Scientific Names and Descriptions for Organisms on the Semantic Web

Nathan Wilson[1], Han Wang[2], and Deborah McGuinness[2]

[1] Marine Biological Laborary, 7 MBL St., Woods Hole, MA 02556, USA nwilson@mbl.edu
[2] Rensselaer Polytechnic Institute, 110 8th Street Troy, NY 12180 USA

Abstract. An efficient process for creating precise, accurate, machine-interpretable morphological descriptions of groups of organisms is needed to more effectively gather observations of the world's biodiversity. While morphological descriptions are required for the publication of modern scientific names, it is common for these descriptions to get revised after the initial publication which can lead to data loss. A system for creating and naming machine-interpretable descriptions of groups of organisms, the Semantic Vernacular System, is proposed as a solution for creating such descriptions and managing their relationship to formal scientific names while improving the collection of observational data.

1 Introduction

We envision a world where everyone is empowered to contribute to the scientific observation of biodiversity. Between the ever increasing access to and use of the Internet throughout the world and the well documented extinction crisis [17], this is a key time to effectively take advantage of the 'crowd' to help discover, document and ultimately manage the world's biodiversity.

A number of web-based biodiversity observation systems intended for this purpose already exist such as eBird [15], ArtPortalen [1], iNaturalist [16] or Mushroom Observer [19] and have already collected tens of millions of occurrence records. All of these tools work by connecting biodiversity observations to the scientific literature using scientific names. For all such systems, it would be desirable to provide precise, accurate and ideally machine-interpretable definitions of the scientific names. While a few new species names have been published with semantic descriptions [12], there is no central repository for these and as yet no widely accepted standards for how to create such descriptions. In addition, the occurrence records in the current systems often lack sufficient evidence to validate the identifications.

This position paper argues that both the lack of precise and accurate descriptions, and the failure to gather supporting evidence can be addressed by adding a layer of abstraction between observations and scientific names that consists of named semantic descriptions.

2 Scientific Names and Observations of Biodiversity

Scientific names are critical for understanding the biological literature and provide a valuable way to understand evolutionary relationships. They are also important for making biological information more computable [13]. The scientific names for naturally occurring species are governed by three nomenclatural codes, the ICNAFP [8], the ICZN [5] and the ICNB [4]. In all of these codes, the names are fundamentally defined based on a physical example, typically a museum specimen, known

as the 'type specimen'. For all names, there is an assumption that all the organisms included are part of a single evolutionary lineage. A formal description or 'circumscription' is required for a name to be validly published. The circumscription is intended to separate the described group of organisms from those described by other names at the same level. The primacy of the type specimen and the evolutionary lineage is demonstrated by the frequent revision of circumscriptions when new evolutionary evidence is found with respect to the type specimen. For example, the circumscription of *Laetiporus sulphureus*, the Sulphur Shelf Mushroom, was significantly reduced when mating studies showed that the original circumscription included at least 5 distinct species [6]. The fundamental justification for this approach is to tie scientific names to the evolutionary relationships between organisms. The side effect of these rules is that a given scientific name may have multiple circumscriptions and a given circumscription may apply to multiple scientific names.

While many biodiversity observation systems encourage photographs and notes to document an observation, frequently a scientific name is simply asserted with no explicit evidence. When just a name is provided, there is an immediate loss of information for any name that has multiple circumscriptions in current use. For example, if an occurrence of *Laetiporus sulphureus* is recorded, it is not clear which existing circumscription was intended. This loss would be further compounded by any further revisions unless the users of that data are careful to keep track of when new circumscriptions are created relative to each observation.

One possible approach would be for biodiversity observation systems to require that users provide an explicit reference to the circumscription they used when making an identification. Sites such as the Encyclopedia of Life (EOL) [3] and the Biodiversity Heritage Library [2] are beginning to provide online resources that could be used for this purpose. However, they are still far from comprehensive so finding and documenting the appropriate references remains a time consuming process which is effectively impossible for many observers. Even so this approach is problematic since it would not require recording the observed features to help validate the identifications in the future.

A better approach from a data management standpoint would be to require a person making an identification to be explicit about the features they based their identification on. However, given the lack of even standardized terminology for many groups, this in effect means the identifier would have to write a detailed description for each identification. Validation would be better supported with this approach, but would be very subject to the consistency and thoroughness of the identifier. Both of these approaches would frequently impose a significant overhead burden on the taxonomic experts.

3 Semantic Vernacular System

The proposed alternative is to create the Semantic Vernacular System which enables authoring named, machine-interpretable definitions of groups of organisms that are then associated with sets of scientific names. The Web Ontology Language [11], is a natural fit for such a definitional system and is already being actively used by significant parts of relevant biology communities including members of the NSF Phenotype Ontology Research Collaboration Network.

Just as with scientific names, it would be valuable for the new system to be peer-reviewed, strictly prevent the reuse of names, apply any agreed upon nomenclatural rules, and avoid the unintended re-publishing of an existing description. The Semantic Vernacular Descriptions, or SVDs, envisioned will be 'born digital' in a freely available, global repository. This will allow all of these desirable features to be applied consistently and automatically from the ground up. Once an SVD has been approved through the peer-review process the association of the chosen name with that semantic

description will be considered a strict definition that should never change just as the association between a scientific name and its type specimen should never change. This tight association means that a particular SVD applies to any observed organism that matches that semantic description. SVDs will allow users to apply precise sets of features by name to their observations. Biodiversity observation systems using the system will allow users to automatically review and even explicitly confirm the defining features for any name they apply. While there will, of course, be human error in the application of such names and even specific features, the data recorded by the observer will be explicit and will not degrade over time. For *Laetiporus sulphureus* this means separate names for the historical circumscription as well as names for each morphologically distinct group whether it includes multiple species, a single species, or a distinctive subset of one or more species.

The proposed system will also support the registration and naming of descriptions that correspond to common observational experiences. This will allow users to record what are in effect partial identifications to groups of 'look alikes' that may or may not include all the members of an evolutionary lineage. Currently groups of species that are difficult to tell apart in the field, sometimes referred to as 'sibling speices' [10] or more generally 'cryptic species', are handled in a variety of inconsistent ways including using higher-level taxa such as genera or families, modifying species names by inserting 'aff.' or 'cf.', adding 'group' or 'complex' at the end of the name, or by informal names such 'Comic Tern' or 'Circus macrourus/cyaneus'. Ever improving resolution of genetic information will increase the gap between formally recognized species and recognizable groups of organisms [9] [14]. The Semantic Vernacular System will provide a way to formally define the recognizable groups while maintaining the many-to-many relationship between SVDs and traditional scientific names. These relationships will allow users to continue to use scientific names as an entry point into the system from which to find SVDs to help efficiently describe their observations.

Populating such a system with meaningful descriptions will not be simple. It must start with the development of ontologies that capture and standardize the terminology needed to describe observed features. This work has begun in some groups such as Teleost fish [7] and the Hymenoptera [20]. These efforts have created workflows for collaboratively developing such ontologies. We expect the work of creating SVDs will start with groups of organisms which are not well handled by existing nomenclatural codes such as polyphyletic groups and cryptic species complexes. Familiarity with the emerging ontologies will in turn encourage users to describe and name SVDs for common, larger monophyletic groups and eventually common species. The natural bias towards common observational experiences will focus the system on the areas where it is expected to have the greatest value. Complete coverage of all existing scientific names is neither necessary nor expected.

A demonstration system [18] created in collaboration with the Mushroom Observer and the EOL is available at `http://mushroomobserver.org/semantic_vernacular`. Mushrooms are an excellent example case for exploring these issues since many circumscriptions and observations are based solely on the mushroom which is roughly equivalent to the flower of the larger fungal organism. As a result there are many examples of polyphyletic groups and species complexes that are difficult to identify to species based on easily observed characteristics. The EOL is a natural aggregator for such descriptions as it already has support for polyphyletic, provisional and other non-standard names that are provided to the system through its content providers. In addition, the EOL is actively working to support 'computable' data using semantic web technology.

Finally, the proposed machine-interpretable definitions naturally lead to a novel system for computer-aided identification of observations. As users learn to describe their observations in the same way that the descriptions are stored, it will be straight-forward for the system to indicate what existing SVDs match the given features and the implied set of potential scientific names.

4 Conclusion

This position paper describes the need for a new class of names tightly associated with semantic descriptions of groups of organisms. We outline the creation of the Semantic Vernacular System for managing these new names and descriptions. This system will enable more precise and accurate observations of biodiversity with minimal additional overhead while encouraging the creation of machine-interpretable descriptions with clear connections to traditional scientific names.

References

1. Artportalen, `http://artportalen.se`
2. Biodiversity Heritage Library, `http://biodiversitylibrary.org`
3. Encyclopedia of Life, `http://eol.org`
4. International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. ASM Press, Washington, DC, USA (1992)
5. International Code of Zoological Nomenclature. The International Trust for Zoological Nomenclature, London, UK, 4th edn. (2000)
6. Burdsall, H.H., Bank, M.T.: The Genus *Laetiporus* in North America. Harvard Papers in Botany 6(1), 43–55 (2001)
7. Dahdul, W.M., Lundberg, J.G., Midford, P.E., Balhoff, J.P., Lapp, H., Vision, T.J., Haendel, M.A., Westerfield, M., Mabee, P.M.: The Teleost Anatomy Ontology: Anatomical Representation for the Genomics Age. Systematic Biology 59, 369–383 (2010), doi:10.1093/sysbio/syq013
8. Knapp, S., McNeill, J., Turland, N.J.: Changes to Publication Requirements Made at the XVIII International Botanical Congress in Melbourne - What Does e-Publication Mean for You? PhytoKeys 6(0), 5–11 (2011), doi:10.3897/phytokeys.6.1960
9. Knowlton, N.: Sibling Species in the Sea. Annual Review of Ecology and Systematics 24, 189–216 (1993), doi:10.1146/annurev.es.24.110193.001201
10. Mayr, E.: The Bearing of the New Systematics on Genetical Problems. The Nature of Species. Advances in Genetics 2, 205–237 (1948)
11. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation. (2004), `http://www.w3.org/TR/owl-features/`
12. Mikó, I., Deans, A.R.: *Masner*, a New Genus of Ceraphronidae (Hymenoptera: Ceraphronoidea) Described Using Controlled Vocabularies. ZooKeys 20, 127–153 (2009), doi:10.3897/zookeys.20.119
13. Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L., Remsen, D.P.: Names are Key to the Big New Biology. Trends in Ecology & Evolution 25(12), 686–691 (2010), doi:10.1016/j.tree.2010.09.004
14. Sato, H., Yumoto, T., Murakami, N.: Cryptic Species and Host Specificity in the Ectomycorrhizal Genus *Strobilomyces* (Strobilomycetaceae). American Journal of Botany 94(10), 1630–1641 (2007)
15. Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S.: eBird: a Citizen-based Bird Observation Network in the Biological Sciences. Biological Conservation 142, 2282–2292 (2009), doi:10.1016/j.biocon.2009.05.006
16. Ueda, K., Loarie, S.: iNaturalist, `http://inaturalist.org`
17. Wilson, E.O.: The Future of Life. Random House Digital, Inc. (2002)
18. Wilson, N., Dunn, K., Wang, H., McGuinness, D.L.: Application of Semantic Technology to Define Names for Fungi. Tech. rep., Tetherless World Constellation at Rensselaer Polytechnic Institute (2012), `http://tw.rpi.edu/web/doc/ApplicationofSemanticTechnologytoDefineNamesforFungi`
19. Wilson, N., Hollinger, J.: Mushroom Observer, `http://mushroomobserver.org`
20. Yoder, M.J., Mikó, I., Seltmann, K.C., Bertone, M.A., Deans, A.R.: A Gross Anatomy Ontology for Hymenoptera. PLoS ONE 5(12), e15991 (2010), doi:10.1371/journal.pone.0015991