Miconia tetraspermoides	full	acc	Miconia tetraspermoides Wurdack	0			Miconia tetraspermoides Wurdack	1	Not in country!
2	Miconia tetraspermoides	fuzzy	acc	Miconia tetrasperma Gleason	0			Miconia tetrasperma Gleason	1	Nearest match in region
3	Miconia tetraspermoides	full	acc	Miconia tetraspermoides Wurdack	0			Miconia tetrapermoides Wurdack	1	Not in country!
3	Miconia tetraspermoides	fuzzy	acc	Miconia tetrasperma Gleason	0			Miconia tetrasperma Gleason	1	Nearest match in region
4	Miconia tetraspermoides	full	acc	Miconia tetraspermoides Wurdack	1			Miconia tetraspermoides Wurdack	0	
5	Zauschneria californica ssp. angustifolia	full	syn	Zauschneria californica subsp. angustifolia D.D. Keck	1	1	1	Epilobium canum subsp. angustifolium (D.D. Keck) P.H. Raven	0	Synonymy supported by det_reference+det_date
6	Ceiba speciosa	full	acc	Ceiba speciosa (A. St.-Hil.) Ravenna	1			Ceiba speciosa (A. St.-Hil.) Ravenna	1	see also homonym
6	Ceiba speciosa	full	homonym	Ceiba speciosa A. St.-Hil. ex Brako	1			Ceiba boliviana Britten & Baker f.	1	det_reference+region suggest homonym
7	Ceiba speciosa	full	acc	Ceiba speciosa (A. St.-Hil.) Ravenna	1			Ceiba speciosa (A. St.-Hil.) Ravenna	0	concept supported by det_date+det_reference
8	Ceiba mandonii	full	syn	Ceiba mandonii Britten & Baker f.	1			Ceiba boliviana Britten & Baker f.	0	concept supported by det_date+determiner

Details:

- Record ID #1: Determination (Socratea exhorrhiza) confirmed by occurrence in known country of distribution. Country determined by supplied coordinates.
- Records IDs #2 & 3: Submitted name (Miconia tetraspermoides) is accepted taxonomically, but outside known distribution (not recorded from Peru or Bolivia). Furthermore, a close near-match (Miconia tetrasperma) is know to occur in each region. Therefore, both two results are returned for each record (exact match outside distribution, and near match in region) and the results flagged as ambiguous. User must decide which to use.

- Record ID #4: Name submitted (Miconia tetraspermoides) matches accepted name and is also within known from Brazil.
- Record ID #5: Name submitted (Zauschneria californica ssp. angustifolia) matches to synonym, taxon represented by accepted name (Epilobium canum subsp. angustifolium) is known from lowest political division supplied (Pima County). Old reference and determination date support this conclusion (Zauschneria californica was name in use at time of publication). This is a straightforward taxonomic synonym; records is **not** flagged as ambiguous.
- Record D #6. This is tricky. Name submitted (Ceiba speciosa) matches to accepted name of taxon recorded from region (Ceiba speciosa (A. St.-Hil.) Ravenna). However, it also matches the homonym Ceiba speciosa A. St.-Hil. ex Brako, a nomen nudum incorrectly perpetuated in one of the primary checklists for the region (Brako and Zarucchi 1993) as a synonym of Ceiba mandonii. Thus the determination reference and the date it was used (1994) suggest that the intended meaning may have been C. mandonii. If so, then C. mandonii should be updated to the current accepted name, Ceiba boliviana. This case clearly warrants further inspection by the user, and is flagged as ambiguous.
- Record ID #7. Name submitted (Ceiba speciosa) matches accepted name Ceiba speciosa (A. St.-Hil.) Ravenna. Taxon represented by this name is know within region, and, despite existence of homonym (see record #6 above), determination reference and date of determination support the conclusion that the accepted name was the intended meaning.
- Record ID #8: Name submitted (Ceiba mandonii) is taxonomic synonym of Ceiba boliviana Britten & Baker f. Locality of observation, determination date and determiner (P. Gibbs, author of latest monograph) all support this concept.