Full of beans: a study on the alignment of two flowering plants classification systems

Yi-Yun Cheng and Bertram Ludäscher

School of Information Sciences, University of Illinois at Urbana-Champaign, USA
{yiyunyc2,ludaesch}@illinois.edu

Abstract. Advancements in technologies such as DNA analysis have given rise to new ways in organizing organisms in biodiversity classification systems. In this paper, we examine the feasibility of aligning two classification systems for flowering plants using a logic-based, Region Connection Calculus (RCC-5) approach. The older "Cronquist system" (1981) classifies plants using their morphological features, while the more recent Angiosperm Phylogeny Group IV (APG IV) (2016) system classifies based on many new methods including genome-level analysis. In our approach, we align pairwise concepts X and Y from two taxonomies using five basic set relations: congruence (X=Y), inclusion (X>Y), inverse inclusion (X<Y), overlap (X><Y), and disjointness (X!Y). With some of the RCC-5 relationships among the *Fabaceae* family (beans family) and the *Sapindaceae* family (maple family) uncertain, we anticipate that the merging of the two classification systems will lead to numerous merged solutions, so-called *possible worlds*. Our research demonstrates how logic-based alignment with ambiguities can lead to multiple merged solutions, which would not have been feasible when aligning taxonomies, classifications, or other knowledge organization systems (KOS) *manually*. We believe that this work can introduce a novel approach for aligning KOS, where merged possible worlds can serve as a *minimum viable product* for engaging domain experts in the loop.

Keywords: taxonomy alignment, KOS alignment, interoperability

1 Introduction

With the advent of large-scale technologies and datasets, it has become increasingly difficult to organize information using a stable unitary classification scheme over time. An ideal work classification system, as noted in [1], should be neither too fine-grained, nor too esoteric, to stand the test of time. However, a real-life knowledge organization system (KOS) oftentimes has trade-offs in its stability and granularity, especially when new developments have been made technologically. For example, the use of DNA analysis has provided new data signals that in turn have changed the way biologists classify organisms—traditionally, they may classify organisms based on similarities of surface-level features, while classifying based on similarities in micro-level DNA analysis has become prevalent now. Therefore, the interoperability among different KOSs over time

addressing the same topic has become more and more important in this current era with rapid developments and innovations.

In the biodiversity communities, taxonomies, a type of KOS under the classification schemes [2], has always been one of the main focuses of research; especially in the field of systematics. As such, similar arguments about maintaining a 'unitary classification' over time were made in [3]. The authors [3] stated that it is common for taxonomists to contradict each other's or even their own previous taxonomies. To this end, rather than having a permanent anchor for a specific KOS (taxonomy or classification scheme), a better approach, perhaps, is to embrace the fact that KOS are dynamic, time-specific, and responsive to both empirical signals and human classification interests. Thus, a more principled solution in dealing with the interoperability issues among KOS is perhaps to compare and align the classifications and at the same time presenting the disagreements among them [3][4].

In this paper, we propose the use of a logic-based approach to compare and reconcile two major classification systems in the field of plant systematics, with the aim of demonstrating the feasibility of integrating two classifications that result in numerous possible solutions. Further, we demonstrate the computational power that can aid us in aligning KOS which could not have been possible when working with alignments manually. Specifically, we consider a use case of the flowering plants classification systems by (1) Arthur Cronquist (1981) [5], and (2) Angiosperm Phylogeny Group IV [6] in order to map plant families (concepts) mentioned in both classifications with one of the five Region Connection Calculus relations (RCC-5) using an open source, logic-based tool named Euler/X. The flowering plants we have included in this study are not only some of the most common families that we see in our everyday lives, such as the sunflowers (Asteraceae sec. APG IV[1]), but also those flowering plants by a biologist's definition such as the beans family (Fabaceae sec. APG IV). We hope that this work will further shed light on the possible alignments of the classifications in the information science community and bring a novel approach for aligning KOS in the future.

2 Two Flowering Plant Classifications

In line with the traditions of Bentham and Hooker, Takhtajan, and Bessy, the ***Cronquist system*** [5] is an approach in classifying and identifying flowering plants based on *phyletics* (classifying resemblances based on evolution and *morphological* similarity - similar characters - of the plants). The Cronquist system divides the whole flowering plant world into two phyla, Magnoliopsida and Liliopsida, with approximately 300 families included in the former, and 60 families in the latter [5]; this system is said to be the most "fully developed phyletic system" of flowering plant classification systems[2] by far [7][8].

However, rapid breakthroughs in DNA studies and technologies have given rise to a more recent camp of approaches in classifying plants based on *phylogenetics*. Early

[1] Taxonomic Concept Labels (TCLs): name sec. source

[2] In biodiversity classification systems, the sequence of the concepts is not taken into account.

cladistic analysis or *phylogenetic systematics*, established by Willi Hennig, has put systematics to the task of finding shared, derived character states among any three groups of organisms to find their common ancestors, or *clades* [9]. Modern phylogenetics consists of the use of "both morphological and molecular data and modern methods of data analyses to study evolutionary relationships among organisms" [7]. The ***Angiosperm Phylogeny Group system (APG IV)*** [6], is one classification in this camp that has become the *de facto* standard for the classification of plants of the modern era. The most noticeable differences between the APG IV system and the other plant classification systems are in the higher-level ranks. Instead of using classes or phyla, the APG uses *clades* (e.g. rosid clade, asterid clade), or even other higher-level ranks (e.g. monocots, eudicots) [6]. Though the APG IV system has become the most recent major classification systems, the Cronquist system, established almost 40 years ago, still remains highly influential to this date for its completeness and comprehensiveness, and many legacy papers with key plant data still used Cronquist's classifications.

3 Reasoning about taxonomies and Euler/X

Taxonomies are one type of KOS with hierarchical structures, similar to classification schemes [2]. *Taxonomy alignment* refers to the mapping and reconciliation of two or more taxonomies. In our approach, we employ a logic-based approach called Region Connection Calculus (RCC-5) to align concepts across taxonomies using five possible relationships: congruence (X=Y), inclusion (X>Y), inverse inclusion (X<Y), overlap (X><Y), and disjointness (X!Y). Euler/X (https://github.com/EulerProject/EulerX) is a logic-based reasoning tool for taxonomy alignment based on set constraints, specifically RCC-5 and implemented in answer set programming and direct RCC reasoning.

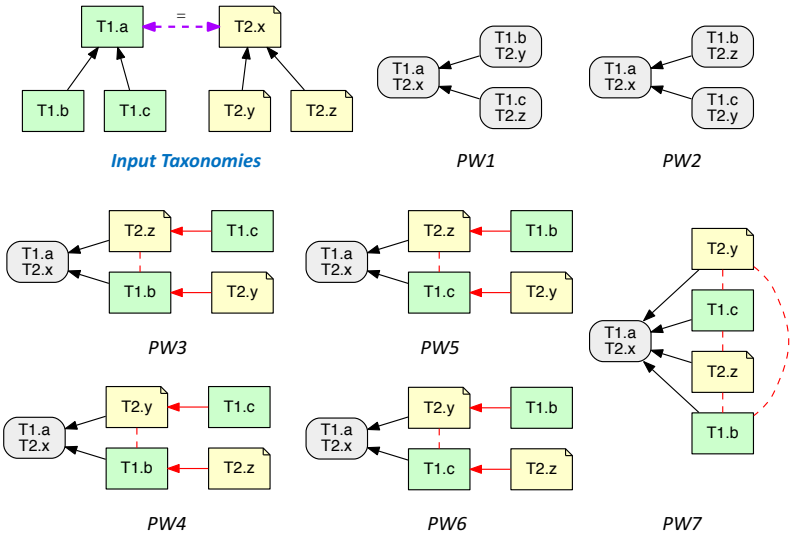

Fig. 1. A simple taxonomy alignment problem with T_1, T_2 and 7 *possible worlds*

Given two taxonomies T_1 and T_2, and a set of articulations (relationships between concept X in T_1 and concept Y in T_2, defined in RCC-5 relations), the *Taxonomy Alignment Problem* (TAP) is to derive a merged taxonomy T_3. A TAP may have zero, one, or many solutions – inconsistent, unique, or ambiguous solutions respectively. In previous work [4], we have addressed the challenges of vocabulary confusion and interoperability between similar but different taxonomies by reconciling the taxonomic disagreement via unique Euler/X solutions, i.e., where there is only one "possible world" that includes all logically inferred inter-taxonomy relationships. In this paper, we further demonstrate features of Euler/X, i.e., its capability to naturally represent ambiguity via multiple possible worlds. A *possible world* is a consistent solution where there are no contradictions on the ways concepts were aligned, and that each concept in the two taxonomies is "sorted out" with exactly one of the five RCC relations. To show a simple TAP example and the multiple possible worlds that exist, consider two taxonomies T_1 and T_2, each with two children (Figure 1). Assume that we only know about these that the highest-level nodes are congruent to each other (T_1.a == T_2.x); how T_1.b, T_1.c relate to T_2.y, T_2.z is unknown. With these underspecified articulations in our TAP, ambiguities arise. Therefore, we end up with seven possible worlds where PW1 and PW2 depict how the concepts can be aligned congruently, PW3 to PW6 depict similar but subtle differences on how T_1.b, T_1.c is included in or includes T_2.y, T_2.z, while PW7 depicts T_1.b, T_1.c, T_2.y, T_2.z all overlapping each other (red dashed lines).

4 Method

It is well-known that the basic taxonomic ranks in biological classification include *kingdom*, *phylum*, *class*, *order*, *family*, *genus*, and *species*. In this research, we are mainly focusing on the alignment of the *family-level* flowering plants. Among the 295,383 flowering plants species [10] within the Magnoliophyta phylum sec. Cronquist, or the Angiosperms sec. APG IV, we are considering only the 40 most common families or subfamilies out of a total of 416 flower families [6]. These families include Magnoliaceae, Ranunculaceae, Papaveraceae, Cataceae, Betulaceae, Fabaceae, Rosaceae, just to name a few.

Both the Cronquist systems and the APG system have these 40 families or some modifications to the names of these families in their classifications. Each family serves as a concept in our alignment. In our first alignment study, if the family in both systems shares the exact same name, we assume (possible incorrectly) that they are congruent to each other. If there are similar but different names, we will leave the concepts unmapped at first. To be more specific, if concept X in the Cronquist system is exactly the same as concept Y in the APG system, we will mark them as [C.X {=} APG.Y]. See **Figure 2** for our initial input taxonomies. The following six articulations are the ones that are uncertain to us because they have different names:

```
[C.Caesalpiniaceae ? APG.Caesalpinioideae]
[C.Mimosaceae ? APG.Mimosoideae]
[C.Fabaceae ? APG.Faboideae]
[C.Aceraceae ? APG.Sapindaceae]
[C.Sapindaceae ? APG.Sapindaceae]
[C.Hippocastanaceae ? APG.Sapindaceae]
```

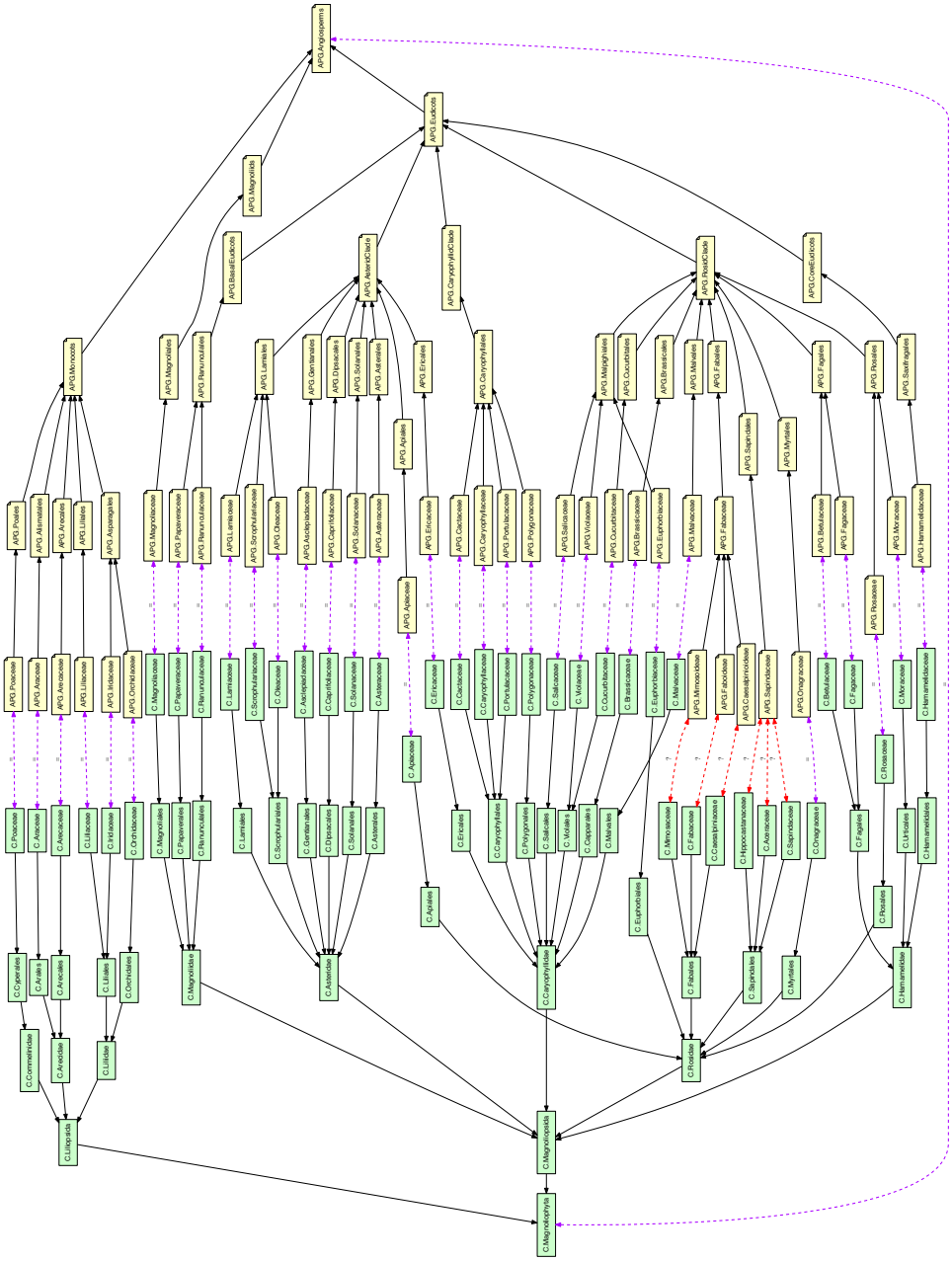

Fig. 2. Initial input taxonomies with six unknown relations, aligning the two classification systems (green=Cronquist; yellow=APG), with the Fabaceae and the Sapinadaceae relationship unknown on both sides (red dotted lines)

Due to the above six uncertain relationships, we expect that the articulations between the concepts of the two classifications in our initial input taxonomies are underspecified and will result in numerous possible worlds. Furthermore, the bold assumption to mark names that are spelled the same in both taxonomies as congruent is often not enough, as noted in [12]. This is the reason we conducted a second round of alignment to seek out consultation from a domain expert to verify our 'congruent' alignments as well as to sort out the underspecified articulations. The possible worlds we have produced in the first stage thus became the *minimum viable product* for us to communicate with the expert and let him/her grasp all possible solutions for the alignment problem. The domain expert in our research asserted that the congruent alignments are indeed correct and that the modified articulations for the Fabaceae and the Sapindaceae families are as follows:

```
[C.Caesalpiniaceae {=} APG.Caesalpinioideae]
[C.Mimosaceae {=} APG.Mimosoideae]
[C.Fabaceae {=} APG.Faboideae]
[C.Aceraceae {<} APG.Sapindaceae]
[C.Sapindaceae {<} APG.Sapindaceae]
[C.Hippocastanaceae {<} APG.Sapindaceae]
```

5 Results

The first round of alignment between the two classification systems resulted in 555 possible worlds—meaning that we have 555 different ways for reconciling these two classifications (see **Figure 3** for a few examples). The alignment resulted in so many solutions because the articulations for the Fabaceae family (the beans family) and the Sapindaceae family (the maple family) were left ambiguous and underspecified.

In the Cronquist system, the three families placed under the order Fabales are:

```
Caesalpiniaceae sec. Cronquist
Mimosaceae sec. Cronquist
Fabaceae sec. Cronquist
```

While in the APG IV system, the order Fabales only consists of one family Fabaceae, and under the Fabaceae there are three sub-families:

```
Caesalpinioideae sec. APG IV
Mimosoideae sec. APG IV
Faboideae sec. APG IV
```

First, the family/subfamily names between the two systems are not exactly the same, the spelling is different in the suffix (-ceae vs. -deae); furthermore, the 'umbrella' family Fabaceae in the APG system is spelled the same as one of the three bean families in the Cronquist system (also Fabaceae). Similarly, we raise doubts on whether the Sapindaceae in the Cronquist system is entirely equivalent to the Sapindaceae sec. APG IV; though they share the exact same name, the Aceraceae sec. Cronquist, and Hippocastanaceae sec. Cronquist, also in the order of Sapindales, were totally void in the APG system. Therefore, in our initial attempt we could not firmly state what the articulations among these families were, echoing the claim that "names are not enough" in [11].

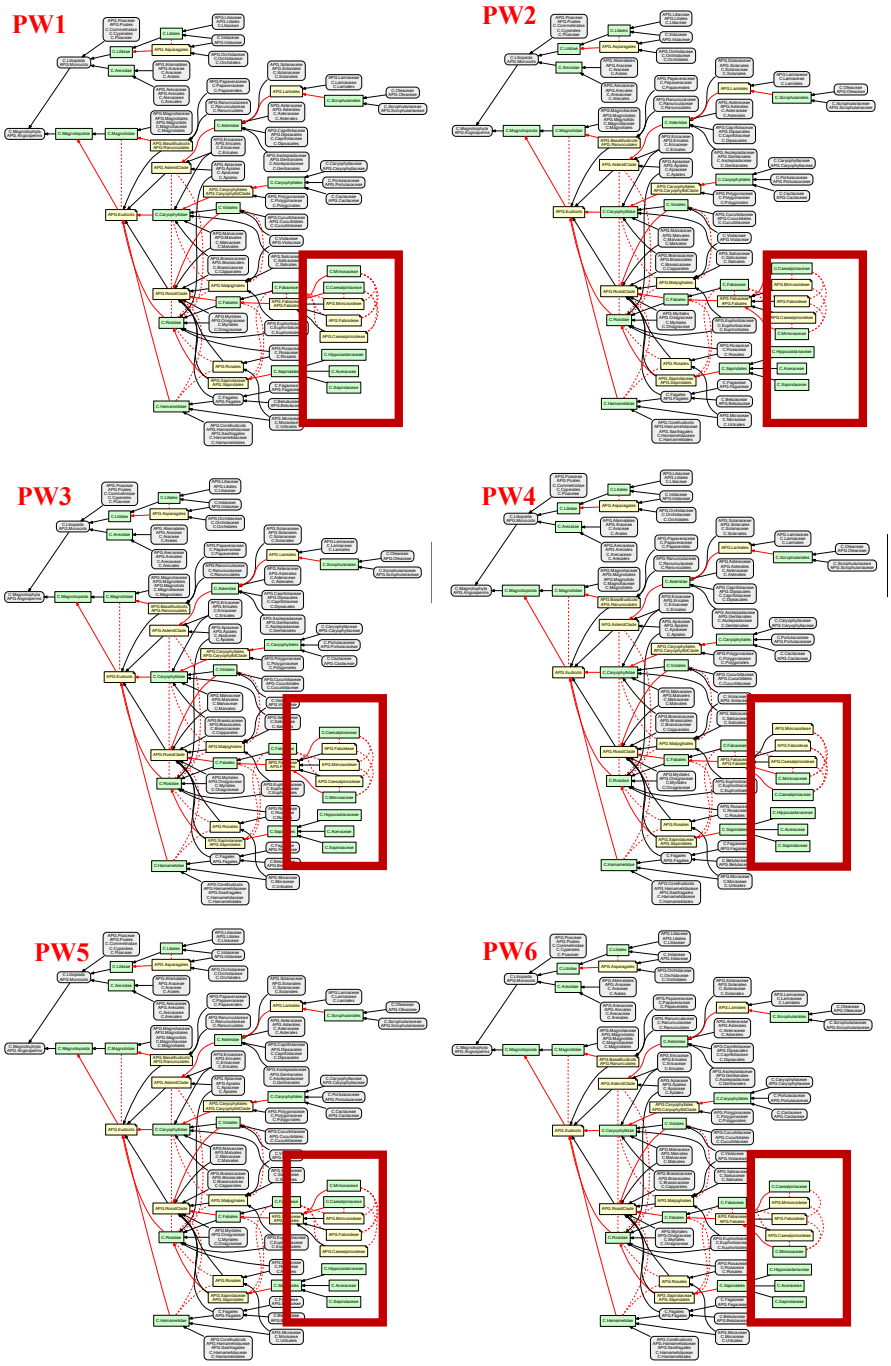

Fig. 3. Six of the 555 possible worlds to model the merged view of the two classification, with the main differences marked in the red boxes.

After consulting the domain expert on our second round of alignment, we were able to refine our alignments between the two taxonomies and make the relationships among the beans families explicit. Despite the slight differences in the suffix, the three families under Cronquist indeed were congruent to the three sub-families within the APG system (**Figure 4**), and that the Sapindaceae sec. Cronquist, Aceraceae sec. Cronquist, and Hippocastanaceae sec. Cronquist are all combined as one as Sapindaceae sec. APG IV now.

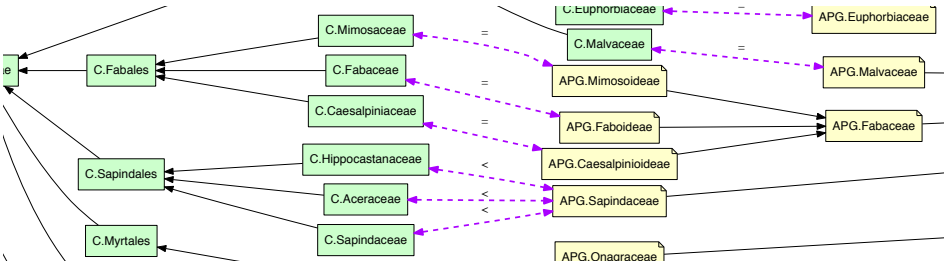

Fig. 4. Refinement of the input alignment to align the Fabaceae and the Sapinadeae family

Given this refinement in the relationship among the beans and the maple families, instead of having 555 possible worlds, we were able to reduce the ambiguities and thus the number of possible worlds to only *one* unique possible world (**Figure 5**), where we can see the congruence in families, the inferred relationship among the higher-level nodes, and the two original classifications. All the results can be retrieved from our Github repository (https://github.com/yiyunyc2/NKOS18).

6 Conclusion

New developments, whether in technologies or in paradigms, always challenge the views of past major knowledge organization systems. In the biodiversity domain, molecular data when examined under careful sampling, can provide valuable pointers to the classification of plants. However, to quote [12]: "... molecular characters are subject to evolutionary convergence, parallelism, and reversal; therefore, molecular methods are not a panacea. Molecular evidence should be used with, not in place of, morphological evidence." Though the APG system has shown more substantial importance in recent years due to the advancement in micro-level analysis of the molecular data, the Cronquist system still maintains its esteemed role for its comprehensiveness and preciseness in morphologically classifying the flowering plants.

This paper serves as an exploratory research on the comparison and alignment of two KOS, specifically classification schemes. Our approach suggests that classifications can coexist with each other while disambiguating the names among concepts in a merged possible world. Our approach also demonstrates the capability of computationally solving complex logic-based alignments for cases where, e.g., due to underspecified relations in the input KOS alignments, manual efforts would likely fail to yield all 555 different ways to merge and reconcile the two KOS.

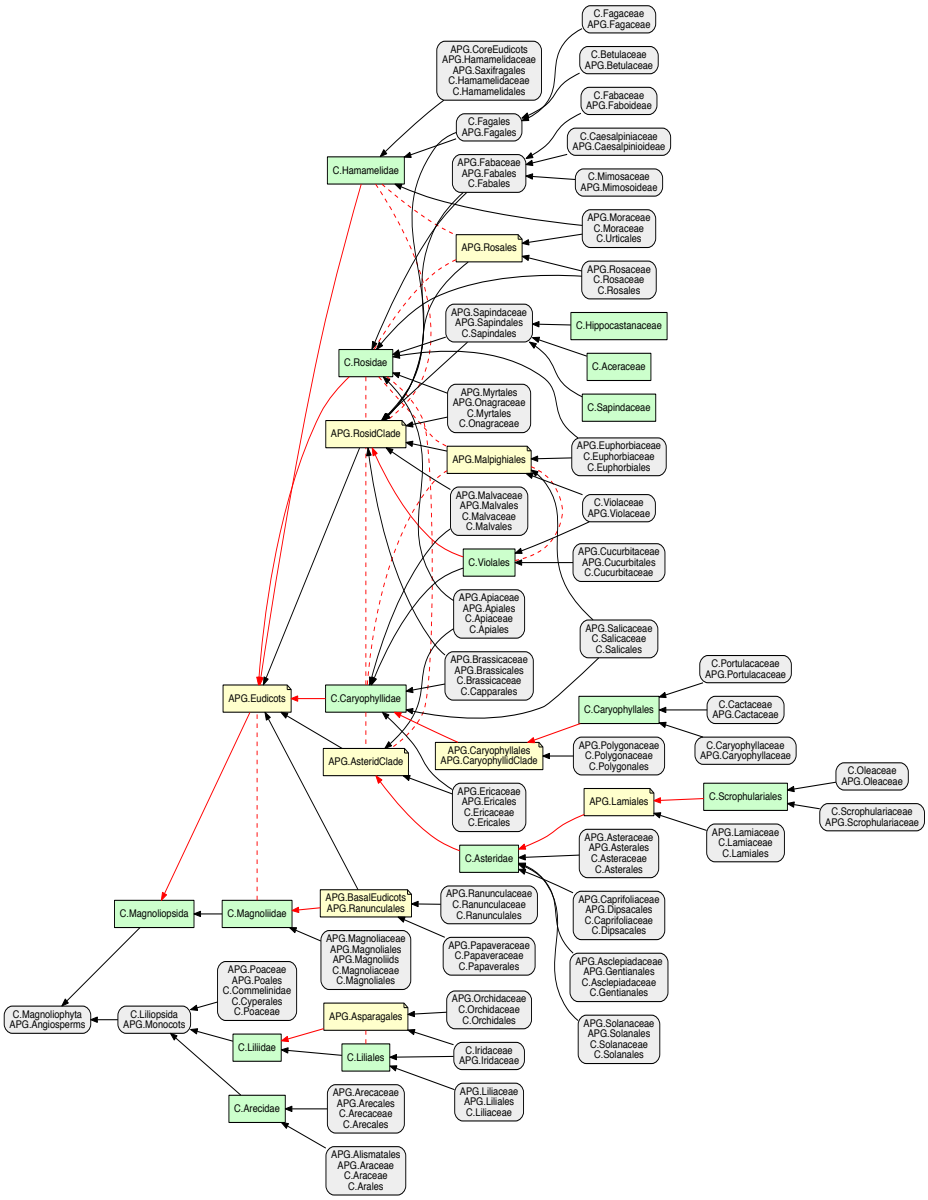

Fig. 5. The final one possible world merged solution for reconciling the two flowering plants classification systems. (green rectangular box=Cronquist; yellow "note" box=APG; grey round box=congruent; red concrete line=inferred parent-child relationship by Euler/X; red dotted line=inferred overlapping relationship from Euler/X)

Somewhat ironically, the limitation in this study also lies in the strong inferential power—in which when a parent node only has one child, the RCC reasoner in Euler/X will collapse the concepts and merge them as the same node. For example, Euler/X derived that the family Ericaceae and the order Ericales are exactly the same and merged Cronquist.Ericaceae, APG.Ericaceae, Cronquist.Ericales, and APG.Ericales as congruent. However, we argue that this limitation is due to the fact that we have only chosen some 40 major flower families instead of all 416 families. We could add missing children or artificial children here, or include the full 416 angiosperms families, but this is beyond the scope of our study. For the purposes of demonstrating the logic-based taxonomy alignment approach, our smaller use case is sufficient and more tractable .

It is also worth noting that domain expert opinions are still needed to differentiate and lead us to the single solution we are looking for. We foresee our logic-based approach for aligning KOS as an essential preliminary processing steps and a *minimum viable product* before bringing the KOS interoperability alignment problems to the domain experts. KOS alignment problems are usually complicated with a slow learning curve; domain experts, in our case, plant systematists, may not fully comprehend at first the reason we need to align different KOS. If we approach them directly with two classifications and ask them one by one what relationships between each concept are, they will probably feel befuddled by the situation. Once we have the Euler/X-generated *possible worlds* in the first round of alignments, whether a few, tens, hundreds, or even thousands of PWs, the alignment problem will become concrete to the domain experts and consulting them for validation of the articulations would be much easier.

This study, we also believe, has further implications on making our classification systems more "full of beans" (here we take on the positive connotation, meaning *full of energy*), meaning that it may open doors to enable semantic interoperability, and enrich diversity in classification systems when we work with KOS alignments using the logic-based RCC-5 approach. We believe that in the future we can implement this approach for semantic interoperability issues among classifications in the information science community, or even other higher-level KOS such as ontologies.

7 Acknowledgement

The first author wishes to thank Dr. Stephen R. Downie for the wonderful introduction to plant systematics. This research is the outcome of the course project of IB335 and the Independent Study taught by Dr. Downie. The authors also thank Dr. Nico M. Franz and Ms. Ly Dinh for support and kind feedback on this research.

References

1. Bowker, G. C., & Star, S. L. (2000). Sorting things out: Classification and its consequences. MIT press.
2. Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Digital Library Federation, Council on Library and Information Resources, 1755 Massachusetts Ave., NW, Suite 500, Washington, DC 20036.

3. Franz, N. M., Peet, R. K., & Weakley, A. S. (2008). On the use of taxonomic concepts in support of biodiversity research and taxonomy. Systematics Association Special Volume, 76, 63.
4. Cheng, Y.-Y., Franz, N., Scheider, J., Yu, S., Rodenhasen, T., Ludäscher, B. (2017). Agreeing to disagree: reconciling conficting taxonomic views using a logic-based approach. In *Proceedings of the ASIS&T 2017 Annual Meeting*, Washington, D.C., USA, October 27-November 1st. IDEALS: http://hdl.handle.net/2142/97907
5. Cronquist, A. (1981). An integrated system of classification of flowering plants. Columbia University Press.
6. Byng, J. W., Chase, M. W., Christenhusz, M. J., Fay, M. F., Judd, W. S., Mabberley, D. J., ... & Briggs, B. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society, 181(1), 1-20.
7. Downie, S. R. (2018). Lecture on Historical Systematics. Personal Collection of S. R Downie, University of Illinois at Urbana-Champaign, Champaign IL.
8. Judd, W. S., Campbell, C. S., Kellogg, E. A., & Stevens, P. F. (2016). Plant systematics. A phylogenetic approach. Sinauer Associates, Sunderland, Mass., USA, 464, 3-4. 3rd edition
9. Vane-Wright, R.I. 2013. Taxonomy, Methods of. In Levin S.A. (ed.), Encyclopedia of Biodiversity (2nd edn) 7: 97–111. Waltham, MA: Academic Press.
10. Christenhusz, M. J., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. Phytotaxa, 261(3), 201-217.
11. Franz, N. M., Chen, M., Kianmajd, P., Yu, S., Bowers, S., Weakley, A. S., & Ludäscher, B. (2016). Names are not good enough: Reasoning over taxonomic change in the Andropogon complex 1. *Semantic Web*, *7*(6), 645-667.
12. Takhtadzhian, A. L. (1997). Diversity and classification of flowering plants. Columbia University Press.