Combining Machine Learning & Reasoning for Biodiversity Data Intelligence

Atriya Sen,[1] Beckett Sterner,[2] Nico Franz,[2] Caleb Powel,[2] Nathan Upham[2]
[1]University of New Orleans
[2]Arizona State University
atriya@atriyasen.com
beckett.sterner@asu.edu, nico.franz@asu.edu, cpowel21@asu.edu, nathan.upham@asu.edu

Abstract

The current crisis in global natural resource management makes it imperative that we better leverage the vast data sources associated with taxonomic entities (such as recognized species of plants and animals), which are known collectively as biodiversity data. However, these data pose considerable challenges for artificial intelligence: while growing rapidly in volume, they remain highly incomplete for many taxonomic groups, often show conflicting signals from different sources, and are multi-modal and therefore constantly changing in structure. In this paper, we motivate, describe, and present a novel workflow combining machine learning and automated reasoning, to discover patterns of taxonomic identity and change – i.e. "taxonomic intelligence" – leading to scalable and broadly impactful AI solutions within the biodata realm.

Introduction

Many of the challenges society faces are being addressed through interdisciplinary collaboration supported by interoperable information resources. A common expectation is that scientists will be able to agree on a shared, stable vocabulary, but consensus classifications are frequently absent in the biodiversity domain and in fact may not be necessary (Sterner, Witteveen, and Franz 2020). Geographic distribution, genetic, and phenotypic traits, as well as ecological interactions of biological entities are collectively known as *biodiversity data*. The recognized species names (i.e., taxonomic names) as well as the associated criteria that circumscribe them (i.e., taxonomic concepts) are the standard methods of classifying these data (Franz and Peet 2009). Biodiversity data are integral components for addressing many contemporary challenges society faces, including: natural resource management, climate change modeling, biodiversity conservation, food security measures, and international treaty enforcement. Unfortunately, there is widespread and persistent variation in the ways that scientists represent biodiversity data, e.g. using hierarchical taxonomies, phylogenies reconstructing evolutionary history, and computer ontologies of terms and their relationships (Franz and Sterner 2018; Vaidya, LePage, and Guralnick 2018; Franz, Musher et al. 2019; Hardisty, Michener et al. 2019; Jeliazkov, Mijatovic et al. 2020). This variation poses major obstacles to robust inter-operability across an increasingly decentralized (Blagoderov et al. 2012) ecosystem of primary biodiversity data sources.

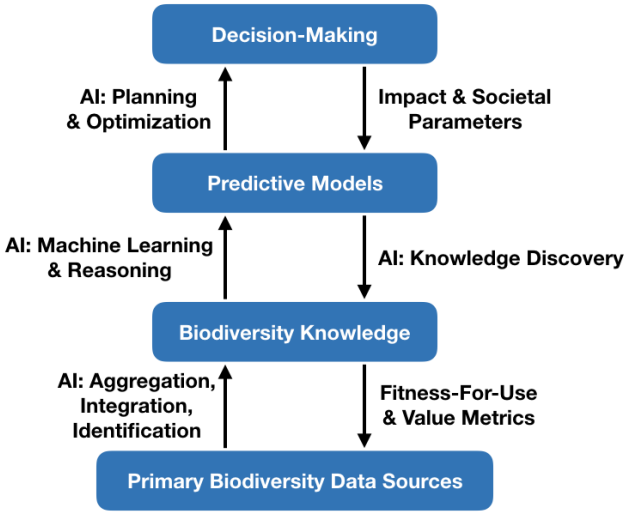

Figure 1: Biodiversity AI "Landscape"

Realizing the societal value of biodiversity data science will require data integration techniques that are aware of the context sensitive variations associated with conflicting taxonomic concepts. Accurate and scalable data intelligence – logically relating data classification terms to their intended yet often implicit meanings – is essential to overcoming these data integration challenges. Artificial intelligence (AI) has an essential role to play in generating the knowledge needed for generating and translating information across conflicting biodiversity data classifications.

Figure 1 illustrates the many roles that AI must play in this domain.

With biodiversity data rapidly expanding in volume, so too is the scope and frequency of its use in research (Nelson and Ellis 2019) causing increasing demand for aggregated, well integrated, inter-operable biodiversity data. In meeting this demand, biodiversity science is ripe for an AI-fueled

transformation.

We address the role of AI methods in leveraging the vast bodies of data that describe historically or currently recognized species (i.e., taxonomic concepts of those species). These include data from "citizen science repositories" such as iNaturalist [1], where enthusiast communities upload images and notes of observations "in the field". Contemporary advances in machine learning, particularly in the field of *deep neural networks*, have recently demonstrated great promise in an important *single aspect* of the general problem of leveraging these data, i.e., the identification of already-recognized species as a *supervised learning* problem, from image data or sound recordings (Wäldchen and Mäder 2018a).

These advancements have not been leveraged to bear upon the broader and impactful problems of biodiversity data aggregation: the assimilation of taxonomic change, monitoring of habitat change, and extinction risk determination – to name just a few – despite the immense and urgent need to do so. Identifying only species concepts presently recognized by a static authority is simply not enough, and falls short of addressing the broader data aggregation challenge for biodiversity data science; indeed, it does not even address the problem of taxonomic translation across evolving perspectives. The pressing problem is how recent AI advances may be harnessed for the good of the planet and the (estimated) 3–10 million species that inhabit it, the vast majority of which are yet unidentified and hence not formally named and described.

Novel Contributions & Outline

In this paper, we introduce and analyze an *inversion* of the well-studied species identification problem. The status quo use of AI in biodiversity research presupposes a well-specified, static taxonomy to which certain members are assigned (Wäldchen and Mäder 2018b). However, advancements in taxonomy and systematics continuously redefine the criteria by which organisms are classified, with the degree to which those criteria are adopted often varying over geography. Therefore, the classes used in this domain are neither stable over time nor universally applied across geographic regions (Franz, Pier et al. 2016; Remsen 2016; Vaidya, LePage, and Guralnick 2018). The novel, inverse problem we introduce is thus to leverage data classified according to a particular taxonomic theory to *learn* its characteristics and its often tacit *alignments* to other taxonomic theories of the same set of species. Hence, instead of falsely taking classification as static background, we develop an AI algorithm to infer and integrate evolving and often conflicting taxonomic perspectives.

The novel contributions of this paper are: (1) the identification of a challenge, i.e. learning from biodiversity data *and* reasoning over existing biodiversity knowledge, to align biological taxonomies: this problem has not at all been considered previously in the AI literature, and, as we describe in Section , has immense potential for social impact; (2) the development of a novel algorithm for this challenge, which exhibits desirable performance on a representative dataset; and (3) the development of publicly available and open-sourced software to address the challenge.

Considered together, these contributions constitute an important advance, addressing multiple calls for an integrative technological approach that amplifies our best biodiversity knowledge to inform policy actions on global societal problems, such as tropical deforestation (Draper et al. 2020) and illegal international wildlife trade (Minin et al. 2019). Our approach is also novel from the perspective of *ontology alignment*, by combining a deep learning method for classifying individual data records, with automated reasoning, in a closed, iterative inference procedure (Angermann and Ramzan 2017).

We introduce publicly available herbarium plant images as a case-study, in Section . In Section , we describe a novel algorithm to discover articulations between two or more taxonomies. A (novel) architecture is proposed; it is summarized in Figure 3. In Section , we add a further layer of generalization, showing how these learnt insights, now leveraging *existing* expert knowledge about the relationships between different taxonomical theories, may be used to study, assimilate, and aggregate these different theories, using automated reasoning. In Section , we describe the broad potential social impact of this work. Section concludes with a discussion of limitations, and outlook. A brief ethics statement follows the references, in Section .

We focus in particular on taxonomic names and their meanings – or, simply – taxonomic intelligence (Franz and Peet 2009; Peterson, Soberón, and Krishtalka 2015; Franz, Pier et al. 2016), for the following reasons: taxonomic names provide a universal currency for exchanging information about empirically inferred, natural biological entities, such as species or larger phylogenetic groupings (Müller-Wille and Charmantier 2012). Our focus is therefore on the *foundational* step of moving from primary biodiversity data sources that are typically generated under multiple systematic classification schemes, to biodiversity *knowledge* ready for modeling applications (see bottom left of Figure 1). Taxonomic names are invariably reused across multiple revisions of biological classifications, where stability and change in name application (nomenclature) and meaning (taxonomy) follow semi-independent rules and domains of practice, leading to an increasingly complex network of many-to-many relationships between name usages and intended meanings. Although informative to some degree, the names of taxonomic groups alone are an insufficient basis for fit-for-purpose data aggregation to standards needed (e.g.) for conservation biology and precision risk management (Mesibov 2013; Guala 2016; Franz and Sterner 2018; Mesibov 2018). In traditional, smaller-scale biology, reliable resolution of this built-in uncertainty of name usages has been the burden of individual human agents (Sterner and Franz 2017).

A Note on the Supplementary Material: Code, Software, Experiments

We have made extensive resources available to supplement this paper. The web application described in Section is

[1] Available at https://www.inaturalist.org. Accessed Sep 1 2020.

provided, with complete instructions for use, and complete open-sourced code, which may be considered a snapshot of the state of the application at the time of submission. All data is publicly available. We note that the web application is immediately available for use. The Supplementary Material also includes some additional experiments described at the end of the next section, which were omitted here for lack of space.

Using Herbarium Image Data

In the particular domain of plant diversity, it has been estimated that more than 50% of unknown species are already represented in herbarium collections (Tan et al. 2019). In this context, we consider the Herbarium Challenge Dataset (Tan et al. 2019), which consists of "$46,469$ digitally imaged herbarium sheets representing 683 species from the flowering plant family Melastomataceae (commonly called melastomes)", from the New York Botanical Garden Herbarium Collections. In this dataset, some 63 genera and 683 species are represented, constituting a subset of the entire family's diversity, according to MELnames[2]. The dataset presents particular challenges for computer vision, including large intra-class variation relative to the small inter-class variation, and the necessity for fine-grained visual recognition. The locally variable process of generating herbarium sheets in collections also strongly alters the physical-visual properties of the imaged specimen, making it difficult to extract taxonomic signals of the same species from those caused by specific curatorial preparation practices.

The aforementioned difficulties originate precisely from the nature of taxonomic theorizing; biologists will often single out fine-grained and subtle visual characteristics, and propose major taxonomic variations on that basis, despite the thus-identified taxonomic concept referring to superficially disparate looking organisms, and despite the presence of superficially highly similar organisms that nevertheless pertain to distinct, non-overlapping taxonomic concepts.

Further, the Herbarium Challenge Dataset was intended for known-species identification – with the winning submission in the associated Kaggle competition achieving 89.8% classification accuracy, employing an ensemble of 5 neural network models, and state-of-the-art optimization techniques. However, the particular taxonomic theory to which the provided classification labels conform, is not evident from the data, or even mentioned in the paper. We were informed, through personal communications with the authors, that the taxonomy in fact corresponds roughly to that found in MELnames.

Some explorations of the feasibility of deep representation learning and feature clustering on herbarium data were performed using the Herbarium Challenge Dataset. For reasons of space, we exclude these experiments here, but describe them in the Supplementary Material.

[2]Available from http://www.melastomataceae.net/MELnames/. Accessed Sep 1 2020.

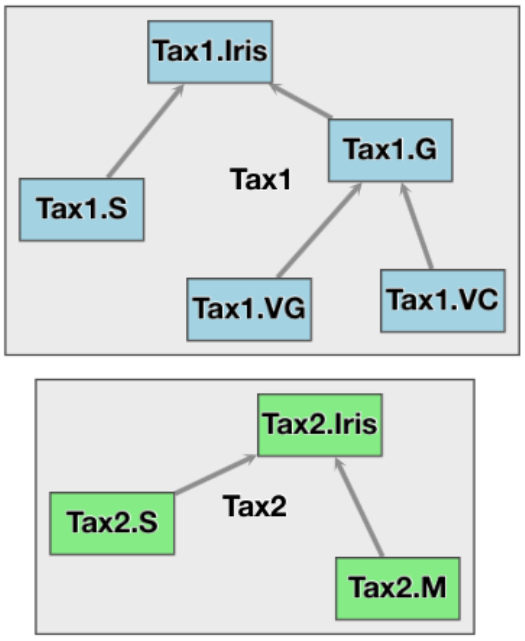

Figure 2: The *Iris* taxonomies $Tax1$ and $Tax2$. G refers to a higher level taxonomic concept that encompasses $Tax1.VC$ and $Tax1.VG$. In this case, it is the *Iris* series *Laevigatae*, a subgroup comprising several species.

Taxonomic Articulation Discovery

Please note that further technical details on the content of this section may be found in the Technical Appendix.

We consider three recognized species in the plant genus *Iris* Linnaeus; viz. *Iris setosa* (henceforth: S), *Iris versicolor* (henceforth: VC), and *Iris virginica* (henceforth: VG). These species are well-known in the AI community, as the constituents of Ronald Fisher's *Iris* dataset. Herbarium specimen images were obtained by querying the API of iDigBio[3]. Pre-processing consisted of converting the images to greyscale, and padding with black pixels on the right and bottom to obtain a square image of dimensions 768 by 768 in pixels. Before training, they were shuffled with a random seed. Standard data augmentation – horizontal and vertical flipping, rotation, and inward zooming – was performed during training.

We considered two illustrative taxonomies $Tax1$ and $Tax2$, with the structures shown in Figure 2. This is designed to simulate the important notion of taxonomic *splitting and merging*, known to be of critical importance in conservation science and policy (Jacobs and Baker 2018; Godfray, Knapp, and Mace 2004). Here, the species concepts VC and VG in $Tax1$ are being *merged* into a single species concept M in $Tax2$. Vice versa, this may be considered a taxonomic *split*, going from $Tax2$ to $Tax1$.

We deployed a deep convolutional auto-encoder (CNN-AE), consisting of 8 strided convolutional layers *each* for encoding and decoding the embedded representation, and a

[3]Available from https://www.idigbio.org. Accessed Sep 1 2020.

dense feature embedding layer consisting of only 10 features $[a, .., j]$. This was done in order to "force" the network to learn a low-dimensional encoding that was hypothesized to be a better representation of taxonomically relevant features, following the intuition of (Guo et al. 2017). The model consisted of approximately 2 million parameters, and was trained end-to-end on the training images for the five species concepts, i.e. $Tax1.S$, $Tax1.VC$, $Tax1.VG$, $Tax2.S$, and $Tax2.M$. Training was done with the $adam$ optimizer (Kingma and Ba 2014) and standard mean square error loss, for 10 epochs on an NVIDIA $K80$ GPU. Hyperparameters, including strides, size of convolutions, number of layers, number of convolution filters, were tuned using standard iterative methods, all using a $70 - 15 - 15$ train-validation-test splitting.

In the absence of *any* provided, underlying taxonomic theory that would serve to annotate, interpret, and weigh the learnt features, our ability to derive a plausible taxonomic interpretation by inducing any hierarchical representation akin to a taxonomy is commensurably limited. This general problem has been studied as *conceptual clustering*, and represents a known, highly challenging problem. Instead, while we choose not to "optimize" the learnt features based on either given taxonomic theory, we may attempt to leverage this theory to learn about differential *taxonomic observations* or the *weighing practices* of the respective authors. This constitutes supervised learning, here not intended to identify known species within a *given* taxonomic context, but instead aimed at shedding light upon the nature of the taxonomic perspective itself.

An *authored practice* constitutes a method for selecting and weighing *criteria* on distinct interpreted features such as life-form, leaf shape, floral and fruit morphology, etc. This is reflected in the formation of a *decision tree* over the selected morphological features. Every known morphological feature may be defined over a set of learnt features; e.g. for a learnt feature vector X and known taxonomic feature f, there is some unknown function σ such that:

$$f = \sigma(X[i], ..., X[j]) \tag{1}$$

We used skope-rules[4] to obtain decision rules for each species class, over the learnt features. This uses a bagging estimator of decision trees along with semantic rule deduplication, and is a trade-off between the interpretability of a decision tree and the predictive power of a random forest. We then used skope-rules to induce decision rules describing the same classes over the morphological features of Fisher's *Iris* dataset. Since for each taxonomic concept, the decision rule over morphological features uniquely identifies the concept, we have $f(A, B, C, ...) \leftrightarrow T.X$, where T is a taxonomy and X is a species concept. Making the assumption that the species concept is uniquely identified by its biodiversity data signal, we have $T.X \leftrightarrow g([a, .., j])$, where $x, y, ...$ are the learnt features. Thus, we have a bi-implication between the rules over learnt features, and the rules over morphological features. We induced such decision rules to classify the images, as opposed to deploying an output layer of the neural network, in order that the resulting rules may be logically relatable in this way to the human-understandable morphological features. This constitutes a "grey box" model which may be solved for the values of morphological features, resulting in an entirely explainable/interpretable conclusion in terms of logical relationships, as described below.

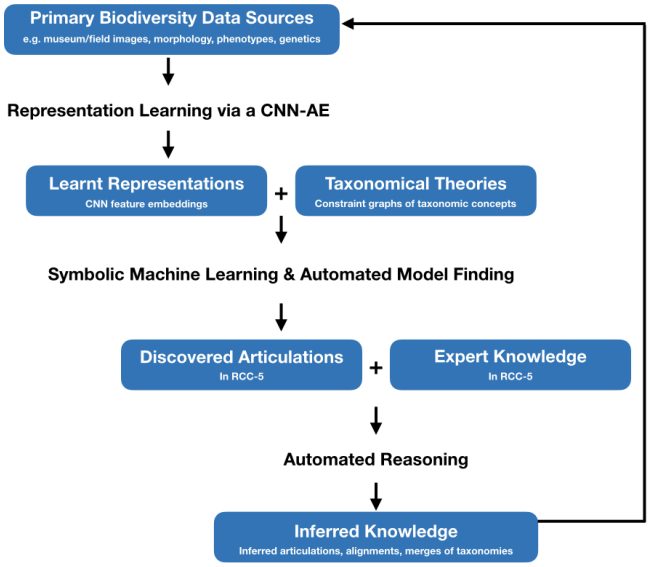

Figure 3: Architecture

We thus obtained a system of 5 bi-implications of linear real arithmetic (LRA) formulae: one for each of $Tax1.S$, $Tax1.VC$, $Tax1.VG$, $Tax2.S$, and $Tax2.M$. These were solved using the Microsoft Z3 SMT solver[5], which uses a dual-simplex solver for this category of problems. For each image in the dataset, the values of the morphological features were obtained.

The *RCC-5* (Cohn and Renz 2008) qualitative spatial reasoning (QSR) calculus consists of 5 relationships: equality ($=$), proper inclusion ($<$), inverse proper inclusion ($>$), overlap ($><$), and disjointness ($!$). Interpreting these relationships over sets (of images, here) instead of over space, we assign the relationship between any two species concepts from different taxonomies to one of these 5, depending on how the specimen images are assigned to each concept. Each such relationship is called a *taxonomic articulation*.

The entire process is summarized in Algorithm 1. Adjusting values of the thresholds $T1$ and $T2$, we obtained the RCC-5 articulations $Tax1.VC < Tax2.M$, $Tax1.VG < Tax2.M$, and $Tax1.S = Tax2.S$, as expected, for our two taxonomies $Tax1$ and $Tax2$ i.e. we have *learnt* that $Tax2$ is characterized by the "merging" of the taxonomic concepts VC and VG into the concept M. We cannot compare this result with other AI approaches to aligning taxonomies using learning from biodiversity data, since none exist in the literature, to the best of our knowledge.

[4]Available from https://github.com/scikit-learn-contrib/skope-rules. Accessed Sep 1 2020.

[5]Available from https://github.com/Z3Prover/z3/wiki. Accessed Sep 1 2020.

Algorithm 1: Articulation Discovery

Result: $RCC-5(Tax1.X, Tax2.Y)$
Given: Taxonomies $Tax1$, $Tax2$, species concepts $Tax1.X$, $Tax2.Y$, images, annotations $[A, B, ..]$;
foreach *taxonomy* T **do**
  train CNN-AE for each species concept of T, obtaining set of features F_S;
  foreach *species concept* X *in* T **do**
    induce $R1$: decision rule for X over F;
    compute $R2$: decision rule for X over $[A, B, ..]$;
    assign S_X: $R1 \leftrightarrow R2$
foreach *image with learnt feature set* F **do**
  Solve system of LRA equivalences S with F, obtaining values of $[A, B, ..]$;
  Using $R1$, classify image in each taxonomy.
switch # *of images co-classified in* $Tax1.X$ *and* $Tax2.Y$*; thresholds* $T1\%$, $T2\%$, $T3\%$ **do**
  case $\geq T1\%$ **do**
    return $=$
  case $\geq T2\%$ *in* $Tax1.X$ *also in* $Tax2.Y$ *or* $\geq T2\%$ *in* $Tax2.Y$ *also in* $Tax1.X$ **do**
    return $<$ *or* $>$ *resp.*
  case $\geq T3\% \leq T1\%$ **do**
    return $><$
  case $\leq T3\%$ **do**
    return !

Reasoning with Taxonomic Articulations

In addition to discovering taxonomic articulations, it is desirable to leverage the considerable multi-taxonomy integration knowledge explicit or implied in primary biodiversity data sources. This takes the form of a human–machine collaboration, with deductive inference from a conjunction of human and machine-learnt knowledge, that informs both future machine learning, and human insight. This is depicted in the "feedback" arrow in Figure 3.

The key technical distinction is that reasoning over a finite number of species concept instances – i.e., those for which data is available – is propositional, whereas reasoning about general articulations between species concepts necessitates *quantification* over all possible instances of that concept.

The challenge of *aligning* multiple taxonomic theories using automated reasoning was formally described by Thau (Thau and Ludäscher 2007). This conception included proving that the alignments are logically consistent, or, if inconsistent, explaining the (joint) causes for inconsistency, and inferring implicit relationships – i.e., the set of *Maximally Informative Relations* – among all entailed concepts, based on limited yet explicit prior expert or source input. Thau (Thau and Ludäscher 2007) deployed the First-Order resolution theorem prover Prover9 and the accompanying model finder Mace4 (McCune 2005–2010). Related work – e.g., the development of the $Euler/X$ system (Chen et al. 2014)) – has both extended the scope of functionality referred to as taxonomy alignment and significantly increased the scalability of the reasoning process, by using a variety of techniques such as *Answer Set Programming* (ASP) and *Qualitative Spatial Reasoning (QSR)* The problem may be equivalently formulated as a QSR problem, definable – with the sole exception of the global constraint of *coverage* (see later in this section) – in the RCC-5 calculus.

However, this approach relies fundamentally on taxonomic insights (referred to as *articulations*) provided by human experts, which must often be meticulously extracted from the literature (Franz and Peet 2009; Franz, Pier et al. 2016; Franz, Musher et al. 2019). As we have shown in this paper, primary biodiversity data sources may be leveraged to inform and explain such taxonomic decisions, illuminating the relationship *between* two (or more) *given* taxonomic theories.

It was shown by Thau (Thau and Ludäscher 2007) that this problem of reasoning over taxonomies can be captured in *Monadic First-Order Logic (MFOL)*. This logic is able to represent not only the taxonomies and RCC-5 articulations, but several taxonomically plausible or necessary global constraints: the non-emptiness of taxonomic concepts, sibling-disjointedness (i.e., the disjointedness of child concepts of the same taxonomic parent concept), and, *coverage*, which denotes the notion that the children denoted by a taxonomic parent concept are completely *covered*, or included in, that parent (which therefore has no further "extension"). Local, selective relaxation of each constraint allows modeling of a wide variety of systematic use cases (e.g., (Franz, Musher et al. 2019)).

The decidability of MFOL was proven long ago by Löwenheim (Löwenheim 1915), and implemented in modern *Satisfiability Modulo Theory (SMT)* solvers such as Microsoft $Z3$, which incorporate a decision procedure, called *Model Based Quantifier Instantiation (MBQI)* (Reynolds et al. 2013), for the *effectively propositional* class of formulae (i.e., the Bernays–Schönfinkel–Ramsey class) that makes it possible to encode problems involving constraints over sets by treating the sets as unary predicates and lifting equalities between sets as formula equivalences. Set constraints are, in turn, equivalent to MFOL in expressivity (Bachmair, Ganzinger, and Waldmann 1993). Thus, reasoning over taxonomies, as defined by Thau, admits a propositional encoding (Ramachandran and Amir 2005) and may be decided by an SMT solver.

We have developed a publicly available, open-sourced web application for aligning taxonomies using Z3, which follows the algorithm by Thau (Thau and Ludäscher 2007), but is *decidable* as described above, as opposed to the semi-decidability of the original first-order encoding. It also enables an interactive workflow as a form of the human–machine collaboration noted earlier: the user's choice of provided articulations is informed by the deductive reasoning. Please see the note on Supplementary Materials at the end of Section .

We employed our software to perform an alignment of the two *Iris* taxonomies $Tax1$ and $Tax2$ described in Section . The resulting visual alignment is shown in Figure 4. The grey arrows represent the taxonomical hierarchy. The green lines represent the provided articulations, which were

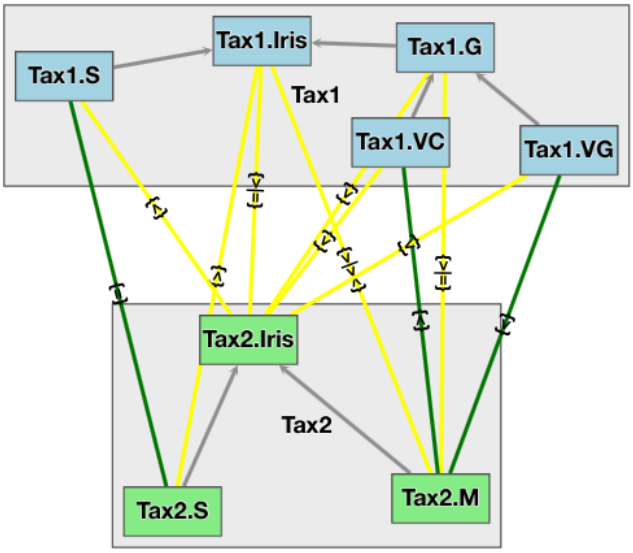

Figure 4: Screenshot from our web-based, interactive, and publicly available tool ATCRL (Automated Taxonomic Concept Reasoner & Learner), aligning the two *Iris* taxonomies using the discovered RCC-5 articulations $Tax1.VC < Tax2.M$, $Tax1.VG < Tax2.M$, and $Tax1.S = Tax2.S$.

learnt by Algorithm 1. The yellow lines represent the *deductively inferred* RCC-5 articulations, which appear in curly braces. The alignment represents one *possible world* that satisfies the set constraints presented to the reasoner; hence, there is uncertainty in the RCC-5 articulation between several concepts, represented by the | symbol. The resultant inferred articulations are exactly as expected, with $Tax1.G = Tax2.M$ and $Tax1.Iris = Tax2.Iris$ as a model.

The complete process described in this section and the previous section, is depicted in Figure 3.

Social Impact

Large aggregated biodiversity datasets are becoming important tools for addressing contemporary social challenges such as modeling the impacts of climate change (Calinger, Queenborough, and Curtis 2013; Willis et al. 2017), identifying medicinal resources from understudied locations (Souza and Hawkins 2017), documenting invasion trends of agricultural *weeds* (Crawford and Hoagland 2009), and food security initiatives (Ng'uni et al. 2019). Major international agreements, such as the Convention on Biological Diversity (CBD) and Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), also regulate the international scientific and commercial exchange of living and preserved organisms ('t Sas-Rolfes et al. 2019; Le Prestre 2017). Wildlife markets in particular have recently been implicated as part of the causal chain linking existing animal reservoirs for the SARS-CoV-2 virus with its spillover into humans, leading to the global pandemic outbreak in early 2020 (Huang et al. 2020).

All of these applications to ecosystem services or threats, law, and policy rely on an integrative approach to information about recognized species (boundaries) and their evolutionary relationships. However, it is widely acknowledged and lamented that human taxonomic expertise is lacking on the massive scale required to simultaneously address these demands for robust evidence (Hoagland 1996; Dar et al. 2012; Draper et al. 2020). The proposed architecture in Figure 2 augments the information and impact provided in expert-curated datasets to bridge gaps of semantic interoperability between alternative classificatory theories and their efficient application to multi-modal and novel data sources.

In this section, we use the problem of monitoring online illegal trade in wildlife to highlight the future value of the AI tools for biodiversity data intelligence we have presented here. Online websites, especially social media platforms such as Facebook, Twitter, and Instagram, have become an increasingly popular place to buy and sell animals and plants across the planet, and researchers are increasingly devoting attention to documenting the intensity of illegal trade in different countries (Siriwat and Nijman 2018; Minin et al. 2019). Approximately 66-84 million caged birds are kept annually by 36 million households in Indonesia (Marshall et al. 2020), for example, and illegal exploitation constitutes a major risk for many endangered species (Scheffers et al. 2019), including *Iris* species such as *Iris boissieri* (Lírio de serra) (Sapir 2016).

Species names provide essential information for buyers and sellers as well as legal regulations and practices since prices are known to vary by species based on their properties, including perceived medicinal properties or cultural uses (Siriwat and Nijman 2020). However, online posts regularly use vernacular names to advertise species for sale, often with an attached image, while legal protections are typically assigned to published taxonomic names – most typically at the species rank and in relation to a particular taxonomic authority. Yet at the global scale, applied national and international laws represent an obscure, undercontextualized *patchwork* of incongruent taxonomic classifications, e.g. for Primates (Svensson et al. 2016). This is partly a result of how irregularly lists of protected species are updated to reflect taxonomic revisions and the adoption of alternative taxonomic authorities for particular groups (Svensson et al. 2016; Panter and White 2020). Across mammals, which include over half of the CITES Appendix I list of trafficked species, the species-level taxonomy was only updated once in the past 15 years (Burgin et al. 2018). Multiple classifications are often in play simultaneously for labeling the same organisms and must be related to automate accurate and salient identification of illegal commercial activity for the different legal regimes pertinent to protecting each species.

Cutting-edge approaches to monitoring online wildlife trade include: (1) descriptive studies that produce expert-curated data records annotated to particular taxonomic classifications; and (2) machine learning methods for reliable species identification using photographs taken at monitoring locations, e.g. (Minin et al. 2019; Olschofsky and Köhl 2020). By leveraging such monitoring-focused datasets and publicly available museum or citizen science data, our ap-

proach has the potential to add the further step of *learning* relationships between names and concepts across taxonomies, filling an important gap for the detection of illegal trade under a dynamic patchwork of inter-/national legal regimes (Jacobs and Baker 2018). One benefit of learning taxonomic alignments is that we gain the ability to reason about whether an organism is protected under different legal frameworks in scenarios where it has been identified manually, with or without image data, and where manual identification is missing but image (or potentially other) data sources are available.

The present use case is emblematic of the broader value of learning and reasoning with taxonomic intelligence to model the impacts of biodiversity scenarios – as reflected in decentralized data sources – on human lives and well-being. Scalable taxonomic intelligence is critical, for example, to enabling trustworthy and fit-for-use biodiversity data aggregation (Sterner, Gilbert, and Franz "In press") for applications such as testing whether certain animal groups pose higher risks of producing novel zoonotic diseases than expected by chance (Guy et al. 2019), and for establishing temporal baselines for species extinction risks as part of the Red List of Threatened Species produced by the International Union for Conservation of Nature (Pacifici et al. 2019).

Conclusions, Limitations, & Outlook

The variety, volume, and velocity of biodiversity data are rapidly growing, with little chance of stable, global consensus on essential metadata categories. While we have restricted our attention to image data associated with species observations – in part because of the great interest in exploiting this data for species identification – we envision a plethora of data sources, including text, DNA sequences, and geo-spatial information, being immediately relevant. As more nations and organizations launch biodiversity monitoring projects, coordinating these decentralized efforts will pose a major challenge that exceeds any foreseeable capacity of humans to address, without assistance from AI.

For example, citizen science platforms such as iNaturalist must regularly update their image-based taxonomic identification algorithms to reflect changes to adopted standards for taxonomic classification. Providers of these standard taxonomies (also known as "taxonomic authorities"), however, are rarely staffed or funded to provide semantically annotated versioning information. As a result, biodiversity data users must manually articulate relations in order to determine how to update their datasets for retraining machine learning models for automated image classification. This manual articulation of logical relations between alternative classificatory systems represents an important but often overlooked dimension of biodiversity knowledge (Sterner and Franz 2017), with significant costs in human labor, lost opportunities, and downstream performance.

Developing an accurate and scalable AI for taxonomic intelligence will also be crucial to downstream computational reasoning for testing the robustness of conservation decision-making in light of conflicting or uncertain taxonomies, which can be in itself sufficient to move a group of organisms in or out of consideration for legal protection as an endangered group. Similarly, accurate data aggregation and the facilitation of alternative classificatory viewpoints is crucial to closing the feedback loop between primary data sources and curation work by data users in siloed computing systems (Franz and Sterner 2018).

Our architecture in Figure 3 is not constrained by one particular taxonomic theory, and is instead explicitly designed to inspire a taxonomic intelligence that agglomerates and aligns different taxonomic theories. It does so by drawing from both articulations in the literature and vast primary biodiversity data sources, constituting what we consider to be the most promising road-map for a scalable and accurate taxonomic intelligence. Inferences drawn from reasoning processes may be used to annotate the primary data, forming a reinforcement loop leveraging both inductive and deductive reasoning. In this sense reasoning can simplify the learning task with both deductively-obtainable conclusions and assertions from the literature.

Effective integration of the architectural components will rely centrally on effective reasoning *with uncertainty* about logical articulations within and across taxonomic hierarchies, which remains outside the scope of this paper. For example, in order to update primary biodiversity databases with new and better aggregated data, it will be important to employ inferred logical articulations with high certainty. Aligning taxonomies regularly uncovers new sub-partitions of the observed dataset that are not recognized in any individual taxonomy, but are logically implied when one concept sub-divides or overlaps with another (Vaidya, LePage, and Guralnick 2018). Automated discovery and characterization of these novel concepts can serve to add finer-grained labels to the initial data sources and drive further successes in taxonomic identification.

Our approach to taxonomic intelligence has the potential to generalize to other types of data with similar, implicit but rich, feature sets, such as species occurrence observations. While a few well-known computer ontologies, e.g. the Gene Ontology (Consortium 2017), have achieved substantial coverage for explicit and combinatorial concept definitions, many important scientific datasets cannot be straightforwardly annotated by a rule-based approach (Bertone et al. 2013). In even fewer cases is it possible to align ontologies by directly identifying logically equivalent terms based on their concept definitions. Interesting other domains in a biodiversity context include taxonomies for ecological biomes and modes of land use, such as urban versus rural or wild habitats.

Acknowledgments

Our thanks to our referees for their detailed and helpful comments, and to funding support from the Arizona State University President's Special Initiative Fund and NSF grant STS-1827993. We also thank Jonathan Rees for his valuable advice.

Ethics Statement

The AI methods motivated and developed in this paper directly aim to better manage and conserve the biodiversity of the planet we live on. The numerous positive societal implications are discussed in Section . We do not foresee any reasonable risk of misuse of this work.

References

Angermann, H.; and Ramzan, N. 2017. *Taxonomy Matching Using Background Knowledge*. Springer.

Bachmair, L.; Ganzinger, H.; and Waldmann, U. 1993. Set constraints are the monadic class. In *[1993] Proceedings Eighth Annual IEEE Symposium on Logic in Computer Science*, 75–83. IEEE.

Bertone, M. A.; Mikó, I.; Yoder, M. J.; Seltmann, K. C.; Balhoff, J. P.; ; and Deans, A. R. 2013. Matching Arthropod Anatomy Ontologies to the Hymenoptera Anatomy Ontology: Results From a Manual Alignment. *Database* .

Blagoderov, V.; Kitching, I. J.; Livermore, L.; Simonsen, T. J.; and Smith, V. S. 2012. No specimen left behind: industrial scale digitization of natural history collections. *Zookeys* 209: 133–146.

Burgin, C. J.; Colella, J. P.; Kahn, P. L.; and Upham, N. S. 2018. How many species of mammals are there? *Journal of Mammalogy* 99(1): 1–14.

Calinger, K. M.; Queenborough, S.; and Curtis, P. S. 2013. Herbarium specimens reveal the footprint of climate change on flowering trends across north-central North America. *Ecology Letters* 16(8): 1037–1044.

Chen, M.; Yu, S.; Franz, N.; Bowers, S.; and Lu´asher, B. 2014. Euler/X: A Toolkit for Logic-based Taxonomy Integration.

Cohn, A. G.; and Renz, J. 2008. Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence* 3: 551–596.

Consortium, T. G. O. 2017. Expansion of the Gene Ontology Knowledgebase and Resources. *Nucleic Acids Research* 45: D331–38.

Crawford, P. H. C.; and Hoagland, B. W. 2009. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? *Journal of Biogeography* 36(4): 651–661.

Dar, G. H.; Khuroo, A. A.; Reddy, C. S.; and Malik, A. H. 2012. Impediment to Taxonomy and Its Impact on Biodiversity Science: An Indian Perspective. *Proceedings of the National Academy of Sciences, India, Section B: Biological Sciences* 82.

Draper, F. C.; Baker, T. R.; Baraloto, C.; Chave, J.; Costa, F.; Martin, R. E.; Pennington, R. T.; Vicentini, A.; and Asner, G. P. 2020. Quantifying Tropical Plant Diversity Requires an Integrated Technological Approach. *Trends in Ecology & Evolution* .

Franz; and Peet. 2009. Perspectives: Towards a Language for Mapping Relationships Among Taxonomic Concepts. *Systematics and Biodiversity* 7: 5–20.

Franz, N. M.; Musher, L. M.; et al. 2019. Verbalizing Phylogenomic Conflict: Representation of Node congruence Across Competing Reconstructions of the Neoavian Explosion. *PLoS Computational Biology* 15: e1006493.

Franz, N. M.; Pier, N. M.; et al. 2016. Two Influential Primate Classifications Logically Aligned. *Systematic Biology* 65(4): 561–582.

Franz, N. M.; and Sterner, B. W. 2018. To Increase Trust, Change the Social Design Behind Aggregated Biodiversity Data. *Database* .

Godfray, H. C. J.; Knapp, S.; and Mace, G. M. 2004. The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359(1444): 711–719. Publisher: Royal Society.

Guala, G. F. 2016. The Importance of Species Name Synonyms in Literature Searches. *PLoS ONE* 11: e0162648.

Guo, X.; Liu, X.; Zhu, E.; and Yin, J. 2017. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, 373–382. Springer.

Guy, C.; Thiagavel, J.; Mideo, N.; and Ratcliffe, J. M. 2019. Phylogeny matters: revisiting 'a comparison of bats and rodents as reservoirs of zoonotic viruses'. *Royal Society Open Science* 6(2): 181182. Publisher: Royal Society.

Hardisty, A. R.; Michener, W. K.; et al. 2019. The Bari Manifesto: an Interoperability Framework for Essential Biodiversity Variables. *Ecological Informatics* 49: 22–31.

Hoagland, K. E. 1996. The taxonomic impediment and the convention on biodiversity. *Association of Systematics Collections Newsletter* 24(5): 61–62.

Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; and Fan, G. e. a. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 395(10223): 497–506.

Jacobs, R. L.; and Baker, B. W. 2018. The species dilemma and its potential impact on enforcing wildlife trade laws. *Evolutionary Anthropology: Issues, News, and Reviews* 27(6): 261–266.

Jeliazkov, A.; Mijatovic, D.; et al. 2020. A Global Database for Metacommunity Ecology, Integrating Species, Traits, Environment and Space. *Scientific Data* 7: 1–15.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization.

Le Prestre, P. G. 2017. *Governing global biodiversity: The evolution and implementation of the convention on biological diversity*. Routledge.

Löwenheim, L. 1915. Über möglichkeiten im relativkalkül. *Mathematische Annalen* 76(4): 447–470.

Marshall, H.; Collar, N. J.; Lees, A. C.; Moss, A.; Yuda, P.; and Marsden, S. J. 2020. Spatio-temporal dynamics of consumer demand driving the Asian Songbird Crisis. *Biological Conservation* 241: 108237.

McCune, W. 2005–2010. Prover9 and Mace4. http://www.cs.unm.edu/ mccune/prover9/.

Mesibov, R. 2013. A Specialist's Audit of Aggregated Occurrence Records. *ZooKeys* 293.

Mesibov, R. 2018. An Audit of Some Processing Effects in Aggregated Occurrence Records. *ZooKeys* 751: 129–46.

Minin, E. D.; Fink, C.; Hiippala, T.; and Tenkanen, H. 2019. A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology* 33(1): 210–213.

Müller-Wille, S.; and Charmantier, I. 2012. Natural History and Information Overload: the Case of Linnaeus. *Studies in the History and Philosophy of Biological and Biomedical Sciences* 43: 4–15.

Nelson, G.; and Ellis, S. 2019. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374(1763): 20170391.

Ng'uni, D.; Munkombwe, G.; Mwila, G.; Gaisberger, H.; Brehm, J. M.; Maxted, N.; Kell, S.; and Thormann, I. 2019. Spatial analyses of occurrence data of crop wild relatives (CWR) taxa as tools for selection of sites for conservation of priority CWR in Zambia. *Plant Genetic Resources: Characterization and Utilization* 17(2): 103–114.

Olschofsky, K.; and Köhl, M. 2020. Rapid field identification of CITES timber species by deep learning. *Trees, Forests and People* 2: 100016.

Pacifici, M.; Cristiano, A.; Burbidge, A. A.; Woinarski, J. C. Z.; Di Marco, M.; and Rondinini, C. 2019. Geographic distribution ranges of terrestrial mammal species in the 1970s. *Ecology* 100(7): e02747.

Panter, C.; and White, R. 2020. Insights from social media into the illegal trade of wild raptors in Thailand. *TRAFFIC Bulletin* 32(1): 5–12.

Peterson, A. T.; Soberón, J.; and Krishtalka, L. 2015. A Global Perspective on Decadal Challenges and Priorities in Biodiversity Informatics. *BMC Ecology* 15: 15.

Ramachandran, D.; and Amir, E. 2005. Compact propositional encodings of first-order theories.

Remsen, D. 2016. The use and limits of scientific names in biological informatics. *ZooKeys* 550: 207–223.

Reynolds, A.; Tinelli, C.; Goel, A.; Krstić, S.; Deters, M.; and Barrett, C. 2013. Quantifier instantiation techniques for finite model finding in SMT. In *International Conference on Automated Deduction*, 377–391. Springer.

Sapir, Y. 2016. IUCN Red List of Threatened Species: Iris boissieri.

Scheffers, B. R.; Oliveira, B. F.; Lamb, I.; and Edwards, D. P. 2019. Global wildlife trade across the tree of life. *Science* 366(6461): 71–76. Publisher: American Association for the Advancement of Science Section: Research Article.

Siriwat, P.; and Nijman, V. 2018. Illegal pet trade on social media as an emerging impediment to the conservation of Asian otters species. *Journal of Asia-Pacific Biodiversity* 11(4): 469–475.

Siriwat, P.; and Nijman, V. 2020. Wildlife trade shifts from brick-and-mortar markets to virtual marketplaces: A case study of birds of prey trade in Thailand. *Journal of Asia-Pacific Biodiversity* 13(3): 454–461.

Souza, E. N. F.; and Hawkins, J. A. 2017. Comparison of Herbarium Label Data and Published Medicinal Use: Herbaria as an Underutilized Source of Ethnobotanical Information. *Economic Botany* 71(1): 1–12.

Sterner, B.; Gilbert, E.; and Franz, N. M. "In press". Decentralized but globally coordinated biodiversity data. *Frontiers in Big Data* .

Sterner, B.; Witteveen, J.; and Franz, N. M. 2020. Coordinating dissent as an alternative to consensus classification. *History and Philosophy of the Life Sciences* 42.

Sterner, B. W.; and Franz, N. M. 2017. Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data. *Biological Theory* 12: 99–111.

Svensson, M. S.; Shanee, S.; Shanee, N.; Bannister, F. B.; Cervera, L.; Donati, G.; and Huck, M. e. a. 2016. Disappearing in the Night: An Overview on Trade and Legislation of Night Monkeys in South and Central America. *Folia Primatologica* 87(5): 332–348. Publisher: Karger Publishers.

't Sas-Rolfes, M.; Challender, D. W.; Hinsley, A.; Veríssimo, D.; and Milner-Gulland, E. 2019. Illegal Wildlife Trade: Scale, Processes, and Governance. *Annual Review of Environment and Resources* 44(1): 201–228.

Tan, K. C.; Liu, Y.; Ambrose, B.; Tulig, M.; and Belongie, S. 2019. The Herbarium Challenge 2019 Dataset.

Thau, D.; and Ludäscher, B. 2007. Reasoning about taxonomies in first-order logic. *Ecological Informatics* 2(3): 195–209.

Vaidya, G.; LePage, D.; and Guralnick, R. 2018. The Tempo and Mode of the Taxonomic Correction Process: How Taxonomists Have Corrected and Recorrected North American Bird Species Over the Last 127 Years. *PLoS ONE* 13: e0195736.

Wäldchen, J.; and Mäder, P. 2018a. Machine Learning for Image Based Species Identification. *Methods in Ecology and Evolution* 9: 2216–25.

Wäldchen, J.; and Mäder, P. 2018b. Machine learning for image based species identification. *Methods in Ecology and Evolution* 9(11): 2216–2225.

Willis, C. G.; Ellwood, E. R.; Primack, R. B.; Davis, C. C.; Pearson, K. D.; Gallinat, A. S.; Yost, J. M.; Nelson, G.; Mazer, S. J.; Rossington, N. L.; Sparks, T. H.; and Soltis, P. S. 2017. Old Plants, New Tricks: Phenological Research Using Herbarium Specimens. *Trends in Ecology & Evolution* 32(7): 531–546.