Resolving Taxonomic Names using Evidence Extracted from Text

Yujie Cao[1,2], Nico Franz[3], James Macklin[4], Jin Mao[1,2], Hong Cui[2 *]

[1]Center for Studies of Information Resources, Wuhan University, Wuhan, China
[2]School of Information, University of Arizona, Tucson AZ USA
[3]School of Life Sciences, Arizona State University, Phoenix AZ USA
[4]Agriculture Agri-Food Canada, Ottawa Canada

*hongcui@email.arziona.edu

Abstract Biological taxonomy is established on organism relationships with scientific names as the primary identifiers; however, resolving various taxonomic names remains one of the greatest challenges in taxonomy and systematic biology overall. We proposed an evidence-based approach that extracts trait (character) evidence from published literature to facilitate the comparison of taxonomic concepts. In this poster, we report an initial set of results from our first case study using the plant genus Rubus. The case study tested the entire pipeline of the Explorer of Taxon Concepts toolkit we have developed and revealed challenging phenomena to be solved in the near future.

Keywords: text mining, taxon concepts, taxonomic descriptions, phenotypic characters, Explorer of Taxon Concepts, Rubus, case study

1 Introduction

For centuries, the fundamental ability to study organisms and their relationships has relied on the use of scientific names. However, name resolution, i.e., sorting out the relationships among various valid and invalid scientific names/synonyms and linking current valid names to appropriate organisms, remains challenging. This is because revisions to the existing taxonomy are happening constantly in various branches, and in general, these revisions are mostly only published in the literature and stored in the brain of true experts as opposed to a searchable database. Biologists use checklists (e.g., [1], [2]) and software services (e.g., [3]) to obtain information on valid names, however, the creation and updates of these tools precisely require the names being resolved first by taxonomists with expertise in the related branches.

Besides taxonomists, the number of which have been steadily declining in the past twenty years [4], another source of taxonomic knowledge is the literature. Our development of the ETC toolkit (Explorer of Taxon Concepts) aims to bring this important source of knowledge to taxonomists using natural language processing and machine-learning methods to facilitate the daunting task of global name resolution.

A taxon concept is one expert's account of a taxon and its relationships with other taxa. Such an account includes a description of organism morphological traits (also called "characters") such as *leaves rounded*. Although genomic evidence has been used, morphological characters remain important in taxonomic studies to identify various taxa. The relationships between two taxon concepts can be one of the five and

each is represented with a symbol: congruent(==), broader than (>), narrower than (<), overlap (><), or disjoint(!).

In this poster, we report a small-scale case study using the ETC toolkit. This case study involves the plant genus Rubus, on which one of the authors is a taxonomic expert. The goal of the case study is to evaluate the usefulness of morphological characters extracted from three Rubus concepts for taxon concept comparison purpose.

2 Data and Method

Three Rubus concepts included in this study are Gleason and Cronquist 1991 [6], abbreviated as C in the text below; Flora of North America V. 9 [7], or FNA; and Weakley 2012 [8], or W. Morphological descriptions of these concepts were extracted.

Table 1. Weakley asserted relationships between the taxa extracted from three sources

No.	Taxon	Relationship	Taxon
1	W: Rubus allegheniensis	==	C: Rubus.allegheniensis
4	W: Rubus.flagellaris	>	C: Rubus.recurvicaulis
12	W: Rubus.pascuus	==	FNA: Rubus.pascuus

Weakley, a renowned plant taxonomist, established relationships among Rubus concepts in [8], where 12 relationships among the three Rubus concepts were extracted. These relationships involve 24 Rubus species. Table 1 lists 3 of the 12 relationships. W:Rubus allegheniensis == C:Rubus allegheniensis says W's Rubus allegheniensis and C's Rubus allegheniensis are the same taxon concept. If two concepts are congruent, we would expect morphological characters of the two to be very similar.

Source text: Flowers bisexual; petals usually white to pale-pink, obovate or elliptic to orbiculate, (8-)10-15 mm.

Extracted characters:

Organ	Character Name	Character Value	Modifier
flower	reproduction	bisexual	
petal	coloration	white to pale-pink	usually
petal	shape	obovate | elliptic to orbiculate	

Fig. 1. An example of a description sentence and the characters extracted

ETC Text Capture, Ontology Building, and Matrix Generation tools [4-5] form a pipeline and were used to extract and standardize morphological characters from the 24 descriptions. Next, ETC Taxonomy Comparison tool was used for character similarity comparison. Fig. 1 shows a descriptive sentence and selected characters that were extracted from the sentence. Vertical bars (|) separate multiple character values.

Taking a character as a set of values that include the organ, character name, individual character values, and individual modifiers (Fig. 1), the Taxonomy Comparison tool computes the similarity score between two characters using the Jaccard Index of two sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

A non-zero similarity score is assigned only to the "comparable" characters that have common organ and character names, for example, J(*leaf length 3cm*, *stem length 3cm*) = 0. Fig. 2 shows the graphic interface of the tool visualizing the character similarity scores and the similarity of two taxa based on their characters.

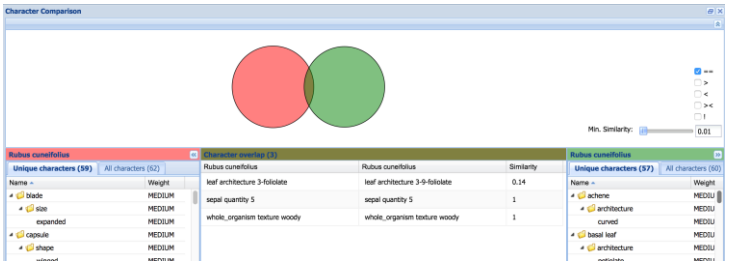

Fig. 2: The character comparison interface in the ETC Taxonomy Comparison tool

Table 2. Challenging characters and similarity scores provided by experts

Character 1	**Character 2**	**Score1**	**Score2**
Comparison of number with expression			
carpel quantity 5-150	carpel quantity 5-many | many	1	1
carpel quantity numerous	carpel quantity 5-many	0.5	1
carpel quantity 2-150	carpel quantity 5-many | many	1	1
leaf width 1dm - 2dm	leaf width >1dm	1	1
leaf width 2dm - 7dm	leaf width 1dm -3 dm	1	1
Hyphen used in character values			
leaf size small	leaf size small – large	0.5	1
flower coloration white - pink | red	flower coloration white | pink | purplish	1	1
hypanthium shape flat - hemispheric	hypanthium shape flattish | hemispheric	1	1
primocane orientation trailing - low-arched	primocane orientation prostrate | creeping | low-arching	1	1
Use (non-use) of modifiers			
whole_organism texture fibrous | more or less woody	whole_organism texture woody	1	1
leaflet margin shape *moderately to coarsely* doubly serrate | *rarely singly* serrate	leaflet margin shape serrate | doubly serrate	1	1
whole_organism texture woody	whole_organism texture woody *at base*	1	1
whole_organism texture woody *at top*	whole_organism texture woody *at base*	1	1
whole_organism texture *rarely* woody	whole_organism texture *usually* woody	1	1
Different levels of details			
leaf architecture long-petiolate| 3-foliate |stipulate	leaf architecture 3-9-foliolate	0.5	1
leaf architecture palmately compound | rarely ternate	leaf architecture 3-9-foliolate | compound	0.2	1

3 Results and Discussion

A total of 1553 characters, include 98 “comparable” pairs, were extracted from the 24 taxonomic descriptions by the ETC pipeline mentioned above. 39 of the 98 pairs of characters were assigned a non-zero score by the ETC Taxonomy Comparison tool, and 59 pairs were assigned a zero score incorrectly. In addition, some of the 39 similarity scores were also deemed too low by biologists.

A manual analysis of these issues revealed some significant challenges associated with characters that were expressed differently but overlapping in meaning (Table 5). Two biologist co-authors were asked to manually score these characters and their scores are listed in Table 5 as Scores 1 and 2. Both biologists demonstrated a high level of tolerance and consistently assigned scores much higher than the scores that would be produced by Jaccard (note: J scores may not be produced for some cases, e.g., “5-many” doesn’t have a upper bound so number of values for the character cannot be found).

4 Conclusion

The Rubus case study shows that the tool is capable of extracting trait evidence from source taxonomic descriptions. All 98 character pairs were extracted. This case study also revealed categories of challenging characters, and more importantly, differences between the scoring mechanisms used by domain experts and by the machine. We will further investigate experts scoring mechanisms and implement several options useful for taxon concept comparison.

Acknowledgment National Science Foundation Grant No. DBI-1147266.

References

1. Roskov, Y., Kunze, T., Paglinawan, L., Orrell, T., Nicolson, D., Culham, A. & Hernandez, F. (2013). Species 2000 & ITIS Catalogue of Life, 2013 Annual Checklist.
2. Pyle, R. L. (2016). Towards a Global Names Architecture: The future of indexing scientific names. ZooKeys, (550), 261.
3. Drew, L. W. (2011). Are We Losing the Science of Taxonomy? As need grows, numbers and training are failing to keep up. BioScience, 61(12), 942-946.
4. Cui, H. (2012). CharaParser for fine-grained semantic annotation of organism morphological descriptions. Journal of the Association for Information Science and Technology, 63(4), 738-754.
5. Cui, H., Xu, D., Chong, S. S., Ramirez, M., Rodenhausen, T., Macklin, J. A., ... & Koch, N. M. (2016). Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. BMC bioinformatics, 17(1), 471-480.
6. Gleason, H. A & Cronquist, A. 1991. Manual of Vascular Plants of Northeastern United States and Adjacent Canada, New York City: New York Botanical Garden Press.
7. Lawrence A.A, Goldman, D.H., Macklin, J.A. & Moore, G. 1997. Rubus. In: Flora of North America Editorial Committee, eds. 1993+. Flora of North America North of Mexico. 20+ vols. New York and Oxford. Vol. 9.
8. Weakley, A. 2015. Flora of the Southern and Mid-Atlantic States http://www.herbarium.unc.edu/flora.htm