Research Ideas and Outcomes 2: e10610
doi: 10.3897/rio.2.e10610

NSF Grant Proposal

Controlling the taxonomic variable: Taxonomic concept resolution for a southeastern United States herbarium portal

Nico Franz[‡], Edward Gilbert[§], Bertram Ludäscher[|], Alan Weakley[¶]

‡ Arizona State University, Tempe, United States of America
§ Arizona State University, Symbiota, Tempe, United States of America
| University of Illinois Urbana-Champaign, Champaign, United States of America
¶ The University of North Carolina at Chapel Hill, Chapel Hill, United States of America

Corresponding author: Nico Franz (nico.franz@asu.edu)

Reviewable v1

Received: 22 Sep 2016 | Published: 30 Sep 2016

Citation: Franz N, Gilbert E, Ludäscher B, Weakley A (2016) Controlling the taxonomic variable: Taxonomic concept resolution for a southeastern United States herbarium portal. Research Ideas and Outcomes 2: e10610. doi: 10.3897/rio.2.e10610

Executive summary

Overview. Taxonomic names are imperfect identifiers of specific and sometimes conflicting taxonomic perspectives in aggregated biodiversity data environments. The inherent ambiguities of names can be mitigated using syntactic and semantic conventions developed under the taxonomic concept approach. These include: (1) representation of taxonomic concept labels (TCLs: name sec. source) to precisely identify name usages and meanings, (2) use of parent/child relationships to assemble separate taxonomic perspectives, and (3) expert provision of Region Connection Calculus articulations (RCC–5: congruence, [inverse] inclusion, overlap, exclusion) that specify how data identified to different-sourced TCLs can be integrated. Application of these conventions greatly increases trust in biodiversity data networks, most of which promote unitary taxonomic 'syntheses' that obscure the actual diversity of expert-held views. Better design solutions allow users to control the taxonomic variable and thereby assess the robustness of their biological inferences under different perspectives. A unique constellation of prior efforts – including the powerful Symbiota collections software platform, the Euler/X multi-taxonomy alignment toolkit, and the "Weakley Flora" which entails 7,000 concepts and more than

75,000 RCC–5 articulations – provides the opportunity to build a first full-scale concept resolution service for SERNEC, the SouthEast Regional Network of Expertise and Collections, currently with 60 member herbaria and 2 million occurrence records.

Intellectual merit. We have developed a multi-dimensional, step-wise plan to transition SERNEC's data culture from name- to concept-based practices. (1) We will engage SERNEC experts through annual, regional workshops and follow-up interactions that will foster buy-in and ultimately the completion of 12 community-identified use cases. (2). We will leverage RCC–5 data from the Weakley Flora and further development of the Euler/X logic reasoning toolkit to provide comprehensive genus- to variety-level concept alignments for at least 10 major flora treatments with highest relevance to SERNEC. The visualizations and estimated > 1 billion inferred concept-to-concept relations will effectively drive specimen data integration in the transformed portal. (3) We will expand Symbiota's taxonomy and occurrence schemas and related user interfaces to support the new concept data, including novel batch and map-based specimen determination modules, with easy output options in Darwin Core Archive format. (4) Through combinations of the new technology, enlisted taxonomic expertise, and SERNEC's large image resources, we will upgrade minimally 80% of all SERNEC specimen identifications from names to the narrowest suitable TCLs, or add "uncertainty" flags to specimens needing further study. (5) We will utilize the novel tools and data to demonstrate how controlling for the taxonomic variable in 12 use cases variously drives the outcomes of evolutionary, ecological, and conservation-based research hypotheses.

Broader impacts. Our project is focused on just one herbarium network, but the potential impact is as wide as Darwin Core or even comparative biology. We believe that trust in networked biodiversity data depends on open and dynamic system designs, allowing expert access and resolution of multiple conflicting views that reflect the complex realities of ongoing taxonomic research. Taking well over 1 million SERNEC records from name- to TCL-resolution will show that "big" specimen data can pass the credibility threshold needed to validate the substantive data mobilization investment. We will mentor one postdoctoral researcher (UNC), two Ph.D. students (ASU, UIUC), and at least 15 undergraduate students (ASU). Each of our workshops will capacitate 10-15 SERNEC experts, who in turn can recruit colleagues and students at their home collections. We will incorporate the project theme and use cases into undergraduate courses taught at six institutions and reaching an estimated 300-500 students annually (10-40% minority students). At each institution, project members will make a systematic effort to recruit new students from underrepresented groups. Our group's leadership of Symbiota (with close ties to iDigBio), SERNEC, and local biodiversity projects and centers will further promote the new data culture. We will create a feature story "Where do plant species occur?" for ASU's popular "Ask A Biologist" website, and a series of undergraduate student-led "How-To" videos that illustrate the use case workflows, including the creation of multi-taxonomy alignments.

Keywords

Aggregation, concept taxonomy, conflict, flora, herbarium, logic, reasoning, Region Connection Calculus, specimens, synthesis

List of participants

Mac Alford, Mark Fishbein, Alan Franck, Nico Franz, Edward Gilbert, Michael Lee, Zack Murrell, Bertram Ludäscher, Pamela Soltis, Alan Weakley.

Data management plan

Types of Data Produced

Data to be produced and managed for the project include: (1a) Software code written for the Symbiota content management system (primarily written in PHP and with heavy use of JavaScript libraries; and connecting to the open source MariaDB SQL database platform) and (1b) for the Euler/X logic reasoning toolkit (primarily written in Python); (2) specimen occurrence records (with new identifications) managed in the Symbiota-operated SERNEC herbarium portal, and formatted in compliance (where possible; see details below) with the Taxonomic Working Group (TDWG) -endorsed Darwin Core (DwC) and Taxonomic Concept Transfer Schema (TCS) standards (https://github.com/tdwg); and (3) Euler/X toolkit input/output files, presently stored in simple .csv, .gv (GraphViz), .pdf, .txt, and .yaml file formats. We will also (4) author web posts (.html) and instructional videos (.mp4) (see Broader Impacts).

Data and Metadata Standards

The Symbiota-based SERNEC portal occurrence data are fully Darwin Core-compatible. These data can be bundled through easy-to-use platform functions to yield Darwin Core Archive files for wider sharing. We note, however, that Darwin Core does not presently support all syntactic and semantic conventions of the taxonomic concept approach. In particular, a modularized and flexible management of taxonomic concept labels (TCLs) in conjunction with parent/child relationships and RCC–5 articulations – in some instances under multiple extensional or intensional readings (Section 8.II.1) – is out of scope for DwC. Certain aspects are covered by the TCS. However, this 2005-ratified standard needs revision and expansion, particularly in connection with a fully functional specimen data environment such as Symbiota.

We will adhere to DwC and TCS as much as is conducive to our representation needs. At the same time, this part of the project (Section 8.3.I: taxonomy/occurrence module expansion) is properly viewed as new work required for updating and expanding the TCS ("2.0"). Other services (e.g., GBIF, iDigBio) that 'just' manage DwC syntax and semantics, while not incompatible with our data, will nevertheless be unable to replicate our TCL-based specimen resolution services that critically require RCC–5 integration signals. As a stop-gap solution, we will provide links to alignments on GitHub and/or in DataOne in the "dynamicProperties" field.

At present Euler/X input and output data formats, including the input constraint .txt files and resulting .csv MIR files, are not covered by ratified standards (TDWG or other entities). However, both are ASCII-based, largely translatable into TCS terms and relationships, and easily manageable through standard control version systems (such as Git) that can automatically visualize version differences. The scale of this project – 2,000-3,000 alignments – presents an opportunity to create more formalized input and output data standards. The UIUC team will develop a simple alignment archive format (.aarc). We will also generate an associated and self-contained viewer tool to make taxonomy alignment products (i.e., input, output, and inference rules used to logically connect these products) transparent and reproducible.

Policies for Access and Sharing

Our project operates fully in the Public Domain. The Symbiota software code is published under the GNU General Public License (Version 3, June 2007), whereas the Euler/X code is published with the BSD license (also used by the Open Tree of Life project). All Symbiota-/SERNEC-held data and the new Euler/X alignments are published under the CC0 license (or similar, given certain collections records and image artefacts; see https://www.idigbio.org/content/idigbio-intellectual-property-policy; http://choosealicense.com/licenses/). UIUC's Ludäscher is a member of the DataONE Leadership Team and will work with colleagues in the DataONE Semantics and Provenance Working Group to explore sharing taxonomically (TCL) annotated datasets through DataONE.

Policies for Re-use and Distribution

Collection- and use case-based data will be published as Darwin Core Archive files. To disseminate DwC–A packages, we will use well-established and separate publication pathways from Symbiota to GBIF (http://www.gbif.org/dataset/) and iDigBio (https://www.idigbio.org/portal/publishers), as preferred by these aggregators. The transformed SERNEC portal will also publish our datasets, as DwC-A files and additionally using the expanded schema (syntax, semantics) for multi-TCL-to-specimen resolution that we will generate. This ensures that our use case results remain accessible and reproducible. Specific data packages authored in relation to the use case publications will be disseminated via means sanctioned by open access (option) journals, using repositories such as Dryad (http://datadryad.org/), figshare (https://figshare.com/), and Zenodo (http://zenodo.org/).

Plans for Archiving and Preservation

New software code will be published as releases through GitHub or similar openly accessible source code repositories (e.g., http://gitlab.com). SERNEC portal and use case data will be archived through redundant back-ups at ASU, in addition to GBIF and iDigBio. Data persistence will be further assured by establishing a new archival service relationship with DataONE, facilitated by Ludäscher, and specifically through addition of our Project data to the DataONE member node Knowledge Network for Biocomplexity (KNB) Data Respository (https://search.dataone.org/#profile/KNB).

Roles and Responsibilities

ASU (Franz, Gilbert) assume primary responsibility for project-based managing of data for Symbiota, SERNEC, and the Euler/X alignment repository on GitHub (https://github.com/taxonomic-concept-alignments). All Symbiota code (https://github.com/Symbiota) and contingent software for portal operation are open source. For select code testing purposes, ASU maintains an experimental portal on an institutionally supported VM server (http://hasbrouck.asu.edu/sandbox/). However, all actual SERNEC data are hosted only and directly by the NSF-supported iDigBio infrastructure, which has dedicated Symbiota data servers for multiple hosted data portals. We commit to iDigBio's rules for collaboration, particularly with regards to creating and resolving globally unique specimen identifiers; see https://www.idigbio.org/content/collaborating-idigbio-grant-proposals. UIUC (Ludäscher) is responsible for maintaining the new Euler/X code on GitHub (https://github.com/EulerProject/).

Project description

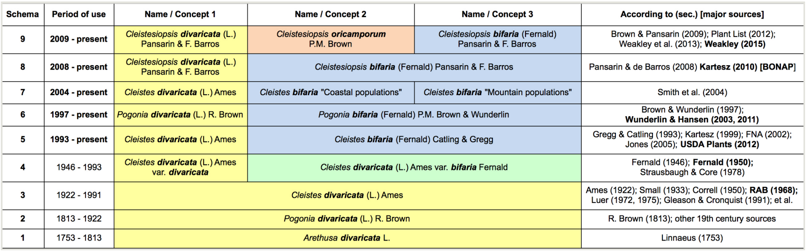
This ABI Development proposal is concerned with building a culture that increases trust in aggregated biodiversity data. We show that the meanings of taxonomic names are a variable in this context that needs to be explicitly modeled and controlled for. We will build a novel, multi-taxonomy conflict resolution service into a herbarium portal, as a pioneering effort that can be applied and propagated more widely.

1. Taxonomic names are (ambiguous) taxonomic concept lineage identifiers

To motivate a complex theme – names, taxa, and concepts – we start with a concrete example. The species epithets "*bifaria*" (coined by Fernald 1946a, Fernald 1946b) and "*divaricata*" (Linnaeus 1753) are very good identifiers for some purposes, and poor ones for others (Witteveen 2015, Franz and Sterner 2015, Franz et al. 2016). Each name is rigidly assigned to an individual type specimen (LINN-HL 1059-3; GH00056705), providing "an objective link between the real world of organisms and the world of language (nomenclature)" (Dubois 2005: 381-382). Moreover, in a particular floristic treatment, *bifaria* and *divaricata* function as identifiers for endangered orchid species that occur in the southeastern Unites States (Fernald 1950, Radford et al. 1968, Wunderlin and Hansen

2011, Weakley 2015). Depending on the treatment, these epithets may be combined with one of four generic names (*Arethusa, Cleistes, Cleistesiopsis*, or *Pogonia*), a varying set of nomenclatural synonyms, a diagnosis or description, and other taxonomic and biological information, such as links to additional specimens or DNA data.

But here we should pause. The phrase "identifiers for species" could imply that we have converged on stable and accurate circumscriptions of two orchid species. It could even imply that we had 'gotten them right' since Linnaeus (1758). Hence the two epithets could reliably *stand for* the species (Ogden and Richards 1923, Peirce 1998, Franz and Sterner 2015). This turns out to be false. Indeed, with the creation of the epithet *bifaria* in 1946, a taxonomic subset of what previous authors (e.g., Ames 1922) had placed under *divaricata* was differentiated from that preceding, more inclusive set (Fig. 1).



Schema	Period of use	Name / Concept 1	Name / Concept 2	Name / Concept 3	According to (sec.) [major sources]
9	**2009 - present**	*Cleistesiopsis* ***divaricata*** (L.) Pansarin & F. Barros	*Cleistesiopsis* ***oricamporum*** P.M. Brown	*Cleistesiopsis* ***bifaria*** (Fernald) Pansarin & F. Barros	Brown & Pansarin (2009); Plant List (2012); Weakley et al. (2013); **Weakley (2015)**
8	**2008 - present**	*Cleistesiopsis* ***divaricata*** (L.) Pansarin & F. Barros	*Cleistesiopsis* ***bifaria*** (Fernald) Pansarin & F. Barros		Pansarin & de Barros (2008) **Kartesz (2010) [BONAP**]
7	**2004 - present**	*Cleistes* ***divaricata*** (L.) Ames	*Cleistes* ***bifaria*** "Coastal populations"	*Cleistes* ***bifaria*** "Mountain populations"	Smith et al. (2004)
6	**1997 - present**	*Pogonia* ***divaricata*** (L.) R. Brown	*Pogonia* ***bifaria*** (Fernald) P.M. Brown & Wunderlin		Brown & Wunderlin (1997); **Wunderlin & Hansen (2003, 2011)**
5	**1993 - present**	*Cleistes* ***divaricata*** (L.) Ames	*Cleistes* ***bifaria*** (Fernald) Catling & Gregg		Gregg & Catling (1993); Kartesz (1999); FNA (2002); Jones (2005); **USDA Plants (2012)**
4	1946 - 1993	*Cleistes* ***divaricata*** (L.) Ames var. ***divaricata***	*Cleistes* ***divaricata*** (L.) Ames var. ***bifaria*** Fernald		Fernald (1946); **Fernald (1950);** Strausbaugh & Core (1978)
3	1922 - 1991	*Cleistes* ***divaricata*** (L.) Ames			Ames (1922); Small (1933); Correll (1950); **RAB (1968);** Luer (1972, 1975); Gleason & Cronquist (1991); et al.
2	1813 - 1922	*Pogonia* ***divaricata*** (L.) R. Brown			R. Brown (1813); other 19th century sources
1	1753 - 1813	*Arethusa* ***divaricata*** L.			Linnaeus (1753)

Figure 1.

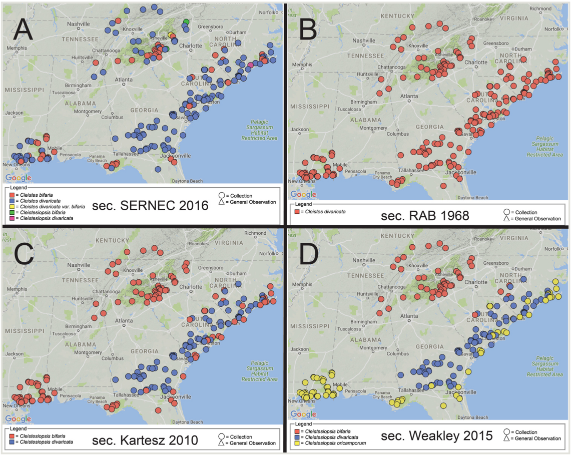
Taxonomic concept labels and concept-to-concept articulations, represented in a tabular alignment of nine schemata, for the *Cleistes* use case (sec. A.S. Weakley). The vertical column position and width of taxonomic concept labels indicates taxonomic non-/congruence. The colors approximate taxonomic name lineages, e.g. blue for *bifaria* and yellow for *divaricata.*

Until 1946, *divaricata* had a wide taxonomic *referent* (= entity for which the name stands), whereas subsequently *divaricata* started to also stand for a narrower referent. Following Pansarin and Brown (2009), *bifaria* received yet another less inclusive meaning (Weakley 2015).

If names are potentially ambiguous, then how should we model the evolving relationships between identifiers, meanings, and natural entities? We propose the following definitions (Franz and Sterner 2015). Genus- and species-level names are anchored by rigidly chosen types – either type names or type specimens (Franz et al. 2008, Witteveen 2015, Witteveen 2016). At any given time, taxonomic names (= *symbols*) also stand for taxa, which we define here as historically, evolutionarily coherent sets of organisms in nature. However, history shows that we need a third element in our model, i.e., taxonomic concepts. We define a taxonomic concept as an empirically informed theory of the identity and definitional boundaries of the *perceived* taxon, expressed by a particular author at a particular time (Berendsohn 1995, Geoffroy and Berendsohn 2003, Franz and Peet 2009). "Name usage" or "taxonomic circumscription/ meaning" (= *reference*; see the semiotic triangle of Ogden and Richards 1923) are common synonyms of "taxonomic concept". We say "perceived taxon" because we represent the interaction between taxonomic concepts – the theories proposed by human authors – and natural taxa as one of successive approximation. Instead of saying "Linnaeus (1758) lumped the species (taxon) *divaricata*", we say "the taxonomic concept *divaricata* sec. (*secundum*, according to) Linnaeus (1758) is more inclusive than the taxonomic concept *divaricata* sec. Pansarin and Brown (2009)". In other words, we are realists about the evolutionary identities of taxa (which are unlikely to have changed much in the past 258 years), but conceptualists about human taxonomic making (which progresses rapidly). Eventually our scientific process is expected to 'get there', within epistemic limits.

It follows that taxonomic names have three roles in our data systems (Franz and Sterner 2015): (1) they identify types, (2) they pragmatically stand for taxa, but with the possibility of imprecision or error, and (3) they identify *lineages* of one to (very) many taxonomic concepts proposed in the history of human taxonomic making.

The third role is critical for querying non-type specimens. Fig. 1 shows that the meanings of *bifaria* and *divaricata* are not stable, instead becoming less inclusive over time. Thus, if a user queries an aggregated data system for specimens identified to *Cleistes bifaria*, the system's immediate response should be: "According to any source, or only according to a particular one? Here are your options [showing five or more schemata] – please specify further." And, depending on the user's query refinement, they would receive distinctly labeled and partially overlapping sets of specimens in return. These sets (if georeferenced) would present different signals on visualized distribution maps (Fig. 2).


A
Legend
🟥 = *Cleistes bifaria*
🟦 = *Cleistes divaricata*
🟨 = *Cleistes divaricata var. bifaria*
🟩 = *Cleistesiopsis bifaria*
🟪 = *Cleistesiopsis divaricata*
sec. SERNEC 2016
○ = Collection
△ = General Observation
B
Legend
🟥 = *Cleistes divaricata*
sec. RAB 1968
○ = Collection
△ = General Observation
C
Legend
🟥 = *Cleistesiopsis bifaria*
🟦 = *Cleistesiopsis divaricata*
sec. Kartesz 2010
○ = Collection
△ = General Observation
D
Legend
🟥 = *Cleistesiopsis bifaria*
🟦 = *Cleistesiopsis divaricata*
🟨 = *Cleistesiopsis oricamporum*
sec. Weakley 2015
○ = Collection
△ = General Observation

Such parent/child relationships are explicit in Figure 2.

"Where do these endangered orchid species occur?" – visualizing the taxonomic variable for aggregated herbarium data. Mappings for the same 250 SERNEC specimens (not all resolved at this geographic scale) according to four distinct taxonomies. **(A)** sec. SERNEC (2016) 'consensus', **(B)** sec. Radford et al. (1968), **(C)** sec. Kartesz (2010), **(D)** sec. Weakley (2015). In A, C, and D, unequal sets of specimens labeled as *bifaria* are red; those labeled as *divaricata* are blue. In B, all specimens are identified as *divaricata*, and hence a query for *bifaria* would not return any specimens.

Likely, one or another specification would lead the user to make distinct biological inferences based on these derivative maps. This is how the user can assess the robustness of their hypotheses vis-à-vis the taxonomic variable.

2. New syntax and semantics for identifying and articulating taxonomic concepts

What we describe is hard to do (Koperski et al. 2000, Boyle et al. 2013, Lepage et al. 2014, Franz et al. 2016a, Franz et al. 2016b). Few if any biodiversity data providers excel at resolving specimen information so granularly. The vast majority are name-based, and not yet concept-based. Creating a "which concept?" query-/counter-query service that reaches to the herbarium specimen level is our overarching objective.

To begin building a solution, we need a new term for the identifier "*Cleistesiopsis divaricata* [name author, year] sec. Pansarin and Brown (2009)". We call these *taxonomic concept labels* (TCLs) (Franz et al. 2015, Jansen and Franz 2015, Franz et al. 2016a). The cardinality relationship between TCLs and taxonomic concepts is one-to-one (Remsen 2016), i.e., they select exactly one taxonomic concept out of a system that may represent many potentially conflicting concepts. We say that the taxonomic name *participates* in the TCL. But the name, by itself, is not suited to pick out a particular taxonomic concept, instead selecting – via nomenclatural (type) identity – the corresponding lineage of concepts (Franz and Sterner 2015). Often the cardinality relationship between taxonomic names and taxonomic concepts is one-to-many, where the "many" will differ in terms of their meanings (Geoffroy and Berendsohn 2003, Kennedy et al. 2005, Remsen 2016). In that sense *bifaria* is a poor identifier: it identifies a lineage of non-congruent meanings. We can also have the inverse situation, where multiple nomenclatural synonyms have a congruent taxonomic reference (Remsen 2016). However, we model these data not as 'one and the same concept', but instead as multiple taxonomic concepts – each with a unique TCL – that happen to have *congruent* meanings (Franz and Peet 2009). So, we would not say "*Cleistes divaricata* var. *bifaria* sec. Fernald (1950) and *Cleistesiopsis bifaria* sec. Kartesz (2010) is the 'same' concept". Instead we manage two TCLs whose meanings are taxonomically congruent. The two *names* involved also happen to have a nomenclatural synonymy relationship, at least according to recent treatments. However, it is not deductively sound to equate synonymy with congruence in our sense, because the former is type-based or may only be "pro parte" (Geoffroy and Berendsohn 2003, Franz et al. 2016a, Franz et al. 2016b). *Arethusa divaricata* is a synonym of *Cleistesiopsis divaricata* (FNA 2015), whereas *Arethusa divaricata* sec. Linnaeus (1758) is congruent with the sum of three species-level concepts sec. Weakley (2015), including *Cleistesiopsis divaricata* sec. (Fig. 1).

Thus, in addition to modeling TCLs, we need a new language to express concept-to-concept relations (Koperski et al. 2000, Geoffroy and Berendsohn 2003, Franz and Peet 2009). These relationships are of two kinds: (1) between concepts authored in a single treatment, or (2) between concepts authored in multiple treatments. The former are parent/child relationships. Example:

- *Cleistesiopsis* sec. Weakley (2015)**is a parent of** *Cleistesiopsis oricamporum* sec. Weakley (2015)

Such parent/child relationships are explicit in the hierarchy asserted by the particular treatment. And the latter, between-hierarchies relationships are RCC–5 articulations, where "RCC" stands for Region Connection Calculus (Randell et al. 1992, Thau and Ludäscher 2007, Thau et al. 2008). RCC–5 is a generic language for expressing to what extent two regions overlap. The five types of pairwise articulations are: congruence (A == B), proper inclusion (A > B), inverse proper inclusion (A < B), overlap (A >< B), exclusion (A {| or !} B). This allows us to say (compare with Fig. 1).

- *Cleistesiopsis bifaria* sec. Weakley (2015) < *Cleistesiopsis bifaria* sec. Kartesz (2010)
- *Cleistesiopsis divaricata* sec. Weakley (2015) ! *Cleistes divaricata* var. *bifaria* sec. Fernald (1950)
- *Cleistesiopsis oricamporum* sec. Weakley (2015) == *Cleistes bifaria* "Coastal populations" sec. Smith et al. (2004)

Armed with the new syntax (TCLs) and semantics (parent/child relationships and RCC–5 articulations), we are much closer to responding to the counter-query "Please specify your preferred name usage".

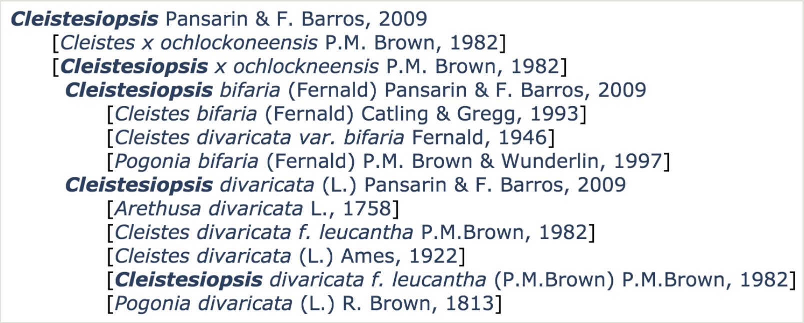
3. Design and trust - a taxonomic 'synthesis' that nobody believes in

In Fig. 2 we approximate the impact of representing and linking taxonomic concepts in an aggregated system. The specimen data are from the **S**outh**E**ast **R**egional **N**etwork of **E**xpertise and **C**ollections portal (SERNEC 2016), as of August, 2016. SERNEC is supported by Symbiota (Gries et al. 2014) and is the focus of our proposed work. The 250 records were obtained by querying the portal for "*Cleistes, Cleistesiopsis*". Georeferencing was added to ensure that all specimens are Google-mapped. The dataset was then imported four times independently into a local Symbiota sandbox. One of the four imports **(2A)** was left unchanged; with specimen identifications as currently represented "sec. SERNEC (2016)". For each of the other imports, specimens were re-identified according to one of the following three floristic treatments: (**2B**) Radford et al. (1968), (**2C**) Kartesz (2010), and (**2D**) Weakley (2015). For the latter two treatments, this step was not done very thoroughly yet, because we are visualizing an argument, not a final product. We estimate 5-10% error of identifications, especially along eastern coastal regions. We then specified queries in such a way that all specimens will be mapped according to the most granular TCL(s) present in each import.

We need to be cautious in interpreting these 'mostly real' data visualizations that make SERNEC (2A) look dismal. The 250 specimens are housed in 33 different herbaria. They were vouchered over the period of 1869 to 2011, which likely means that they were variously re-/identified using any/all relevant treatments starting with Brown (1813) (Fig. 1). Many identifications are evidently not recent, because only four specimens are identified to the genus-level name *Cleistesiopsis*. This name was not coined until Pansarin and Barros

(2008), Pansarin and Brown (2009), and in these treatments *Cleistes* was initially placed into synonymy of *Cleistesiopsis*.

But this proposal is as much about the *design* of aggregating systems (trust) as it is about promoting more, and more accurate, identifications (quality). If we look at the SERNEC (2016) Taxonomic Thesaurus in Fig. 3, we note that its syntax and semantics are *systemically misdesigned* to represent the taxonomic concept label and articulation information shown Fig. 1.



Cleistesiopsis Pansarin & F. Barros, 2009
[*Cleistes x ochlockoneensis* P.M. Brown, 1982]
[***Cleistesiopsis*** *x ochlockneensis* P.M. Brown, 1982]
Cleistesiopsis *bifaria* (Fernald) Pansarin & F. Barros, 2009
[*Cleistes bifaria* (Fernald) Catling & Gregg, 1993]
[*Cleistes divaricata var. bifaria* Fernald, 1946]
[*Pogonia bifaria* (Fernald) P.M. Brown & Wunderlin, 1997]
Cleistesiopsis *divaricata* (L.) Pansarin & F. Barros, 2009
[*Arethusa divaricata* L., 1758]
[*Cleistes divaricata f. leucantha* P.M.Brown, 1982]
[*Cleistes divaricata* (L.) Ames, 1922]
[***Cleistesiopsis*** *divaricata f. leucantha* (P.M.Brown) P.M.Brown, 1982]
[*Pogonia divaricata* (L.) R. Brown, 1813]

Figure 3.

SERNEC (2016) Taxonomic Thesaurus for the genus-level name *Cleistesiopsis*, as of August 2016. [] = synonym.

In particular, the system will not permit users to submit specimen queries in accordance with a particular taxonomic perspective, other than 'the portal consensus'. And we note that 'the consensus' is actually an evolving body of data, yet without adequate version tracking through time (Scoble 2004, Berendsohn and Geoffroy 2007, Cheney et al. 2007, Franz and Thau 2010, Midford et al. 2013. At some point, someone with access to the Thesaurus may add "*Cleistesiopsis oricamporum*". There will be no special flag, however, that we now have a new 'consensus' *version.* No ability to switch back and forth between versions, or to show the differential. This is commonplace (Page and Valiente 2005, Leonelli 2013, O'Malley 2013, Winsberg et al. 2014, Hinchliff et al. 2015): often when we hear terms like consensus or synthesis in the context of aggregation, what we actually have is a poorly sourced and connected set of unitary, taxonomically incongruent classification snapshots. Each snapshot fails to represent the actual diversity of taxonomic perspectives in use at the time, and obscures the localized practice of identifying herbarium specimens to particular (preferred) sets of taxonomic concepts.

What needs to change? The prevailing name-based designs of aggregating systems improperly conflate two semi-independent processes. One might say with reasonable accuracy that, given a particular taxonomic perspective, the application of valid names and nomenclatural relationships is an undemocratic, logically contingent process. However, *adherence* to this or that perspective *is* democratic. At present, herbaria networked in SERNEC (2016) may follow one or more of five variously authored and endorsed perspectives that relate incongruently to the orchid specimens in question (Figs 1, 2).

Because the business of aligning our concepts with taxa is not finished, a consensus perspective should not be 'dictated by design' (Fig. 3). Instead, the system should explicitly model the alternative views, resolve incongruences as much as possible, and allow users to empirically assess the impact of varied outcomes.

This proposal has conceptual, technical, social, and hence trust-related implications for biodiversity data science. The difference between the four visualizations (Fig. 2) is not just that (**2A**) resolves semantically ambiguous taxonomic concept lineages, whereas (**2B–2D**) resolve unique sets of TCLs (with increasing granularity). Surely it matters in some hypothesis-testing contexts whether we apply one perspective or another (Flanagan et al. 2006). A more fundamental difference is that, *qua the process of data aggregation*, the individual, time-stamped, consenting or dissenting 'voice' of the taxonomic expert(s) has been designed away in (**2A**). The aggregation design has disenfranchised the expert authors, whose unique views are no longer accessible to configure specimen queries. But SERNEC (2016) herbaria use at least five incongruent taxonomic schemata to identify the specimens in question! In contrast, noone in particular owns the 'consensus', which may not align with any treatment used in a herbarium to perform identifications. This is not just a matter of data quality, i.e., of updating all records once to 'the latest consensus'. Instead it is a matter of poor data system design that affects trust in aggregated data. The design must improve first to enable a better data culture.

4. Intellectual merit: Creating a trusted "big" biodiversity data culture

While our development focuses on SERNEC 2016, a shift to concept-based specimen resolution potentially affects biodiversity informatics at a global scale. We believe that this field is, and should be, part of the big data movement that many sciences are experiencing. But "big data" is a malleable term, and challenges encountered while *transitioning* to the next levels of aggregation are specific to particular domains (Leonelli 2013, O'Malley 2013, Winsberg et al. 2014). *Our* challenge is this: taxonomic names and nomenclatural relationships are by design imprecise identifiers and connectors of conflicting/evolving taxonomic perspectives in aggregating systems. Most systems nevertheless legislate one view at a time, through algorithmic and social processes (Page and Valiente 2005, Jones et al. 2011, Döring 2013, Hinchliff et al. 2015, Döring 2016b). As a result, they create unitary syntheses that noone in particular owns or subscribes to (Figs 2, 3). But few biodiversity scientists would want a future where big data is equated with the disenfranchisement of individual taxonomic expertise, or where poor identifiers support only a subset of reliable inferences *because* of the way in which the data were (mis-)aggregated via so-called synthetic classifications.

The service envisioned in Fig. 2 can be built now for the SERNEC (2016) portal. Doing so is both a milestone for the taxonomic concept approach (Franz et al. 2008) and a powerful demonstration of its impact on specimen-based inferences. Why? – Because specimens are special. Many biologists understand the limitations of name-based aggregating systems 'in principle' (Kennedy et al. 2005, Patterson et al. 2016, Remsen 2016). But experiencing a herbarium portal solution developed from A to Z, where users are shown the effect of multiple incongruent taxonomic representations of *their* specimens on a map of *their* focal region, is not just novel but forceful. For many good reasons including reproducibility (Vink et al. 2012), we care greatly about specimen-based data. Once the options displayed in Figure 2 become reality, we predict that it will be hard to return to 'just names'.

Below we describe why SERNEC (2016) is ideally suited for this purpose. Thereafter we present our specific ABI Development objectives, implementation and management plans, and broader impacts.

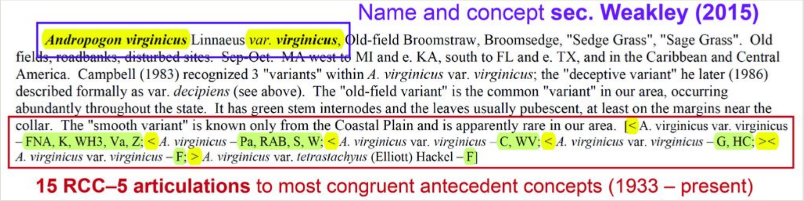
5. Why SERNEC? Review of relevant prior work

A constellation of prior efforts in four different areas uniquely identifies SERNEC as the target for developing a concept-based system.

1. **SERNEC** is an active, NSF-supported Thematic Collections Network (TCN), with currently 60 member herbaria (goal: 200) and nearly 1.97 million herbarium records (goal: 3–4 million). More than 85% of SERNEC's specimens are from the Mid-Atlantic and Southeastern United States region. Led by Zack Murrell (Appalachian State University), the community is primed to shift to taxonomic concept resolution.
2. The SERNEC (2016) portal and underlying data are sustained by the open source **Symbiota** software platform (Gries et al. 2014), hosted by the Integrated Digitized Biocollections project (iDigBio; see https://www.idigbio.org). Symbiota is the go-to solution for the majority of NSF-supported TCNs (August, 2016: 11/17 TCNs; ~ 25 portals; > 750 collections; > 32 million records; > 9 million images; > 7,300 users including 2,875 active contributors). It is the most impactful mid-level aggregator for millions of distributed, Darwin Core-compliant records (Wieczorek et al. 2012) that nevertheless allows full data access and authorship by individual collections and expert contributors. Co-PI (PI = Principal Investigator) Edward Gilbert (Arizona State University; ASU) is the lead Symbiota developer, and is singularly qualified to transform Symbiota as part of his developing biodiversity informatics research program.
3. Co-PI Bertram Ludäscher (University of Illinois Urbana-Champaign; UIUC) and PI Franz (ASU) are co-leaders of the **Euler/X** toolkit (Chen et al. 2014b, Dang et al. 2015, Ludäscher et al. 2016) for achieving logically consistent multi-taxonomy alignments (Section 6). Application of this toolkit will allow us to provide RCC–5 articulations at scale. Lastly,
4. Co-PI Alan Weakley (University of North Carolina; UNC) is the author of the **Weakley Flora** (Weakley 2015; see http://www.herbarium.unc.edu/flora.htm and Fig. 4). This 1320-page treatment covers nearly the entire SERNEC region, except southern Florida (expansion in progress). Weakley specifies as valid 7,000 taxonomic concepts at the sub-/specific level. In addition, he provides RCC–5 articulations to taxonomic concepts authored in multiple relevant preceding treatments (Fig. 4; http://tinyurl.com/wf-rcc5). These logically actionable concept-to-concept reconciliation data are unique and invaluable to the project. We have on

hand 75,621 RCC–5 articulations that reconcile each of Weakley's 7,000 concepts (on average nearly 11 times) with alternative concepts stemming from 465 sources, starting as early as Coulter and Rose (1900) and

5. and reaching the present time (Schilling et al. 2015). The 20 most comprehensive treatments – e.g. Small (1933), Wunderlin and Hansen (2011), or FNA (2015) – have 1,402–5,543 articulations, for a total of 66,996 or 88.6% of all articulations.



Name and concept **sec. Weakley (2015)**

Andropogon virginicus Linnaeus *var.* ***virginicus***, Old-field Broomstraw, Broomsedge, "Sedge Grass", "Sage Grass". Old fields, roadbanks, disturbed sites. Sep-Oct. MA west to MI and e. KA, south to FL and e. TX, and in the Caribbean and Central America. Campbell (1983) recognized 3 "variants" within *A. virginicus* var. *virginicus*; the "deceptive variant" he later (1986) described formally as var. *decipiens* (see above). The "old-field variant" is the common "variant" in our area, occurring abundantly throughout the state. It has green stem internodes and the leaves usually pubescent, at least on the margins near the collar. The "smooth variant" is known only from the Coastal Plain and is apparently rare in our area. [< A. virginicus var. virginicus – FNA, K, WH3, Va, Z; < *A. virginicus* – Pa, RAB, S, W; < *A. virginicus* var. *virginicus* – C, WV; < *A. virginicus* var. *virginicus* – G, HC; >< *A. virginicus* var. *virginicus* – F; > *A. virginicus* var. *tetrastachyus* (Elliott) Hackel – F]

15 RCC–5 articulations to most congruent antecedent concepts (1933 – present)

Figure 4.

Species concept entry Weakley's Flora (Weakley 2015: 355), with 15 RCC–5 articulations to name usages endorsed in incongruent, preceding treatments (such as Radford et al. 1968 ["RAB"]).

What do Weakley's RCC–5 articulations signal? Weakley's articulations measure the performance of taxonomic names as identifiers of taxonomic meanings (Table 1; Geoffroy and Berendsohn 2003, Franz et al. 2016a, Franz et al. 2016b). The signal is good for the majority of names: 57.1% of the pairwise RCC–5 relationships involving the same name(s) are also taxonomically congruent (==). For the remaining cases (42.9%), either the meanings are incongruent (2.9%), or the names are different (18.3%), or both (21.7%). We note, however, that these are pairwise articulations, in each case directed from Weakley (2015) to one preceding flora. The percentage of reliable names decreases if longer concept lineages are measured (Franz et al. 2016a).

Table 1.

Name-to-meaning reliability analysis of Weakley's RCC–5 data. ***Bold & italized font*** = reliable names (in pairwise alignments); regular font = name and/or meaning change; <u>underlined font</u> = totals.

Relationship (RCC–5 / names)	==	>	<	><	?	<u>**Totals**</u>
Same name(s)	***43,185***	625	1,540	15	24	<u>45,389</u>
Different names	13,836	6,433	9,000	228	735	<u>30,232</u>
<u>**Totals**</u>	<u>57,021</u>	<u>7,058</u>	<u>10,540</u>	<u>243</u>	<u>759</u>	<u>75,621</u>

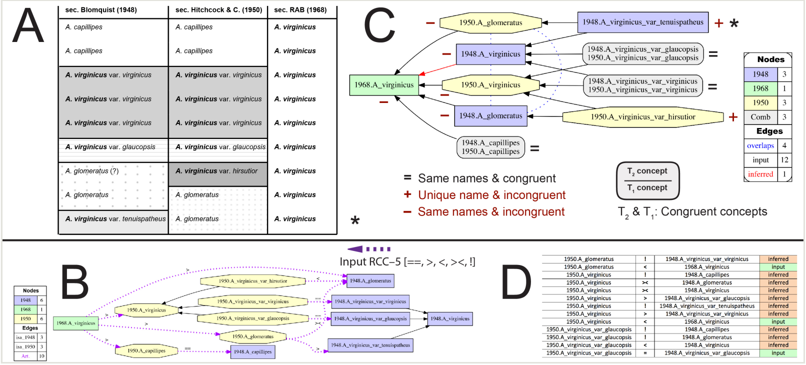
Furthermore, Weakley's work focuses on providing *one* lowest level, closest matching articulation to a concept in another treatment. This has numerous implications. **(1)** Weakley's articulations do not directly address the genus level, although often species-level incongruences will propagate up (Fig. 1), affecting genus-level SERNEC queries (Fig. 2).

The proportion of taxonomic conflict increases at this level, and hence names perform worse (Franz et al. 2016b). **(2)** Weakley does not directly provide RCC–5 articulations between concepts authored in two non-Weakley treatments. However, such articulations are needed for the service envisioned in Fig. 2. In the great majority of cases the articulations are logically implied, due to transitivity constraints (e.g., if A == B and B > C, then A > C; (Geoffroy and Güntsch 2003, Thau et al. 2009, Thau et al. 2010, Chen et al. 2014a, Chen et al. 2015). Euler/X is designed precisely to remove the two deficiencies. **(3)** Weakley's articulations are effectively a 'heat map' of taxonomic conflict. If we filter the RCC–5 data for "><" (overlap) or "?" (uncertainty), we get exactly that set of 1,002 articulations where controlling for the taxonomic variable will have the most impact. Such a 'trouble' filter has two invaluable functions: it concentrates our work on a feasible subset of all records, and it provides an empirically preselected set of use cases where the positive effects of transitioning to concept-based resolution of specimen data will be maximal. For these reasons, Weakley's Flora and RCC–5 articulations, and hence Euler/X, SERNEC, and Symbiota – are uniquely indicated to carry concept taxonomy to the specimen level, at scale.

6. Achieving scale with the Euler/X reasoning toolkit

Application of Euler/X to data explicit and implicit in Weakley's Flora will yield 1.5–2.5 *billion* additional RCC–5 articulations. Here is how. The toolkit ingests two or more taxonomies ($T_1,T_2,\ldots,T_N$) at a time (Franz et al. 2015). These are read off the respective floras, manually if needed. The concepts entailed in each have parent/child (*is_a*) relationships. Further input includes the between-taxonomy RCC–5 articulations (*A*) by Weakley, and finally a set of covering constraints (*C*) applicable to taxonomic hierarchies (e.g., no parent is childless; siblings are reciprocally disjoint) (Thau and Ludäscher 2007). Euler/X is a powerful, specialized, logic constraint solver (Chen et al. 2014b, Franz et al. 2016a, Franz et al. 2016b). The toolkit represents the entirety of the input ($T_{1\text{–}N},A,C$) as a set of constraints which, if consistent (not self-contradictory), will generate one or more output *alignments* of the input taxonomies, as dictated by the RCC–5 articulations and all logically implied relations. The toolkit's main functionality is to interactively infer – driven by user inputs – output alignments that are both consistent and maximally precise (Chen et al. 2014a, Chen et al. 2015, Franz et al. 2015).

The input constraints and derived alignments can be visualized (Fig. 5:A–C). The alignment visualizations (Fig. 5:C) are directly interpretable by users as multi-taxonomy integration maps, i.e., they are logic-vetted analogs of Fig. 1 that communicate how the input taxonomies are congruent or not, and therefore can or cannot be integrated. Just as critical is the reasoner-inferred set of *Maximally Informative Relations* (MIR), output either as a simple spreadsheet (Fig. 5:D) or via a scalable matrix visualization tool (Dang et al. 2015). For any concept pair, there are 32 possible RCC–5 articulations (2^5) if we allow for uncertainty to be expressed by disjunctions (A {== *or* >} B; etc.). The MIR are exactly that set of relations from which the truth or falseness of any relation in the "R_{32} lattice" for all concept pairs can be deduced. The formula for the number of MIR inferred via reasoning is simple: If T_1 has *m* concepts and T_2 has *n* concepts, then we obtain *m* * *n* MIR. This means, e.g., that aligning Weakley (2015) with the Flora of North America (FNA 2015) generates 7000 * 5408 = 37,856,000 MIR logically implied by Weakley's input. Thus if we took the 20 most articulated non-Weakley treatments and produced a matrix of reciprocal pairwise alignments for all of them, we would obtain 2,586,545,944 or nearly 2.6 billion MIR.


A
sec. Blomquist (1948)
sec. Hitchcock & C. (1950)
sec. RAB (1968)
A. capillipes
A. capillipes
A. virginicus
A. capillipes
A. capillipes
A. virginicus
A. virginicus var. virginicus
A. virginicus var. virginicus
A. virginicus
A. virginicus var. virginicus
A. virginicus var. virginicus
A. virginicus
A. virginicus var. virginicus
A. virginicus var. virginicus
A. virginicus
A. virginicus var. glaucopsis
A. virginicus var. glaucopsis
A. virginicus
A. glomeratus (?)
A. virginicus var. hirsutior
A. virginicus
A. glomeratus
A. glomeratus
A. virginicus
A. virginicus var. tenuispatheus
A. glomeratus
A. virginicus
*
C
–
1950.A_glomeratus
1948.A_virginicus_var_tenuispatheus
+ *
–
1948.A_virginicus
1948.A_virginicus_var_glaucopsis
1950.A_virginicus_var_glaucopsis
=
1968.A_virginicus
–
1950.A_virginicus
1948.A_virginicus_var_virginicus
1950.A_virginicus_var_virginicus
=
–
–
1948.A_glomeratus
1950.A_virginicus_var_hirsutior
+
1948.A_capillipes
1950.A_capillipes
=
Nodes
1948 3
1968 1
1950 3
Comb 3
Edges
overlaps 4
input 12
inferred 1
= Same names & congruent
+ Unique name & incongruent
– Same names & incongruent
T₂ concept
T₁ concept
T₂ & T₁: Congruent concepts
Input RCC–5 [==, >, <, ><, !]
B
Nodes
1948 6
1968 1
1950 6
Edges
isa_1948 3
isa_1950 3
Art. 10
1968.A_virginicus
1950.A_virginicus
1950.A_virginicus_var_hirsutior
1950.A_virginicus_var_virginicus
1950.A_virginicus_var_glaucopsis
1950.A_glomeratus
1950.A_capillipes
1948.A_glomeratus
1948.A_virginicus_var_virginicus
1948.A_virginicus_var_glaucopsis
1948.A_virginicus
1948.A_virginicus_var_tenuispatheus
1948.A_capillipes
>
<
==
==
><
>
>
>
==
>
D

1950.A_glomeratus	!	1948.A_virginicus_var_virginicus	inferred
1950.A_glomeratus	<	1968.A_virginicus	input
1950.A_virginicus	!	1948.A_capillipes	inferred
1950.A_virginicus	><	1948.A_glomeratus	inferred
1950.A_virginicus	><	1948.A_virginicus	inferred
1950.A_virginicus	>	1948.A_virginicus_var_glaucopsis	inferred
1950.A_virginicus	!	1948.A_virginicus_var_tenuispatheus	inferred
1950.A_virginicus	>	1948.A_virginicus_var_virginicus	inferred
1950.A_virginicus	<	1968.A_virginicus	input
1950.A_virginicus_var_glaucopsis	!	1948.A_capillipes	inferred
1950.A_virginicus_var_glaucopsis	!	1948.A_glomeratus	inferred
1950.A_virginicus_var_glaucopsis	<	1948.A_virginicus	inferred
1950.A_virginicus_var_glaucopsis	=	1948.A_virginicus_var_glaucopsis	input

Figure 5.

Euler/X toolkit products. **(A)** Part of the complex "*Andropogon* use case", with 1948/1950/1968 input sec. Weakley (2015) (see also Fig. 1). **(B)** Euler/X visualization of input constraints. Each concept taxonomy is uniquely colored. Ten RCC–5 articulations are provided. **(C)** Euler/X alignment visualization, showing congruent (grey, rounded) and incongruent (colored) alignment regions. Arrows represent proper inclusion; blue dashed lines show overlap. **(D)** Subset of 48 inferred, *Maximally Informative Relations* (MIR). The "*" represents the 1948. *Androgon*_var_*tenuispatheus* concept lineage throughout (A) and (C).

Application of Euler/X will generate vast numbers of concept-to-concept relations that speak directly to the query: "To what extent can these two concepts be integrated?" The toolkit will create a comprehensive corpus of RCC–5 signals that will newly drive name-based integration for SERNEC specimen data.

7. ABI Development objectives: Taxonomic concept resolution for SERNEC

We target the ABI Development level because our innovations consist primarily of key increments to well-established service components, and in making the newly integrated infrastructure work in conjunction with SERNEC's specimen data.

1. Through annual workshops and continuous research on specific use cases, we will foster **community engagement** to transform SERNEC into the first concept-based herbarium specimen data culture.
2. We will apply the Euler/X reasoning toolkit to generate comprehensive genus- to variety-level **concept alignments** for at least 10 major treatments with relevance to SERNEC. Both visual and MIR-based toolkit products (> 10^9 RCC–5 articulations) will be stored in a GitHub repository for user access from the portal, and the RCC–5 articulations will drive the future reconciliation of specimen data.

3. We will expand and optimize the **Symbiota** taxonomy and occurrences modules, and critical related user interfaces, to represent both **name- and concept-based** information. We will build **modules** to manage multiple concept taxonomies, utilize Euler/X reasoning products for specimen queries and visualizations, and perform semi-automated upgrades of identifications to entirely transition to TCLs.
4. Using Symbiota's new identification module and the expertise of the UNC team and SERNEC collaborators, we will **augment** minimally 80% of the in-region **specimen identifications** (up to 1.6 million) from the current name-level to the most granular available TCL. Where needed, the identifications (and hence specimens) will be flagged with an "uncertainty signal" (need for study).
5. We will research and publish at least 12 **use cases** "on the impact of controlling for the taxonomic variable" that showcase the diverse scientific (evolution, ecology) and societal (global change, conservation) significance of further developing a more robust, concepts-to-specimens data culture.

8. Research and implementation plan: Realizing objectives 1-5

8.1. Community engagement. Community engagement is absolutely critical because we aim to build a new data culture by example. Working with the SERNEC (2016) community is advantageous because this community is highly active and coherent, with ASU (Franz, Gilbert) presently acting as a TCN grant subawardee in charge of all Symbiota-related development. SERNEC botanists are deeply familiar with Weakley's Flora, and his concept mapping efforts are well known, valued, and used (Weakley 2009, Weakley 2014).

To further deepen the engagement, we have identified a core of 12 use cases that will be taken up from the planning to the publication stage by leaders within the SERNEC community (Section 8.5). To directly engage SERNEC scientists, we will hold **annual workshops** (2nd quarter of each project year) with as many as 10 non-local invitees plus 10-20 local participants at UNC (year 1), the University of Florida and iDigBio (year 2), and Appalachian State University (year 3). Workshop goals will evolve with the advancement of use cases. Each workshop will run for two full days, plus travel. During the interim periods, we will communicate virtually with use case groups (e.g. via iDigBio's Adobe Connect) and through monthly updates to the SERNEC (2016) community and the iDigBio/Symbiota Working Group (https://www.idigbio.org/wiki/index.php/Symbiota_Working_Group). We budget funds to cover presentation costs at inter-/national **conferences**, and publication costs for peer-reviewed **publications** of use case outcomes in impact-maximizing open access (option) journals.

8.2. Euler/X concept alignments. ASU and UIUC will concentrate on this task, with Franz mentoring undergraduate students at ASU to create and publish the alignments, and Ludäscher mentoring a graduate (Ph.D.) student at UIUC to develop new toolkit capabilities for special reasoning and visualization challenges of the SERNEC use case. Weakley's group (UNC) will provide expert input as needed.

I. Scope. We will produce comprehensive – all with all – alignments for the 6-12 most abundantly applied treatments for SERNEC, given the taxonomic subgroup (see Weakley 2015: 7–10). The alignments will be partitioned by family rank sec. Weakley (2015), yielding 413 high-level partitions. This means that input data for, e.g., Orchidaceae sec. Weakley (2015) will include 33 genus-level concepts and their respective children – to be reconciled with orchid concepts authored in one or more (up to 12) additional treatments. However, we will represent each taxonomy as a multi-rooted hierarchy with (in this example 33) unconnected genus-level parents. This is because we will *not* infer family-level congruence with RCC–5; at that level, name identity 'takes over'. Instead we will model concepts exclusively at the genus- to variety-level, in two complementary ways called "extensional" and "intensional" (Franz and Peet 2009, Franz et al. 2015, Franz et al. 2016b). Roughly, extensional alignments compute genus-level congruence *only* from the joint signal provided by the children. Sampling of children matters greatly then: if one treatment lists one additional species-level concept – perhaps an introduced entity or one that occurs within the treatment's greater geographic range – that logically means *non*-congruence at the genus level. Intensional alignments, in turn, respond to the question: "regional sampling issues aside, are these two treatments in obvious conflict in terms of feature-based diagnostics, or not?" As an example, Arizona and Florida have different sets of oak species (non-congruence, extensional), but two concepts of *Quercus* sec. the respective flora treatments (Arizona, Florida) likely congruently entail "all plants that have acorns and no plants that lack acorns" (congruence, intensional). Both types of alignments are valuable for different purposes, either (1) for bringing out fine differences among parent-level taxonomic concepts or (2) for maximizing concept integration at the generic level across differentially sampled floras. We have optimized Euler/X translations of either kind (Franz et al. 2015).

II. Feasibility. The task of producing two types of 413 family-level alignments that are 6-12 taxonomies deep and reciprocally comprehensive may appear daunting. We are certain that it is not, given prior experience, efforts, and project resources allocated to this task. The reasoning capacity is already there (Franz et al. 2016b). At present, Euler/X can complete on a laptop (> 13 hours) pairwise alignments of taxonomies that are 7-11 ranks deep, have 3,200 input concepts, and yield 2.46 million MIR. In comparison, the most speciose family Asteraceae sec. Weakley (2015) has a total of 1,120 concepts. Most other treatments are less granular, but even doubled (2,240 concepts) these alignments are already well within range (< 2 hours) if done pairwise. In fact they will be faster because we only represent three ranks (genus to variety), which greatly diminishes the reasoning burden. In cases nearing the scale of the Asteraceae, we will be limited to aligning 2-3 treatments at a time. But for many less diverse cases we can align 5-10 taxonomies at once with current capacity, which we expect to increase with further development at UIUC.

III. Approach. We have run thousands of successful alignments with Euler/X, including larger sub-alignments (all Gymnospermae, all Rosaceae) of Weakley (2015) versus Radford et al. (1968). Based on these experiences, and given (1) Weakley's comprehensive RCC–5 articulations and (2) the predominance of congruent (==) signals therein (75.4%; see Table 1), we are confident that a majority of consistent, well-specified alignments are achievable using simple translation scripts. The RCC–5 spreadsheet can be directly reconfigured (e.g., with Python scripts) to extract, hierarchically arrange, and pre-articulate all concept sets as Euler/X input (Franz et al. 2015), leaving only minimal cleaning and completion work for us to infer the alignments. We will generate these scripts (UIUC, UNC), build up efficient workflows, and scale them from less to more complex alignments. The majority of the hands-on work will take place at ASU, under Franz' lead, carried out by a rotating group of undergraduate students who will be trained in systematic theory, toolkit use, and the interpretation/resolution of conflicting treatments. We aim to generate **2,000-3,000 alignments** – on average 6-10 per family sec. Weakley (2015) – and more than **10^9 logically inferred RCC–5 articulations.** The latter will drive specimen data integration in the transformed SERNEC portal. These products will be openly accessible – also via links from portal interfaces – in a structured GitHub repository (https://github.com/taxonomic-concept-alignments/SERNEC).

IV. New Euler/X development. SERNEC 2016 poses special challenges for reasoning and visualization. We identify three critical needs. **(1)** In 2016 Ludäscher's UIUC team released the first toolkit version with a custom RCC–5 reasoner, on which the 3000+ input concepts per alignment capacity is based. However, much of the process of generating alignments relies on constraint specifications, logic-based diagnostics, and visualization functions that remain operationally linked to off-the-shelf "Answer Set Programming" reasoners (Chen et al. 2014a, Chen et al. 2014b, Chen et al. 2015). One example is the representation of hybrid concepts (Fig. 3; *Cleistesiopsis* x *ochlockneensis*), achieved by locally relaxing the sibling disjointness constraint (Thau and Ludäscher 2007). To fully deploy all needed Euler/X functions at increasingly large scales, we will require **further toolkit development** in connection with custom RCC–5 reasoning. **(2)** The SERNEC alignments have few ranks per taxonomy (range: 2-3), but are deep in terms of numbers of taxonomies (range: 2-12). Moreover, the RCC–5 input articulations are "star-like", i.e., always directed from Weakley 2015 to one alternative treatment. We need to generate matrices of all possible pairs. This will require new logic development and application of **RCC–5 *transitive* reasoning.** We will review and advance how each element of the R_{32} lattice (see Section 6) can be propagated along a chain of three or more concepts. This problem was initially explored by Geoffroy and Güntsch 2003 under the theme of "navigating the potential taxon graph". But the theoretical work remains under-developed, and solutions have not yet been implemented (Thau 2010, Thau et al. 2010). The subject is therefore highly suitable for a Ph.D. thesis in computer science, to be mentored at UIUC by Ludäscher. The SERNEC use case is ideal for optimizing the task of RCC–5 propagation beyond paired taxonomies, which addresses general and specific needs to develop comprehensive RCC–5 coverage for all SERNEC taxonomies. **(3)** Lastly, we need to visualize alignments of concept-rich but rank-poor taxonomies. The task entails special color/shape assignment and spatial efficiency constraints (Graham and Kennedy 2010, Dang et al. 2015). Indeed, if we limit ourselves to the most granular level, Weakley's tabular views are most effective (Figs 1, 5:A). To find the best pragmatic solutions for 2-3 levels and 2-12 taxonomies, we will create **custom stylesheets** that translate the toolkit products (Fig. 5) into easy-to-interpret GraphViz (Gansner and North 2000) and PDF outputs for users.

8.3. Adding taxonomic concept representation to Symbiota and SERNEC. This objective requires a large part of the project's resources for new Symbiota development at ASU. Three major task domains are involved. (1) Symbiota's underlying taxonomic and occurrence schemas must be expanded to support TCLs, source-specific parent/child relationships, and RCC–5 articulations. (2) A subset of Symbiota's graphic user interfaces will be changed accordingly, and new interfaces will be created to manage multiple taxonomies and efficiently upgrade specimen identifications to TCLs. (3) A name-to-concept transition plan for the SERNEC portal will be executed such that (a) existing named-based data are not functionally compromised and (b) new concept-based data become the portal norm – most immediately to support our use cases. Below we describe the sequence of actions that will achieve this transition.

I. Schema expansion. The expansion of Symbiota's taxonomic and occurrence modules will be guided by the remarkable example of Avibase (Lepage et al. 2014; http://avibase.bsc-eoc.org) which manages more than 1.5 million taxonomic concepts. Symbiota's current taxonomic module (http://symbiota.org/docs/taxonomic-schema-2) is a well-developed branch of the data structure that already permits source-specific views on valid versus synonymous names. As in the case of Avibase, we will transition to maintain a single, specific higher-level classification for higher taxonomic levels (kingdom to family; Weakley 2015, Stevens 2016), but represent taxonomic concepts at lower levels where the impact on SERNEC specimen searches is most significant. We will complete and release a new "references" module, and add new tables to explicitly model TCLs, parent/child relationships, and RCC–5 articulations. We will also 'elevate' the representation of synonymy relationships to the concept level, meaning that nomenclatural relationships between valid and synonymous names (each rendered as TCLs) will be displayed as a specific function of the treatment that recognizes them (unlike Fig. 3). Once transitioned, we will propagate default settings for occurrences that require using TCLs from the expanded taxonomic module.

II. User interfaces. We will upgrade a critical subset of Symbiota's interfaces to enable concept use. In particular, we will modify the **taxonomy viewing and editing interfaces** to only accept the new syntax and semantics. New concept taxonomies can be uploaded piece-meal or through batch functions. We will also create simple formatting and loading tools to ingest multiple taxonomies into the Euler/X alignment toolkit and re-integrate the outcomes (MIR) into the RCC–5 table. Based on the latter, we will generate a new "incongruence alert" table that entails precisely those taxonomic names that, when searched by users, require additional specification of a TCL to identify a consistent circumscription (Figs 1, 2).

Symbiota already has an effective visualization interface for single taxonomies. Rather than building a new multi-tree visualization interface – which is both difficult and redundant (Graham and Kennedy 2010, Dang et al. 2015, Hinchliff et al. 2015) – we will (1) provide drop-down menus for users to select and view a taxonomy at a time, and (2) links for each genus-level concept to the corresponding set of Euler/X alignments with other treatments stored in the GitHub repository. This approach is more modular and feasible than dynamically integrating Euler/X alignment visualizations within Symbiota, while allowing users to explore SERNEC's classifications interactively.

We will reconfigure the **occurrence identification interface** to interact with the new taxonomy module. Again, this will include drop-down options to select preferred sources, view alignments, and populate a TCL. Very substantive upgrades will be made to the **add batch determinations** interface, which presently permits selecting names or individual specimens. In collaboration with SERNEC experts, we will expand this interface to represent the subset of Darwin Core fields most decisive for filtering occurrences so that batch updates can follow. Target fields will include (e.g.) the source collection, collecting locality and date, collector/identifier information (who/when?), and references (where available). Combinations of these fields will facilitate smart queries of the kind: "show me all specimens in this region, collected in this time period, and identified to this name by members of this herbarium community". These queries, combined with expert knowledge and specimen images, will facilitate upgrades of many identifications to TCLs at once. A second, innovative **map-based determinations** function will be developed as an extension of Symbiota's Map Search module. It will allow experts to gather specimen sets for batch determinations directly through the map interface (http://tinyurl.com/sernec-mapint), by using area shape selectors. Because granular taxonomic concepts are often geographically separated (Fig. 2:D), the function will allow finding *all* specimens in need of TCL identification conveniently, and for any given region.

Lastly, we will transform the primary **search and display specimens** interfaces. Key goals are to promote TCL-based specimen queries and mappings, with the option to *relabel* an initially queried set according to an alternative treatment (Fig. 2). Superficially, the specimen query interface will not change much. However, if for instance the user enters the name "*Cleistes divaricata*", this name is pre-identified in the "alert" table to have incongruent meanings (see above). Hence the user will be prompted with the counter-query "please specify further". They will see a drop-down list of treatments, and links to alignments on GitHub that will explain the concept resolution issues. Once a TCL is chosen (which remains optional – we will never prohibit name searches), the output is rendered accordingly (Fig. 2). At this point users have two additional options to assess robustness. (1) They can rerun the query with alternative TCLs and compare maps, which may show different specimen sets; or (2) they can rerun the query on the same specimen set but with different TCLs. We will highlight options to view and save TCL-based query and relabeling outcomes. Symbiota is known for the relative ease of moving data; typically only 2-3 clicks are needed to extract datasets in various formats, including Darwin Core Archive (GBIF 2010). This ensures that explorations of the taxonomic variable are fully documented and reproducible.

III. Transition plan. Realizing the above changes requires a sound transition plan. It is critical not to break existing services while building new ones for transition. We also recall that 57.1% of the RCC–5 articulations identify reliable name usages that (at present) need no additional specification (Table 1).

Once the taxonomy module is expanded, it is necessary to 'reify' the SERNEC consensus taxonomy (Fig. 3) *at a specific date* to become static. This means, for instance, that a specimen identified only to the name "*Cleistes divaricata*" will now be automatically updated to the TCL "*Cleistes divaricata* sec. SERNEC consensus, March 1st, 2018", where the latter is an immutable classification. Multiple (6-12) rigorously sourced and aligned classifications will be added under the new conventions. Submitting 'just names' or new TCLs without RCC–5 articulations is no longer admissible. We thereby shift to the new culture. However, for reasons explained in Sections 1–3, the reified 'TCLs' of the SERNEC consensus cannot be precisely aligned with TCLs of specific treatments. Initially, we will have created a syntax change without actually improving the core semantics (!). Yet this gives us an excellent success metric: the degree to which we will have transitioned from name- to TCL-based specimen resolution is directly proportional to the number of "sec. SERNEC 2018" TCLs that are *re-identified* to a specific source ("sec. Weakley 2015", etc.). In other words, the syntactic shift to TCLs for SERNEC specimens is instantaneous (March 1st, 2018) and will affect *all* specimens identified only to a name at that time. In the following, gaining semantic resolution becomes a function of gradually replacing 'consensus' TCLs with (typically the most granular) externally sourced TCLs. The use cases will facilitate the beginnings of this transition.

8.4. Augmenting SERNEC specimen identifications to TCLs. Using the new tools, our goal is to upgrade minimally **80% of all specimen identifications** from "sec. SERNEC 2016" to either (1) a specific, maximally granular treatment or (2) an added "uncertainty signal". For the use cases, which – as shown in Table 2 – compromise 35,422 specimens, we will aim for close to 100% granular TCL identifications.

Table 2.

Overview of 12 use cases. Headers: valid names sec. Weakley 2015; species- and variety-level diversity sec. Weakley 2015; number of SERNEC specimens (August, 2016); number of corresponding SERNEC names; reliability ratio from Weakley's RCC–5 data (see Table 1: ***bold italics*** versus regular font cells), showing reliable vs. unreliable names and the percentage of reliable names; biological significance for differential concept resolution (**Con**servation, **Cul**tivars, **Dis**tribution, **Div**ersity, **Eco**logy, **Evo**lution, **Exo**tics, **G**lobal **C**hange **B**iology, **H**istorical **B**io**G**eography, **Phy**logeny, **Pol**lination); and leading SERNEC expert.

#	**Names sec.** Weakley 2015	**Taxonomic diversity sec.** Weakley 2015	**Specimens in** SERNEC 2016	**Names in** SERNEC 2016	**Reliability ratio**	**Impact**	**Use case lead**
1	***Andropogon* "complex"**	7 species | 4 varieties	2,696	16	**14 : 90** (13.5%)	Dis - Div - Evo	Weakley
2	***Asarum* & *Hexastylis***	14 species | 5 varieties	3,564	36	**87 : 110** (44.2%)	Dis - Div - Phy	Murrell
3	***Cleistes* & *Cleistesiopsis***	3 species	250	12	**8 : 47** (14.6%)	Con - Dis - Phy	Weakley

4	***Coreopsis***	23 species | 11 varieties	4,561	56	**185 : 155** (54.4%)	Eco - Evo - HBG	Weakley
5	***Cornus***	11 species	5,575	40	**104 : 63** (62.2%)	Eco - Evo - HBG	Murrell
6	***Euphorbia***	50 species	5,747	190	**247 : 213** (53.7%)	Con - Dis - Eco	Alford
7	***Galactia***	7 species | 1 variety	1,408	23	**61 : 49** (55.5%)	Dis - Div - Eco	Franck
8	***Gonolobus* & *Matelea***	9 species | 2 varieties	1,571	28	**48 : 43** (52.7%)	Eco - Evo - Phy	Fishbein
9	***Lantana***	4 species | 1 variety	659	22	**22 : 19** (53.6%)	Eco - Exo - GCB	Franck
10	***Liatris***	28 species | 4 varieties	4,200	70	**121 : 185** (39.6%)	Con - Evo - Pol	Alford
11	***Magnolia***	9 species | 4 varieties	5,135	45	**46 : 114** (28.8%)	Cul - Evo - HBG	Weakley
12	***Monotropis***	2 species	56	4	**13 : 19** (40.6%)	Eco - Evo - GCB	Weakley

Improving identifications will be facilitated by technology, but is only feasible because of the direct involvement of experts. Some 15-30% of the SERNEC records have partial identification-related information recorded (expert, year, taxonomic reference used). We will utilize these data to identify the best-fitting TCL. Collective experience strongly indicates that, even for problematic cases, a remotely working expert can confidently assert TCL identifications by drawing on their sophisticated background knowledge of spatially/ temporally localized identification practices. For instance, a very large number of non-Floridian SERNEC herbaria have treated Radford et al. 1968 as "the bible" during the past four decades. Weakley's concepts tend to be more granular, but in geographically and/or ecologically structured patterns (Fig. 2:D). Thus, all specimens identified as "*Cleistes divaricata*" but documented from Gulf Coast savannas can be confidently updated to "*Cleistesiopsis oricamporum* sec. Weakley 2015". The scope of the SERNEC TCN grant is critical here: all newly digitized herbarium sheets are scanned/uploaded as high-resolution images (http://tinyurl.com/sernec-at-idigbio). Presently, this includes 1.25 million or 63.6% of all records – a proportion that will only increase with time. On average, our use case experts will have these images available to make maximally granular identifications in 2/3 of all ambiguous cases.

Thousands of herbarium sheets will nevertheless remain "uncertain" with regards to the narrowest TCL. Removing uncertainty may require direct study of morphological/molecular data, likely in the context of new revisions. Although hypotheses are weakened in such cases, we regard this as a positive contribution to explicitly identify 'problem specimens'. In analogy to the "alert" table for incongruent name usages, we will create a special flag for uncertain TCL identifications. Flagged specimens will be retrievable by query, and uniquely colored on maps, with an option to show only non-ambiguous specimens.

8.5. Use case selection, approach, and impact. We have enlisted five SERNEC botanists (plus the UNC postdoc; Section 9) to lead 12 use cases (Table 2). The selection rationale is as follows. Use cases (1) must be feasible in terms of taxonomic and specimen volume; (2) entail significant conflicts in taxonomic concepts; (3) have availability of taxonomic experts; and (4) differential concept resolution should affect the robustness of basic and applied biological inferences. We are confident that several hundreds of such use cases exit within the SERNEC domain. Hence the current selection and teams are subject to much expansion, but have emerged as a strategy to best demonstrate impact. We briefly describe each.

I. Use case particulars. The ***Andropogon glomeratus-virginicus*** **"complex**" is notorious for its taxonomic instability (Campbell 1983, Weakley et al. 2011, Weakley 2015). An Euler/X analysis of 11 classifications (Franz et al. 2016a) revealed a name:meaning cardinality ranging from 1:6 to 4:1, with only 1/36 names being reliable in the sense of Table 1. Different taxonomic schemata dramatically alter the recognition of species- and variety-level diversity, ranges, ecological preferences (hydrology), and hybridization hypotheses for these "bluestem" grasses. Weakley will guide this paradigmatic concept use case. ***Asarum*** and ***Hexastylis*** (birthwort family) have longstanding 'porous' generic limits (Kelly 1998), disagreements regarding species-level diversity, and hence also distribution patterns (Weakley 2015). Murrell, who has extensively studied "concept creep" in this group using electrophoretic and morphological methods (Murrell et al. 1998), will lead. ***Cleistes*** and ***Cleistesiopsis*** sec. auctorum entail concepts for rare, endangered "Pogonia" orchids (Pansarin and Barros 2008, Pansarin and Brown 2009, Smith et al. 2004). Alternative views have disparate implications for distribution ranges (Figs 1, 2) and thus for conservation management decisions (Gregg and Kéry 2006). This use case, headed by Weakley, is particularly suited to replicate the seminal (but pre-Darwin Core) Peterson and Navarro-Sigüenza 1999 analysis of species concepts altering conservation priorities – at the *specimen* level. The composite ***Coreopsis*** is diverse and poorly delimited from *Bidens* (Kim et al. 1999, Crawford and Mort 2005, Sorrie et al. 2013). Our hypotheses of narrow ecological specialization, the evolution of polyploidy, and historical biogeography in the Southeast are contingent on the taxonomic variable. Weakley and postdoc will lead. The use case for ***Cornus*** (dogwood family) will be coordinated by specialist Murrell (Murrell 1993, Xiang et al. 2006). Several complex subgroups and possible hybrids are involved; with implications for understanding Tertiary continental migration patterns.

Euphorbia (spurge family) is the most speciose complex, including recent introductions not yet keyed out by Weakley 2015. Many species-level groups are geographically limited and require specialized edaphic conditions, with implications for biogeography, population genetics, conservation, and even medicinal discovery (Park 1998, Dorsey et al. 2013, Ernst et al. 2016). Mac Alford (University of Southern Mississippi) will coordinate. ***Galactia*** is a legume complex (Duncan 1979, Isely 1998, Ward and Hall 2004) and "one of the worst genera ever" (Alan Franck, University of South Florida, pers. comm.), with incongruences in species delimitation and distribution that will only increase after an upcoming revision co-authored by Franck who will take on this use case. ***Gonolobus*** and ***Matelea*** (dogbane family) are variously 'joined' or not and entail several contentious subgroups, with differential outcomes for our understanding of their distributions, reproductive biology, and breeding system evolution (Lipow and Wyatt 1998, Krings 2008). Mark Fishbein (Oklahoma State University) will concentrate on this complex. ***Lantana*** is a small use case, with Franck leading, where incongruent concepts involve ecological interactions of endemic/ native and invasive/exotic groups (Sanders 2006, Sanders 2012) – an important theme for the flora of Florida (Lee et al. 2009). The composite ***Liatris*** (Mayfield 2002, FNA 2015, Weakley 2015) is the second most speciose use case, with Alford as lead, where several inconsistently recognized groups are threatened in eastern prairie habitats. The group is also a model for understanding density-dependent foraging (Figlar and Nooteboom 2004). For ***Magnolia***, 12 incongruent schemata have been published (Figlar and Nooteboom 2004, Sima and Lu 2012, Weakley 2015) – a situation that requires concept resolution for these abundant, cherished, and often cultivated and then naturalized groups. Weakley has aligned all views and will head the use case. Alternative schemata will affect deep- and shallow-time inferences of disjunctions and dispersal processes between temperate and tropical regions (Azuma et al. 2001, Azuma et al. 2011). Lastly, ***Monotropis*** (heath family) is the smallest use case, though very complex due to different family- to species-level assignments. Members have specialized heteromycotrophic interactions with fungi and their (oak or pine) hosts, and these are variously captured in taxonomic perspectives (Klooster and Culley 2009, Rose and Freudenstein 2014). Weakley will analyze.

II. Research approach. We expect that use case leaders will engage many additional SERNEC members. Although the TCL identification efforts (Section 8.4) will be similar in each case, the ultimate research goals will vary greatly. Some may take on the form of a review – though rooted in specimen-level data and visualizations – of taxonomic inconsistencies affecting our basic understanding of biodiversity and distribution. Others may reassess the specimen-level evidence and inter-taxonomic robustness of very specific ecological or evolutionary hypotheses. Still others may characterize to what extent conservation and global change assessments are contingent on a specific taxonomic commitment (Peterson and Navarro-Sigüenza 1999, Rosenberg 2014, Rylands and Mittermeier 2014, Borsch et al. 2015). This variety of motivations is deliberate: we will thereby demonstrate that the need to control for the effect of the taxonomic variable on integrating specimen-level data is potentially as broad as comparative biology.

III. Innovative impact. Rather than specifying each worthwhile research question, we exemplify the kinds of questions that our development will facilitate, and why this matters. Accordingly, in the case of *Andropogon*, users can query (often sequentially):

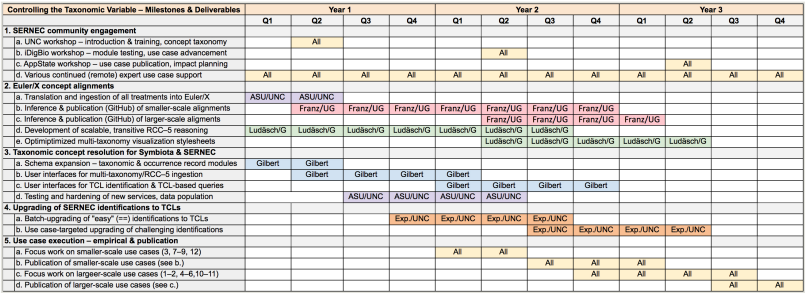
1. "Show me all specimens identified to the taxonomic name *Andropogon virginicus* in the Carolinas" [returns many records, resolved only with the ambiguity of an incongruent taxonomic concept lineage].
2. "Now show me all specimens identified to the TCL *Andropogon virginicus* sec. Weakley 2015" [returns a subset of these records, reflecting a choice for this particular, granular TCL resolution].
3. "Now me all specimens of *Andropogon virginicus* sec. Radford et al. 1968, yet translated into the TCLs sec. FNA 2015" [returns (again) many records, but specifically represents and contrasts two treatments, as opposed to providing the ambiguous lineage view of (1)].
4. "Show me all specimens whose identifications are ambiguous with regards to FNA 2015 versus Weakley 2015" [using the "uncertainty" flag, points to specimens needing further study].
5. "For the Carolinas and for this inclusive specimen set, show me the *composite* least versus most granular taxonomic perspective(s) available" [returns a potentially multi-taxonomy 'composite' that represents opposite extremes in resolution granularity].
6. "Save and output all results in Darwin Core Archive format".

This is what we mean by "controlling the taxonomic variable". The services will be basic, as dictated by realism, and the control offered to users is not explicitly of a statistical kind. Yet we are confident that queries (2-6) – which are not supported by any existing herbarium specimen network – will yield impactful outcomes when applied to the aforementioned use cases and research goals. We will work to carry each of these to publication in international journals with open access options that variously expand the reach of our approach, such as Biodiversity Data Journal, Conservation Biology, Global Ecology and Biogeography, PeerJ, PLoS ONE, Systematic Biology, Taxon, and Trends in Ecology and Evolution.

9. Lead personnel and management

The project lead personnel – Franz, Gilbert, Ludäscher, and Weakley – is introduced in Section 5. Franz will mentor 15 undergraduate students at ASU to achieve the Euler/X alignments (Section 8.2.I–III). At UIUC, Ludäscher will mentor a Ph.D. student who will concentrate on the reasoning and visualization challenges (Section 8.2.IV). At UNC, Weakley will mentor a postdoc (year 1), and utilize applications analyst Michael Lee to provide critical support at the interface of Symbiota module development, data population, and service optimization for the use cases. Weakley and the postdoc will play an immense role in overseeing the rapid integration of the floristic legacy, new tools, and expert contributions. Gilbert and Lee will translate new conceptual and TCL identification-related functions and data from various sources into the transformed SERNEC portal (Sections 8.3 & 8.4). We invest significant resources to support expert co-leadership of the use cases (Sections 8.1 & 8.5).

Fig. 6 lists the development milestones and deliverables corresponding to the five main objectives and for the three allocated project years. Symbiota's structure is modular, and the development release approach is such that all new code is immediately functional in portals that need the new functions. Select modular functions will go live in SERNEC within weeks of the project start, although the complete service as described in Section 8.4.III will take more time to be built and fully sustain the use cases.



Controlling the Taxonomic Variable – Milestones & Deliverables	Year 1				Year 2				Year 3			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1. SERNEC community engagement												
a. UNC workshop – introduction & training, concept taxonomy		All										
b. iDigBio workshop – module testing, use case advancement						All						
c. AppState workshop – use case publication, impact planning										All		
d. Various continued (remote) expert use case support	All	All	All	All	All	All	All	All	All	All	All	All
2. Euler/X concept alignments												
a. Translation and ingestion of all treatments into Euler/X	ASU/UNC	ASU/UNC										
b. Inference & publication (GitHub) of smaller-scale alignments		Franz/UG	Franz/UG	Franz/UG	Franz/UG	Franz/UG	Franz/UG	Franz/UG				
c. Inference & publication (GitHub) of larger-scale aligments						Franz/UG	Franz/UG	Franz/UG	Franz/UG			
d. Development of scalable, transitive RCC–5 reasoning	Ludäsch/G	Ludäsch/G	Ludäsch/G	Ludäsch/G	Ludäsch/G	Ludäsch/G	Ludäsch/G					
e. Optimiptimized multi-taxonomy visualization stylesheets						Ludäsch/G	Ludäsch/G	Ludäsch/G	Ludäsch/G	Ludäsch/G		
3. Taxonomic concept resolution for Symbiota & SERNEC												
a. Schema expansion – taxonomic & occurrence record modules	Gilbert	Gilbert										
b. User interfaces for multi-taxonomy/RCC–5 ingestion		Gilbert	Gilbert	Gilbert	Gilbert							
c. User interfaces for TCL identification & TCL-based queries					Gilbert	Gilbert	Gilbert	Gilbert				
d. Testing and hardening of new services, data population			ASU/UNC	ASU/UNC	ASU/UNC	ASU/UNC						
4. Upgrading of SERNEC identifications to TCLs												
a. Batch-upgrading of "easy" (==) identifications to TCLs				Exp./UNC	Exp./UNC	Exp./UNC	Exp./UNC					
b. Use case-targeted upgrading of challenging identifications							Exp./UNC	Exp./UNC	Exp./UNC	Exp./UNC		
5. Use case execution – empirical & publication												
a. Focus work on smaller-scale use cases (3, 7–9, 12)					All	All						
b. Publication of smaller-scale use cases (see b.)							All	All	All			
c. Focus work on largeer-scale use cases (1–2, 4–6,10–11)								All	All	All	All	
d. Publication of larger-scale use cases (see c.)											All	All

Figure 6.

Project management time line; see also Sections 5 & 8. G = graduate student; UG = undergraduate students.

10. Broader impacts - scientific and educational

I. Scientific. The intellectual case was presented in Sections 1–4. Our project is focused on just one herbarium network, but the potential impact is much wider. Space does not permit reviewing the many aggregators that concede, in one form or another, that taxonomic concept resolution is needed but seemingly out of reach. The list includes (e.g.) Catalogue of Life, GenBank, Global Biodiversity Information Facility, Global Names Architecture, iDigBio, Open Tree of Life, and the Taxonomic Name Resolution Service (Jones et al. 2011, Federhen 2012, Boyle et al. 2013, Döring 2013, Hinchliff et al. 2015, Döring 2016b, Patterson et al. 2016). Meanwhile, research communities are viewing the inherent ambiguity of taxonomic names in networked environments as a perpetual Achilles' heel for aggregation – our weakest semantic link in the entire system (Por 2007, Bortolus 2008, Peet et al. 2012, Hjarding et al. 2015, Maldonado et al. 2015, Franklin et al. 2016, Wiser 2016).

This project is designed to advance a global agenda, by demonstrating that conceptual and technical challenges can be addressed at scale if communities are willing to engage in concept taxonomy. Trust in data is also a design feature of allowing expert access and resolving multiple conflicting views that reflect the realities of ongoing taxonomic research. If only 1% of SERNEC's data display the issues shown in Fig. 2, trust in all data suffers. We believe that better science will follow from our demonstrations, first reaching other Symbiota portals and eventually large-scale aggregators. In light of vast investments into vouchered data mobilization, our project is a highly cost-effective way to improve data quality and trust. Taking well over 1 million SERNEC records from name- to TCL-identification will show that "big" specimen data can pass the credibility threshold needed to validate the mobilization investment.

II. Educational. We will directly train one postdoctoral researcher (UNC), two Ph.D. students (ASU, UIUC), and at least 15 undergraduate students (ASU). Each of our workshops will capacitate 10-15 SERNEC experts, who in turn can inform and recruit colleagues and students at their home herbaria. Project members Alford, Fishbein, Franck, Franz, Gilbert, Murell, Soltis, and Weakley regularly offer plant/biodiversity courses to undergraduate students at their respective institutions, reaching an estimated **300-500 students per year**, with ca. 10-40% minority students (range: Oklahoma – Mississippi). Each has committed to integrating our project's theme and use cases as new sections into their future biodiversity teaching plans. At ASU, this will include two new three-hour sections of the undergraduate-focused biodiversity informatics course "Discovering Biodiversity – Field to Database", offered in the spring of 2017 and 2019 to 25 students. At each institution, project members will make a sustained, systematic effort to recruit new students from underrepresented groups, working through institutional (e.g., sponsored STEM minority mentor programs) and local student organizations to advertise project opportunities and thereby proactively broaden participation.

Murrell's leadership of SERNEC will promote our advances with nearly 200 herbarium scientists in the region. Alford's involvement in the Magnolia grandiFLORA project (http://www.mississippiplants.org/), which has an educational component for K–12 teachers, will add exposure. At ASU, Franz and Gilbert will promote the project through virtual and personal outreach, aided by their leadership of the Biodiversity Knowledge Integration Center (BioKIC; https://biokic.asu.edu/). We will publish a BioKIC monthly blog post with project updates, to be shared with the iDigBio/Symbiota Working Group. Conference presentations will mainly target the global TDWG community (http://www.tdwg.org/).

We budget funds for two additional forms of outreach. The first will be a **feature story** "Where do plant species occur?" (see Fig. 2) – with accessible web content, learning activities, and videos – for ASU's popular "Ask A Biologist" learning resource, visited 22,000 times daily (https://askabiologist.asu.edu/). The second is a comprehensive series of "**How-To" videos** for creating multi-taxonomy alignments. Connecting the primary taxonomic literature with the RCC–5 logic is not trivial. Applying the concepts and tools is a bottleneck for further adoption. We will produce 10-15 well-structured and -narrated videos that exemplify steps the entire project workflow, including efficient use of the Euler/X toolkit. The video projects will be presented by ASU undergraduate students, with input from project experts. The series will be published and promoted through iDigBio's Vimeo channel (https://vimeo.com/idigbio).

11. Sustainability

Mid-term prospects (~ 5-15 years) for our development and data innovations are very strong. Our project operates inside an upward-trending information culture (Gries et al. 2014, Jones et al. 2014, Nelson et al. 2015). The SERNEC community is 233 herbaria strong, active since 2005, and recently organized as the SouthEast Chapter of the Society of Herbarium Curators, creating a region-wide research/training/proposal engine. In the past year the portal was accessed ca. 86x daily by a total of 13,410 users (Google

Analytics). SEINet's metrics (http://swbiodiversity.org/seinet/) are 10x higher – a realistic SERNEC target. The data, portal, and software are sustained by standing commitments from SERNEC, ASU, and iDigBio (University of Florida). Odds of sustaining the concept taxonomy innovations are less clear; thus we prioritize community engagement and use case-driven development.

Results from prior NSF support

Details are not provided here; however, the following NSF-funded projects were reviewed (intellectual merit, broader impacts) for each Co-/PI. This information is publicly available through the NSF website (links provided here).

1. PI Franz. NSF DEB-1155984. CAREER: Systematics of eustyline and geonemine weevils: connecting and contrasting Caribbean and Neotropical mainland radiations. http://nsf.gov/awardsearch/showAward?AWD_ID=1155984
2. Co-PI Gilbert. NSF-DBI 0743827. Symbiota, a virtual flora model for the Southwestern United States. http://www.nsf.gov/awardsearch/showAward?AWD_ID=0743827
3. Co-PI Ludäscher. NSF IIS-1118088. III: Small: A logic-based, provenance-aware system for merging scientific data under context and classification constraints. http://nsf.gov/awardsearch/showAward?AWD_ID=1118088
4. Co-PI Weakley. NSF-DBI 1410439. Digitization TCN: Collaborative Research: The key to the cabinets: building and sustaining a research database for a global biodiversity hotspot. http://www.nsf.gov/awardsearch/showAward?AWD_ID=1410439

Acknowledgements

The authors acknowledge critical proposal and input by Mac Alford, Mark Fishbein, Alan Franck, Michael Lee, Zack Murrell, and Pamela Soltis. This acknowledgment need not imply that the collaborators are responsible for the overall goals and content of the proposal.

Funding program

National Science Foundation (U.S.A.): Division of Biological Infrastructure: Advances in Biological Informatics.

Grant title

Collaborative Research: ABI Development: Controlling the taxonomic variable: Taxonomic concept resolution for a southeastern herbarium portal.

Call

URL: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5444

Hosting institution

Arizona State University, with the University of Illinois Urbana-Champaign and the University of North Carolina collaborating.

Ethics and security

None apparent.

Author contributions

NMF had primary content-related and organizational responsibilities, however all authors contributed (variously) to most proposal aspects. ASW contributed the RCC–5 data upon which much of the concept alignment objectives are based.

Conflicts of interest

None apparent.

References

- Ames O (1922) Orchidaceae: Illustrations and Studies of the Family Orchidaceae, Issuing from the Ames Botanical Laboratory, Fascile 7. Ames Botanical Laboratory, North Easton, Massachusetts, 174 pp. URL: http://dx.doi.org/10.5962/bhl.title.15433 DOI: 10.5962/bhl.title.15433
- Azuma H, García-Franco JG, Rico-Gray V, Thien LB (2001) Molecular phylogeny of the Magnoliaceae: the biogeography of tropical and temperate disjunctions. American Journal of Botany 88 (12): 2275‑2285. DOI: 10.2307/3558389
- Azuma H, Figlar RB, Tredici Pd, Camelbeke K, Palmarola-Bejerano A, Romanov MS (2011) Intraspecific sequence variation of cpDNA shows two distinct groups within

Magnolia virginiana L. of Eastern North America and Cuba. Castanea 76 (1): 118-123. DOI: 10.2179/10-018.1

- Berendsohn WG (1995) The concept of "potential taxa" in databases. Taxon 44 (2): 207-212. DOI: 10.2307/1222443
- Berendsohn WG, Geoffroy M (2007) Networking taxonomic concepts – uniting without 'unitary-ism'; pp. 13-22. Biodiversity Databases: from Cottage Industry to Industrial Networks. Curry, G., Humphries, C. (Eds.). Systematics Association Special Volume 73. Taylor & Francis, Boca Raton. URL: http://dx.doi.org/10.1201/9781439832547.ch3 DOI: 10.1201/9781439832547.ch3
- Borsch T, Hernández-Ledesma P, Berendsohn WG, Flores-Olvera H, Ochoterena H, Zuloaga FO, Mering Sv, Kilian N (2015) An integrative and dynamic approach for monographing species-rich plant groups – building the global synthesis of the angiosperm order Caryophyllales. Perspectives in Plant Ecology, Evolution and Systematics 17 (4): 284-300. DOI: 10.1016/j.ppees.2015.05.003
- Bortolus A (2008) Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. AMBIO: A Journal of the Human Environment 37 (2): 114-118. DOI: 10.1579/0044-7447(2008)37[114:ecitbs]2.0.co;2
- Bowers S, McPhillips TM, Ludäscher B (2008) Provenance in collection-oriented scientific workflows. Concurrency and Computation: Practice and Experience 20 (5): 519-529. DOI: 10.1002/cpe.1226
- Boyle B, Hopkins N, Lu Z, Raygoza Garay JA, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, Mckay SJ, Lowry S, Freeland C, Peet RK, Enquist BJ (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14 (1): 16. DOI: 10.1186/1471-2105-14-16
- Brown R (1813) Pogonia, p. 203. In: Aiton WT (Ed.) *Hortus Kewensis*; or, a Catalogue of the Plants Cultivated in the Royal Botanic Garden at Kew, Second Edition, Volume 5. Longman, Hurst, Rees, Orme, and Brown, London, 568 pp. URL: http://dx.doi.org/10.5962/bhl.title.105339 DOI: 10.5962/bhl.title.105339
- Campbell CS (1983) Systematics of the Andropogon virginicus complex (Gramineae). Journal of the Arnold Arboretum 64 (2): 171-254. DOI: 10.5962/bhl.part.27406
- Cheney J, Chiticariu L, Tan W- (2007) Provenance in databases: why, how, and where. Foundations and Trends in Databases 1 (4): 379-474. DOI: 10.1561/1900000006
- Chen M, Yu S, Franz N, Bowers S, Ludäscher B (2014a) A hybrid diagnosis approach combining Black-Box and White-Box reasoning; 127-141. In: Bikakis A, Fodor P, Roman D (Eds) Rules on the Web. From Theory to Applications. Proceedings of the 8th International Symposium, RuleML 2014, Co-located with the 21st European Conference on Artificial Intelligence, ECAI 2014, Prague, Czech Republic, August 18-20, 2014. Lecture Notes in Computer Science 8620. URL: http://dx.doi.org/10.1007/978-3-319-09870-8_9 DOI: 10.1007/978-3-319-09870-8_9
- Chen M, Yu S, Franz N, Bowers S, Ludäscher B (2014b) Euler/X: a toolkit for logic-based taxonomy integration; arXiv:1402.1992 [cs.LO]. arXiv 2014: 1-8. URL: http://arxiv.org/abs/1402.1992
- Chen M, Yu S, Kianmajd P, Franz N, Bowers S, Ludäscher B (2015) Provenance for explaining taxonomy alignments; pp. 258-260. In: Ludäscher B, Plale B (Eds) Provenance and Annotation of Data and Processes. Revised Selected Papers of the 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany,

June 9-13, 2014. Lecture Notes in Computer Science 8628. URL: http://dx.doi.org/10.1007/978-3-319-16462-5_27 DOI: 10.1007/978-3-319-16462-5_27

- Coulter JM, Rose JN (1900) Monograph of the North American Umbelliferae. Contributions of the United States National Herbarium 7 (1): 1-256. DOI: 10.5962/bhl.title.38223
- Crawford DJ, Mort ME (2005) Phylogeny of Eastern North American *Coreopsis* (Asteraceae-Coreopsideae): insights from nuclear and plastid sequences, and comments on character evolution. American Journal of Botany 92 (2): 330-336. DOI: 10.3732/ajb.92.2.330
- Dang TN, Franz NM, Ludäscher B, Forbes AG (2015) ProvenanceMatrix: a visualization tool for multi-taxonomy alignments. CEUR Workshop Proceedings 1456: 13-24. URL: http://ceur-ws.org/Vol-1456/paper2.pdf
- Döring M (2013) i4Life - Indexing for Life. D6.1. Integrated access to CoL in GBIF, Workpackage 6, pp. 1-4. Capacities Programme of Framework 7: EC e-Infrastructure Programme - Virtual Research Communities - INFRA-2010-2. 4 pp. URL: http://www.i4life.eu/i4lifewebsite/wp-content/uploads/2012/12/D6-1-Integrated-Access-to-CoL-in-GBIF_V4.pdf
- Döring M (2016a) GBIF Checklist Bank and the Backbone Taxonomy. http://www.slideshare.net/mdoering/checklist-bank-and-backbone. Accession date: 2016 8 15.
- Döring M (2016b) Updating the GBIF Backbone. http://gbif.blogspot.co.uk/2016/04/updating-gbif-backbone.html. Accession date: 2016 8 15.
- Dorsey BL, Haevermans T, Aubriot X, Morawetz JJ, Riina R, Steinmann VW, Berry PE (2013) Phylogenetics, morphological evolution, and classification of *Euphorbia* subgenus *Euphorbia* . Taxon 62 (2): 291-315. DOI: 10.12705/622.1
- Dubois A (2005) Proposed rules for the incorporation of nomina of higher-ranked zoological taxa in the International Code of Zoological Nomenclature. 1. Some general questions, concepts and terms of biological nomenclature. Zoosystema 27 (2): 365-426. URL: http://sciencepress.mnhn.fr/sites/default/files/articles/pdf/z2005n2a8.pdf
- Duncan WH (1979) Changes in *Galactia* (Fabaceae) of the southeastern United States. SIDA, Contributions to Botany 8 (2): 170-180. URL: http://www.jstor.org/stable/23909679
- Ernst M, Saslis-Lagoudakis CH, Grace OM, Nilsson N, Simonsen HT, Horn JW, Rønsted N (2016) Evolutionary prediction of medicinal properties in the genus *Euphorbia* L. Scientific Reports 6: 30531. DOI: 10.1038/srep30531
- Federhen S (2012) The NCBI Taxonomy database. Nucleic Acids Research 40: D136-D143. DOI: 10.1093/nar/gkr1178
- Fernald ML (1946a) Some orchids of the Manual range. Rhodora 48 (572): 184-197. URL: http://www.jstor.org/stable/23304761
- Fernald ML (1946b) Some orchids of the Manual range. Rodora 48 (571): 161-162. URL: http://www.jstor.org/stable/i23303797
- Fernald ML (1950) Gray's Manual of Botany, Eigth (Centennial) Edition. American Book Company, New York, 1632 pp.
- Figlar R, Nooteboom HP (2004) Notes on Magnoliaceae IV. Blumea - Biodiversity, Evolution and Biogeography of Plants 49 (1): 87-100. DOI: 10.3767/000651904x486214
- Flanagan NS, Peakall R, Clements MA, Otero JT (2006) Conservation of taxonomically difficult species: the case of the Australian orchid, *Microtis angusii* . Conservation Genetics 7 (6): 847-859. DOI: 10.1007/s10592-006-9119-8

- FNA (2015) Flora of North America Editorial Committee, eds. 1993+. Flora of North America North of Mexico. 16+ vols. New York and Oxford. Vol. 1, 1993; vol. 2, 1993; vol. 3, 1997; vol. 4, 2003; vol. 5, 2005; vol. 7, 2010; vol. 8, 2009; vol. 19, 2006; vol. 20, 2006; vol. 21, 2006; vol. 22, 2000; vol. 23, 2002; vol. 24, 2007; vol. 25, 2003; vol. 26, 2002; vol. 27, 2007; vol 28, 2014; vol. 9, 2014; vol. 6, 2015. http://floranorthamerica.org. Accession date: 2016 8 15.
- Franklin J, Serra-Diaz JM, Syphard AD, Regan HM (2016) Big data for forecasting the impacts of global change on plant communities. Global Ecology and Biogeography 2016: 1-12. DOI: 10.1111/geb.12501
- Franz NM, Peet RK (2009) Towards a language for mapping relationships among taxonomic concepts. Systematics and Biodiversity 7 (1): 5-20. DOI: 10.1017/s147720000800282x
- Franz NM, Sterner BS (2015) Taxonomy - for computers. biorXiv 2015: 1-19. DOI: 10.1101/022145
- Franz NM, Thau D (2010) Biological taxonomy and ontology development: scope and limitations. Biodiversity Informatics 7 (1): 45-66. DOI: 10.17161/bi.v7i1.3927
- Franz NM, Peet RK, Weakley AS (2008) On the use of taxonomic concepts in support of biodiversity research and taxonomy, pp. 63-86. In: Wheeler QD (Ed.) The New Taxonomy, Systematics Association Special Volume, Series 74. Taylor & Francis, Boca Raton. URL: http://dx.doi.org/10.1201/9781420008562.ch5 DOI: 10.1201/9781420008562.ch5
- Franz NM, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B (2015) Reasoning over taxonomic change: exploring alignments for the *Perelleschus* use case. PLoS ONE 10 (2): e0118247. DOI: 10.1371/journal.pone.0118247
- Franz NM, Chen M, Kianmajd P, Yu S, Bowers S, Weakley AS, Ludäscher B (2016a) Names are not good enough: reasoning over taxonomic change in the *Andropogon* complex. Semantic Web (IOS) 7: 1-23. DOI: 10.3233/SW-160220
- Franz NM, Pier NM, Reeder DM, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B (2016b) Two influential primate classifications logically aligned. Systematic Biology 65 (4): 561-582. DOI: 10.1093/sysbio/syw023
- Gansner ER, North SC (2000) An open graph visualization system and its applications to software engineering. Software: Practice and Experience 30 (11): 1203-1233. DOI: 10.1002/1097-024x(200009)30:113.3.co;2-e
- GBIF (2010) Global Biodiversity Information Facility. Darwin Core Archives – How-To Guide, Version 1, Released on 1 March 2011 (contributed by Remsen, D., K. Braak, M. Döring & T. Robertson), Copenhagen. http://links.gbif.org/gbif_dwca_how_to_guide_v1. Accession date: 2016 8 15.
- Geoffroy M, Berendsohn WG (2003) The concept problem in taxonomy: importance, components, approaches. Schriftenreihe für Vegetationskunde 39: 4-15. URL: http://www.nhbs.com/series/64263/schriftenreihe-fur-vegetationskunde
- Geoffroy M, Güntsch A (2003) Assembling and navigating the potential taxon graph. Schriftenreihe für Vegetationskunde 39: 71-82. URL: http://www.nhbs.com/series/64263/schriftenreihe-fur-vegetationskunde
- Graham M, Kennedy J (2010) A survey of multiple tree visualisation. Information Visualization 9 (4): 235-252. DOI: 10.1057/ivs.2009.29

- Gregg KB, Kéry M (2006) Comparison of size vs. life-state classification in demographic models for the terrestrial orchid *Cleistes bifaria* . Biological Conservation 129 (1): 50-58. DOI: 10.1016/j.biocon.2005.09.044
- Gries C, Gilbert EE, Franz NM (2014) Symbiota – a virtual platform for creating voucher-based biodiversity information communities. Biodiversity Data Journal 2: e1114. DOI: 10.3897/bdj.2.e1114
- Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz L, Laughinghouse HD, McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences 112 (41): 12764-12769. DOI: 10.1073/pnas.1423041112
- Hjarding A, Tolley KA, Burgess ND (2015) Red List assessments of East African chameleons: a case study of why we need experts. Oryx 49 (4): 652-658. DOI: 10.1017/s0030605313001427
- Isely D (1998) Native and Naturalized Leguminosae (Fabaceae) of the United States (Exclusive of Alaska and Hawaii). Monte L. Bean Life Science Museum, Brigham Young University, Provo, Utah, 1007 pp. [ISBN 978-0842523967]
- Jansen MA, Franz NM (2015) Phylogenetic revision of *Minyomerus* Horn, 1876 sec. Jansen & Franz, 2015 (Coleoptera, Curculionidae) using taxonomic concept annotations and alignments. ZooKeys 528: 1-133. DOI: 10.3897/zookeys.528.6001
- Jones AC, White RJ, Orme ER (2011) Identifying and relating biological concepts in the Catalogue of Life. Journal of Biomedical Semantics 2 (1): 7. DOI: 10.1186/2041-1480-2-7
- Jones TM, Baxter DG, Hagedorn G, Legler B, Gilbert E, Thiele K, Vargas-Rodriguez Y, Urbatsch LE (2014) Trends in access of plant biodiversity data revealed by Google Analytics. Biodiversity Data Journal 2: e1558. DOI: 10.3897/bdj.2.e1558
- Kartesz J (2010) Floristic Synthesis of North America, Version 9-15-2010. Biota of North America Program (BONAP), Chapel Hill. http://www.bonap.org/. Accession date: 2016 8 15.
- Kelly LM (1998) Phylogenetic relationships in *Asarum* (Aristolochiaceae) based on morphology and ITS sequences. American Journal of Botany 85 (10): 1454-1467. DOI: 10.2307/2446402
- Kennedy JB, Kukla R, Paterson T (2005) Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration, pp. 80-95. In: Ludäscher B, Raschid L (Eds) Data Integration in the Life Sciences: Proceedings of the Second International Workshop, San Diego, DILS 2005, LNBI 3615. URL: http://dx.doi.org/10.1007/11530084_8 DOI: 10.1007/11530084_8
- Kim S-, Crawford DJ, Tadesse M, Berbee M, Ganders FR, Pirseyedi M, Esselman EJ (1999) ITS sequences and phylogenetic relationships in *Bidens* and *Coreopsis* (Asteraceae). Systematic Botany 24 (3): 480. DOI: 10.2307/2419701
- Klooster MR, Culley TM (2009) Comparative analysis of the reproductive ecology of *Monotropa* and *Monotropsis*: two mycoheterotrophic genera in the Monotropoideae (Ericaceae). American Journal of Botany 96 (7): 1337-1347. DOI: 10.3732/ajb.0800319
- Koperski M, Sauer M, Braun W, Gradstein SR (2000) Referenzliste der Moose Deutschlands. Schriftenreihe für Vegetationskunde 35: 1-519. URL: https://

www.amazon.de/Referenzliste-Moose-Deutschlands-Schriftenreihe-Vegetationskunde/dp/3784335047

- Krings A (2008) Synopsis of *Gonolobus s. l.* (Apocynaceae, Asclepiadoideae) in the United States and its territories, including lectotypification of *Lachnostoma arizonicum* . Harvard Papers in Botany 13 (2): 209-218. DOI: 10.3100/1043-4534-13.2.209
- Lee DJ, Adams D, Kim CS (2009) Managing invasive plants on public conservation forestlands: application of a bio-economic model. Forest Policy and Economics 11 (4): 237-243. DOI: 10.1016/j.forpol.2009.03.004
- Leonelli S (2013) Classificatory theory in biology. Biological Theory 7 (4): 338-345. DOI: 10.1007/s13752-012-0049-z
- Lepage D, Vaidya G, Guralnick R (2014) Avibase – a database system for managing and organizing taxonomic concepts. ZooKeys 420: 117-135. DOI: 10.3897/zookeys.420.7089
- Linnaeus C (1753) Species plantarum, exhibentes plantas rite cognitas, ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas. Tomus 1. Laurentii Salvii, Holmiae, 560 pp. URL: http://dx.doi.org/10.5962/bhl.title.669 DOI: 10.5962/bhl.title.669
- Linnaeus C (1758) Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis, Editio 10. Laurentii Salvii, Holmiae, 824 pp. URL: http://dx.doi.org/10.5962/bhl.title.542 DOI: 10.5962/bhl.title.542
- Lipow SR, Wyatt R (1998) Reproductive biology and breeding system of *Gonolobus suberosus* (Asclepiadaceae). Journal of the Torrey Botanical Society 125 (3): 183-193. DOI: 10.2307/2997216
- Ludäscher B, Chen M, Yu S, Kianmajd P, Bowers S, Franz N, McPhillips T (2016) Euler Project – Reasoning Over Taxonomies. Release date: 2016 8 01. URL: https://github.com/EulerProject
- Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Global Ecology and Biogeography 24 (8): 973-984. DOI: 10.1111/geb.12326
- Mayfield MH (2002) The varieties of *Liatris elegans* (Asteraceae). SIDA, Contributions to Botany 20 (2): 597-603. URL: http://www.jstor.org/stable/41968077
- Midford PE, Dececchi TA, Balhoff JP, Dahdul WM, Ibrahim N, Lapp H, Lundberg JG, Mabee PM, Sereno PC, Westerfield M, Vision TJ, Blackburn DC (2013) The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes. Journal of Biomedical Semantics 4 (1): 34. DOI: 10.1186/2041-1480-4-34
- Murrell ZE (1993) Phylogenetic relationships in *Cornus* (Cornaceae). Systematic Botany 18 (3): 469. DOI: 10.2307/2419420
- Murrell ZE, Carroll PE, Myers SA, Lawless PJ (1998) Examination of species boundaries in *Hexastylis contracta* Blomquist and *H. rhombiformis* Gaddy. (Abstract.). American Journal of Botany (Supplement) 85: 146-147.
- Nelson G, Sweeney P, Wallace LE, Rabeler RK, Allard D, Brown H, Carter RJ, Denslow MW, Ellwood ER, Germain-Aubrey CC, Gilbert E, Gillespie E, Goertzen LR, Legler B, Marchant DB, Marsico TD, Morris AB, Murrell Z, Nazaire M, Neefus C, Oberreiter S, Paul D, Ruhfel BR, Sasek T, Shaw J, Soltis PS, Watson K, Weeks A, Mast AR (2015)

Digitization workflows for flat sheets and packets of plants, algae, and fungi. Applications in Plant Sciences 3 (9): 1500065. DOI: 10.3732/apps.1500065

- Ogden CK, Richards IA (1923) The Meaning of Meaning: a Study of the Influence of Language upon Thought and of the Science of Symbolism. 1st Edition. Kegan Paul, London, 396 pp. [ISBN 978-0156584463]
- O'Malley MA (2013) When integration fails: prokaryote phylogeny and the tree of life. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 44 (4): 551-562. DOI: 10.1016/j.shpsc.2012.10.003
- Page RD, Valiente G (2005) An Edit script for taxonomic classifications. BMC Bioinformatics 6 (1): 208. DOI: 10.1186/1471-2105-6-208
- Pansarin ER, Barros Fd (2008) Taxonomic notes on Pogonieae (Orchidaceae): *Cleistesiopsis*, a new genus segregated from *Cleistes*, and description of two new South American species, *Cleistes batistana* and *C. elongata* . Kew Bulletin 63 (3): 441-448. DOI: 10.1007/s12225-008-9047-5
- Pansarin RP, Brown PM (2009) A new genus for the North American *Cleistes* . North American Native Orchid Journal 15 (1): 50-58. URL: https://www.scribd.com/document/42865188/March-2009-North-American-Native-Orchid-Journal
- Park K (1998) Monograph of *Euphorbia* sect. *Tithymalopsis* (Euphorbiaceae). Edinburgh Journal of Botany 55 (2): 161-208. DOI: 10.1017/s0960428600002122
- Patterson D, Mozzherin D, Shorthouse DP, Thessen A (2016) Challenges with using names to link digital biodiversity information. Biodiversity Data Journal 4: e8080. DOI: 10.3897/bdj.4.e8080
- Peet RK, Lee MT, Jennings MD, Faber-Langendoen D (2012) VegBank – a permanent, open-access archive for vegetation-plot data. Biodiversity & Ecology 4: 233-241. DOI: 10.7809/b-e.00080
- Peirce CS (1998) The Essential Peirce, Volume 2. Peirce Edition Project (Editor). Indiana University Press, Bloomington, 624 pp. URL: http://www.iupress.indiana.edu/product_info.php?products_id=21333 [ISBN 978-0-253-21190-3]
- Peterson AT, Navarro-Sigüenza A (1999) Alternate species concepts as bases for determining priority conservation areas. Conservation Biology 13 (2): 427-431. DOI: 10.1046/j.1523-1739.1999.013002427.x
- Por FD (2007) A "taxonomic affidavit": why it is needed? Integrative Zoology 2 (2): 57-59. DOI: 10.1111/j.1749-4877.2007.00044.x
- Radford AE, Ahles HE, Bell CR (1968) Manual of the Vascular Flora of the Carolinas. University of North Carolina Press, Chapel Hill, 1245 pp. URL: http://uncpress.unc.edu/books/T-766.html [ISBN 978-0-8078-1087-3]
- Randell DA, Cui Z, Cohn AG (1992) A spatial logic based on regions and connection, pp. 165-176. In: Nebel B, Swartout W, Rich C (Eds) Proceedings of the Third International Conference on the Principles of Knowledge Representation and Reasoning. Morgan Kaufmann, Los Altos.
- Remsen D (2016) The use and limits of scientific names in biological informatics. ZooKeys 550: 207-223. DOI: 10.3897/zookeys.550.9546
- Rose JP, Freudenstein JV (2014) Cryptic and overlooked: species delimitation in the mycoheterotrophic *Monotropsis* (Ericaceae: Monotropoideae). Systematic Botany 39 (2): 578-593. DOI: 10.1600/036364414x680762

- Rosenberg MS (2014) Contextual cross-referencing of species names for fiddler crabs (genus *Uca*): an experiment in cyber-taxonomy. PLoS ONE 9 (7): e101704. DOI: 10.1371/journal.pone.0101704
- Rylands AB, Mittermeier RA (2014) Primate taxonomy: species and conservation. Evolutionary Anthropology: Issues, News, and Reviews 23 (1): 8-10. DOI: 10.1002/evan.21387
- Sanders RW (2006) Taxonomy of *Lantana* sect. *Lantana* (Verbenaceae): I. Correct application of *Lantana camara* and associated names. SIDA, Contributions to Botany 22 (1): 381-421. URL: http://www.jstor.org/stable/41968588
- Sanders RW (2012) Taxonomy of *Lantana* sect. *Lantana* (Verbenaceae): II. Taxonomic revision. Journal of the Botanical Research Institute of Texas 6 (2): 403-441. URL: http://www.jstor.org/stable/41972430
- Schilling EE, Floden A, Schilling DE (2015) Barcoding the Asteraceae of Tennessee, tribe Cichorieae. Phytoneuron 2019 (19): 1-8. URL: http://www.phytoneuron.net/2015Phytoneuron/19PhytoN-TennCichorieae.pdf
- Scoble MJ (2004) Unitary or unified taxonomy? Philosophical Transactions of the Royal Society B: Biological Sciences 359 (1444): 699-710. DOI: 10.1098/rstb.2003.1456
- SERNEC (2016) SouthEast Regional Network of Expertise and Collections, On-Line Portal; Community website available at http://sernec.appstate.edu. http://sernecportal.org. Accession date: 2016 8 15.
- Sima Y-, Lu S- (2012) A new system for the family Magnoliaceae; pp. 55-71. Proceedings of the Second International Symposium on the Family Magnoliaceae; Nianhe, X., Qingwen, Z., Fengxia, X., Qigen, W. (Eds.), Xia Nianhe, Zeng Qingwen, Xu Fengxia, Wu Qigen. 304 pp. [ISBN 9787560973494].
- Small JK (1933) Manual of the Southeastern Flora: Being Descriptions of the Seed Plants Growing Naturally in Florida, Alabama, Mississippi, Eastern Louisiana, Tennessee, North Carolina, South Carolina and Georgia. J.K. Small, New York, 1554 pp. URL: http://dx.doi.org/10.5962/bhl.title.696 DOI: 10.5962/bhl.title.696
- Smith SD, Cowan RS, Gregg KB, Chase MW, Maxted N, Fay MF (2004) Genetic discontinuities among populations of *Cleistes* (Orchidaceae, Vanilloideae) in North America. Botanical Journal of the Linnean Society 145 (1): 87-95. DOI: 10.1111/j.1095-8339.2003.00265.x
- Sorrie BA, LeBlond RJ, Weakley AS (2013) Identification, distribution, and habitat of *Coreopsis* section *Eublepharis* (Asteraceae) and description of a new species. Journal of the Botanical Research Institute of Texas 7 (1): 299-310. DOI: 10.5962/bhl.title.63883
- Stevens PF (2016) Angiosperm Phylogeny Website. Version 12, July 2012 [and more or less continuously updated since]. http://www.mobot.org/MOBOT/research/APweb. Accession date: 2016 8 15.
- Thau D, Ludäscher B (2007) Reasoning about taxonomies in first-order logic. Ecological Informatics 2 (3): 195-209. DOI: 10.1016/j.ecoinf.2007.07.005
- Thau D, Bowers S, Ludascher B (2010) Towards best-effort merge of taxonomically organized data; pp. 151-154. 2010 IEEE 26th International Conference of Data Engineering Workshops (ICDEW). 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010), 344 pp. URL: http://dx.doi.org/10.1109/icdew.2010.5452756 DOI: 10.1109/icdew.2010.5452756
- Thau D, Bowers S, Ludäscher B (2008) Merging taxonomies under RCC-5 algebraic articulations, pp. 47-54. Conference on Information and Knowledge Management.

Proceeding of the 2nd International Workshop on Ontologies and Information systems for the Semantic Web (ONISW), Napa Valley, California. ACM, New York, 118 pp. URL: http://dx.doi.org/10.1145/1458484.1458492 DOI: 10.1145/1458484.1458492

- Thau D, Bowers S, Ludäscher B (2009) Merging sets of taxonomically organized data using concept mappings under uncertainty; pp. 1103-1120. Proceedings of the 8th International Conference on Ontologies, Databases, and the Applications of Semantics (ODBASE 2009). OTM 2009. Lecture Notes in Computer Science 5871. URL: http://dx.doi.org/10.1007/978-3-642-05151-7_26 DOI: 10.1007/978-3-642-05151-7_26
- Thau DM (2010) Reasoning About Taxonomies. Doctoral Dissertation. University of California at Davis, 204 pp. [ISBN 978-1-124-22390-2].
- Vink CJ, Paquin P, Cruickshank RH (2012) Taxonomy and irreproducible biological science. BioScience 62 (5): 451-452. DOI: 10.1525/bio.2012.62.5.3
- Ward DB, Hall DW (2004) Keys to the flora of Florida – 10, *Galactia* (Leguminosae). Phytologia 86 (2): 65-74. URL: http://www.biodiversitylibrary.org/bibliography/12678
- Weakley AS (2009) A practioner's guide to concept mapping. Narrative for "Least Divisible Taxonomic Unit" approach to "concept mapping". http://sernec.appstate.edu/sernec-working-groups/how-map-taxonomic-concepts-alan-weakley. Accession date: 2016 8 15.
- Weakley AS (2014) Applying concept mapping to 7,100 vascular plants. Presentation given at BIGCB Workshop, University of California at Berkeley, November 7-9, 2014. http://taxonbytes.org/wp-content/uploads/2014/10/Weakley-2014-BIGCB-Applying-Concept-Mapping-to-7100-Vascular-Plants.pdf. Accession date: 2016 8 15.
- Weakley AS (2015) Flora of the Southern and Mid-Atlantic States. University of North Carolina Herbarium, Chapel Hill, 1320 pp. URL: http://www.herbarium.unc.edu/flora.htm
- Weakley AS, LeBlond RJ, Sorrie BA, Witsell CT, Estes LD, Mathews KG, Ebihara A, Gandhi K (2011) New combinations, rank changes, and nomenclatural and taxonomic comments in the vascular flora of the southeastern United States. Journal of the Botanical Research Institute of Texas 5 (2): 437-455. URL: http://www.jstor.org/stable/41972288
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE 7 (1): e29715. DOI: 10.1371/journal.pone.0029715
- Winsberg E, Huebner B, Kukla R (2014) Accountability and values in radically collaborative research. Studies in History and Philosophy of Science Part A 46: 16-23. DOI: 10.1016/j.shpsa.2013.11.007
- Wiser SK (2016) Achievements and challenges in the integration, reuse and synthesis of vegetation plot data. Journal of Vegetation Science 27: 868-879. DOI: 10.1111/jvs.12419
- Witteveen J (2015) Naming and contingency: the type method of biological taxonomy. Biology & Philosophy 30 (4): 569-586. DOI: 10.1007/s10539-014-9459-6
- Witteveen J (2016) Suppressing synonymy with a homonym: the emergence of the nomenclatural type concept in nineteenth century natural history. Journal of the History of Biology 49 (1): 135-189. DOI: 10.1007/s10739-015-9410-y
- Wunderlin RP, Hansen BF (2011) Guide to the Vascular Plants of Florida, Third Edition. University Press of Florida, Gainesville, 800 pp. URL: http://upf.com/book.asp?id=wunde004# [ISBN 978-0-8130-3543-7]

- Xiang Q-, Thomas DT, Zhang W, Manchester SR, Murrell Z (2006) Species level phylogeny of the genus *Cornus* (Cornaceae) based on molecular and morphological evidence - implications for taxonomy and Tertiary intercontinental migration. Taxon 55 (1): 9-30. DOI: 10.2307/25065525