Application of Semantic Technology to Define Names for Fungi

by Nathan Wilson[1], Kathryn M. Dunn[2], Han Wang[3], and Deborah L. McGuinness[4]

v1.0 - April 27, 2012

Abstract

The need for well-defined, persistent descriptions of taxa that can be accurately interpreted by computers is becoming increasingly clear. The goal of this work is to develop named descriptions of Fungi that enable automated reasoning by computers. We encode these descriptions using the Web Ontology Language (OWL). The initial target audience is field mycologists using the Mushroom Observer website, who range from professional scientists to beginning mushroom enthusiasts. We describe our mycology ontology and propose developing a transparent, community-based ontology evolution process. The ontology was designed to focus on properties that can be observed in the field, but the framework is proving to be suitable for microscopic, chemical, and genomic properties as well. Concrete examples are provided where field mycologists need names for groups of similar-looking Fungi that are known to belong to different species, and where our approach can significantly increase the precision of information recorded by the observer. Such a system is important for enabling the field mycologist to make more meaningful contributions to the modern scientific literature. In addition, the resulting ontology and descriptions provides a foundation for consistent, unambiguous, computational representations of Fungi. Finally, we expect that such a system will enable more people to become active field mycologists by providing a more robust way to document field observations and connect those observations with information about similar fungi.

Introduction

The original motivation for the proposed system is to fill a gap in the existing naming systems for Fungi between scientific and common names, as well as between names based on evolutionary relatedness and names based on observed similarity. In systems such as the Mushroom Observer website (http://mushroomobserver.org), the breadth of the user community requires traversing this gap relatively seamlessly. Most languages use two types of names for Fungi - scientific names and common names. Scientific names are Latinized words or phrases such as "*Russula*", "*Chroogomphus vinicolor*", or "*Amanita muscaria* var. *guessowii* Veselý". Scientific names of Fungi are introduced by means of a formal peer-reviewed publication process and are governed by the *International Code of Nomenclature for algae, fungi, and plants* (ICN) (Knapp, 2011). Taxonomists express the relatedness between species by defining groups such as genera and families by evolutionary proximity. Common names or "vernaculars", on the other hand, are simply nouns or noun phrases for which there is no formal definition process. Common names have varying levels of acceptance among different groups interested in Fungi, and usage can vary by geographical region. Examples include "Pine Spike", "Kupferroter Gelbfuß", "Porcini" and "Pig's Ears". Common names are typically based on some rough, easily observed similarity between the things that the name

[1] nwilson@eol.org, Marine Biological Laboratory, Woods Hole, MA, USA
[2] dunnk2@rpi.edu, Rensselaer Libraries, Rensselaer Polytechnic Institute, Troy, NY, USA
[3] wangh17@rpi.edu, Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA
[4] dlm@cs.rpi.edu, Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

applies to.

A problem for field mycologists is that many scientific names cannot be accurately applied without careful microscopic work or, increasingly, genetic sequencing. On the other hand, common names for Fungi lack formal definitions, which results in essentially unresolvable ambiguities. For example, the term "Pig's Ears" is used both for *Gomphus clavatus* and *Discina ancilis*, two common edible species which have about as much in common evolutionarily and in appearance as humans and bumblebees. Homonyms can also occur with published scientific names, but in this case there are strict rules to determine which name has priority, or the names are for very different groups such as plants and animals, which are governed by different nomenclatural codes. In addition, it is not unusual for particular species, such as well known edibles, to have a number of common names in use by different groups of people while other less popular species may have no common names even if they common, widespread, and easily recognized.

There have been some efforts to formally define a set of common names for Fungi (Holden, 2003; Redhead, 2003) similar to the *Check-list of North American Birds* (Chesser, 2011) managed by the American Ornithologists' Union. These attempts have focused generally on associating common names of Fungi with a specific scientific name or at most a small collection of closely related species. These efforts have yet to produce a widely accepted result. Even if such a formalized set of common names were adopted, due to the close linkage between such common names and scientific names, there would still be cases that require the users to make distinctions that they are not likely to be able to do because, for example, the distinctions require using sophisticated equipment and/or technical expertise.

This paper proposes a new type of name called “semantic vernacular names”. Semantic vernacular names are distinctive terms that are defined with the assistance of the representational logic underlying the semantic web (http://en.wikipedia.org/wiki/Semantic_Web). Specifically, we encode our definitions in the Web Ontology Language (OWL). OWL is a computationally understandable and unambiguous ontology language that initially received recommendation status in 2004 (McGuinness and van Harmelen, 2004) and went through a revision process, reaching a new recommendation status for the expanded language OWL 2 in 2009. We refer to the combination of a semantic vernacular name and its defining properties as a “semantic vernacular definition”. The basis for a semantic vernacular definition are objective, observable features of the entities being described. As such they may also be thought of as named sets of evidence used to conclude that this name applies. In the case of Fungi, the types of evidence included in these descriptions include features such as the overall shape, presence or absence of particular structures, color or shape of such structures, taste, etc. Such observable features or evidence are explicitly made the basis for these definitions, in contrast to the definition of scientific names, which emphasizes evolutionary relatedness.

Motivating Examples

To understand the need for semantic vernacular definitions, consider the following examples based on observations submitted to the Mushroom Observer website (http://mushroomobserver.org):

1) The Common Pine Spike

Chroogomphus is a genus containing roughly 20 distinct species according to the Index Fungorum website (http://www.indexfungorum.org as of March 9, 2012). Three, *Chroogomphus rutilus*, *C. ochraceus*, and *C. vinicolor*, are commonly reported on the Mushroom Observer website. In practice it is very difficult to distinguish between these species using the naked eye. Historically, in the United States, all three names have been used. *C. ochraceus* has been applied to lighter, ochre colored collections; *C. vinicolor* has been distinguished primarily on the microscopic feature of the unusually thick cell wall of the pleurocystidia; and *C. rutilus* has been applied to the remaining darker collections that lack thick walled pleurocystidia (Arora, 1986).

Recent molecular evidence (Miller, 2003), reveals that *C. rutilus* is a European species that has yet to be verified as occurring in the US. The examined US material previously identified as *C. rutilus* because of its darker coloration was shown to belong to the surprisingly variable *C. ochraceus*. Based on this same study, *C. vinicolor* remains supported as a separate species.

What name should a field mycologist use for collections of these species, assuming they cannot put each collection under the microscope and may be collecting in an area that has not been surveyed molecularly? Any of the three scientific names given above could be accurate, but the collector has no way to decide which is appropriate. Typically in such a circumstance the genus name is applied and perhaps some notes are recorded. However, this loses the information that it was unlikely to be one of the 17 other species of *Chroogomphus*. It also loses the evidence on which the identification was based.

A semantic vernacular definition named "PikeSpike", for example, could avoid this loss of information. "PineSpike" would be defined by the macroscopic features common between these three species. Assuming that the other species of *Chroogomphus* are distinctive enough from these three, their exclusion would be captured. Furthermore, since the definition is explicitly defined by a set of features, it would be possible to explicitly suggest features to observe if those have not been explicitly stated as well as providing a method to verify that the user had observed all of the definitional features of a "PineSpike". An example formal definition for "PineSpike" is given in Appendix A.

2) The Spicy Red *Russula*

Russula is a very large genus containing hundreds of described species (http://www.indexfungorum.org as of March 9, 2012). The genus *Russula* is relatively easy to recognize in the field, but distinguishing species can be very time-consuming and often requires looking at a number of microscopic and chemical features (Hussey, 1855; http://en.wikipedia.org/wiki/Russula as of March 9, 2012). 44% of *Russula* observations on Mushroom Observer are simply recorded as "*Russula sp.*" with no species epithet. In comparison only 22% of observations of genus *Amanita* (which has a similar number of species) are recorded as "*Amanita sp.*".

Many *Russula* species have red caps, white stipes and a very spicy taste (Arora, 1986). In addition, there are many white capped *Russula* species that again have white stipes and a very spicy taste. Mushrooms with these groups of features are readily recognized and distinguished by field mycologists and as such would be logical groups for which to have names. However, red capped and white capped *Russula* species are not evolutionarily distinct. In fact there is known to be at least one species, *Russula cremoricolor*, that reliably produces either white capped or red capped mushrooms (Redecker, 2001) possibly even from the same

individual fungus (personal observation http://mushroomobserver.org/obs/2108). As a result, an evolutionarily-based scientific name cannot be used to distinguish spicy-tasting red-capped *Russula* from spicy-tasting white-capped *Russula.*

On the other hand, two semantic vernacular names, "SpicyRedRussula" and "SpicyWhiteRussula" could be defined. Again these terms would express what was actually observed by the collector rather than making any sort of implied claim regarding the evolutionary history of what the collector saw.

3) A Truffle by Any Other Name...

The term “truffle” here is used loosely for all of the many fungal species that have macroscopic, hypogeous (meaning underground) fruiting bodies. Truffles are a fascinating case of convergent evolution across almost all of the groups of mushroom-producing Fungi (Trappe, 2007). There has been a tradition of creating distinct genera for the hypogeous Fungi affiliated with many of the more familiar species that fruit above ground. Molecular studies have repeatedly shown that these distinctions are artificial from an evolutionary perspective (Miller, 2001; Peintner, 2001). However, these names are still very useful, especially to experts who focus on truffles, and these names continue to be used on Mushroom Observer. This is an easily recognized group for the field mycologist and at least some of the important distinctions can be made without a microscope. Unfortunately, it is very difficult to quantify the use of these hypogeous genera in Mushroom Observer since there is not a recognized mechanism for identifying these names in a consistent manner. Semantic vernacular names, such as "FungalTruffle" or "RussuloidTruffle", on the other hand, would provide a solution since their definitions would include their hypogeous growth habit.

The truffle example highlights another benefit of semantic vernacular definitions. Because the definitions are designed to be easily interpretable by computers, they will enable efficient searches for particular features such as growth habit or color . Furthermore, it will also be straightforward to associate non-definitional properties such as edibility, geographical range, and literature references with the definitions.

Discussion

The ability to search on properties like those proposed for semantic vernacular definitions has long been desired for biological species. A key issue is the lack of accessible, regularized data to enable such searching. One idea has been to attempt to extract such data from the historical literature for scientific names. Fortunately, much of the historical literature is in the process of being digitized through the Biodiversity Heritage Library (http://biodiversitylibrary.org/). However, even overlooking the major issue of accurate letter recognition from such scans, there are still very significant, unresolved challenges related to interpreting that text to accurately collect the data for a single species, dealing with conflicting information, and then regularizing that data in such a way that it can be efficiently searched.

The proposed semantic vernacular system provides an alternative route for generating searchable descriptions. The key is to base the definitions of these names on a standardized, yet expandable vocabulary. In the context of the semantic web, such a vocabulary is part of an “ontology”. A key component of developing the proposed naming system will be the creation and refinement of an ontology for describing Fungi with a focus on macroscopic, field-observable features.

Core to an ontology are the definitions of various properties including their range of allowable values (Noy and McGuinness, 2001; Hitzler, 2009). The information expressed through these properties will be familiar to anyone who has worked with character matrix keys. Semantic vernacular names will be defined through this sort of formal ontology. A sample definition of the term "PineSpike" created with the semantic web tool Protégé is provided in Appendix A.

Formal ontologies provide natural ways to connect the vocabularies to human readable definitions or even diagrams. They can also describe important constraints for automated reasoning, such as controlling how many values something can have for a given property, the type of the value, or even relationships between values. For example, the presence of a particular color may be required for the definition to apply, but another color may optional. Thus for a RedSpicyRussula, it might be that the cap must include some shade of red or pink, but white patches are permitted. However, the definition would not apply if a shade of blue is reported.

The core idea is that a semantic vernacular definition applies to anything that matches its computable definition. This is in contrast to scientific names, where the definition is based on a type, typically a specimen stored in a museum, along with a textual description that circumscribes an expected range of variation around the type. However, the evolutionary relationship between organisms and the type specimen is ultimately of higher importance than the original description to the application of the name over time. This has necessarily resulted in significant changes in the application of many scientific names over time.

The design of the semantic vernacular names is also important. The codes that govern scientific names strive to make them unique. Uniqueness makes terms much easier to use as indexes for related information and generally reduces confusion and avoids miscommunication. For use as indexes, uniqueness is simply required in the context of other published scientific names. The process of ensuring uniqueness within the set of scientific names has been greatly facilitated by the use of computers and various existing large databases of names. To help reduce confusion, it can be useful for terms to be distinctive particularly when they occur in text. Scientific names are made distinctive in text through the traditional use of italics. They are also made distinctive in both text and speech by the use of scientific Latin. However, the use of scientific Latin also makes them more difficult to learn and can present challenges for integrating these names into non-Latin-based text.

The semantic vernacular names are also expected to be unique. They will be developed from the beginning using globally connected databases and uniqueness within the context of other semantic vernacular names will be guaranteed. With regard to distinctiveness, semantic vernacular names are intended to be more familiar to their users and hence easier to learn. Semantic web technology supports full Unicode encodings so it is possible for the semantic vernacular names to use a very wide variety of character sets. In addition, semantic web technology make a clear distinction between the unique identifier for an entity and its "label". In the example, SV1234 vs. “PineSpike”@en. A common use of labels in semantic web technology is to provide one or more human-readable Unicode strings for entities in identified languages. This makes it easy to expand a semantic vernacular definition to have multiple language-specific semantic vernacular names for the same description where the names are explicitly associated with their intended languages. However, the semantic vernacular identifier would be unique for each description.

While familiarity is important, distinctiveness of the term is also desirable. This is particularly

important in written text. The initial proposal for labels that use Latin alphabets is to use camel case (http://en.wikipedia.org/wiki/Camel_case) where the terms are built from two or more reasonably familiar words in the given language. The first letter of each word is then capitalized and they are run together with no spaces. For example, "PineSpike" or "KupferroterGelbfuß".

It is important to recognize that the unique identifier (in the example, SV1234), is distinct from the label "PineSpike"@en, which represents the name in English for the semantic vernacular definition represented by the identifier. The unique identifier is technically a Uniform Resource Identifier or URI. While SV1234 may not seem like a good example of a universally unique identifier, there is an implied URL prefix that in the actual system would guarantee that it is unique. The registration system for semantic vernacular definitions will ensure not only that the identifiers are unique, but that the semantic vernacular names are unique within a particular language, region or country.

Once definitions like these are available in a database, it will then be relatively easy to provide a computer interface that allows users to not only look up definitions by their various semantic vernacular names, but to search the database for existing semantic vernacular definitions that match an observed specimen. Standard OWL reasoning systems can automatically determine hierarchical relationships when they apply. In addition, when someone claims to have collected something that matches a particular semantic vernacular definition, a web-enabled application can ask them to verify the key defining features that separate this definition from all others. This will also act as a way to explicitly collect the observed evidence on which the determination was made. This will improve the quality of observations and will encourage users to be more aware of what they are really saying when they apply a particular semantic vernacular name.

The semantic vernacular system is by no means intended to replace the traditional scientific naming systems. On the contrary, these systems should be viewed as complementary. In fact, making connections between them will be an essential aspect of the semantic vernacular system and will be critical to its success. For any given semantic vernacular description, there will be a list of scientific names that can match that description. For example, SpicyRedRussula and SpicyWhiteRussula would both be associated with *Russula cremoricolor*. These lists will necessarily be manually curated, but will provide an essential bridge between the two systems.

An important difference between the two systems is that the scientific naming system is rooted in the real world through its use of type specimens whereas semantic vernacular descriptions are strictly abstract. Being able to return to an explicit, defining example of a species can be critical to determining some key properties of a species, especially those related to the viability of that species' population such as genetic sequences. However, a consequence of a type-based definition is that application of a scientific name is riskier than application of a semantic vernacular name. Even a detailed comparison to the type may be incomplete or flawed. Furthermore, getting access to type specimens is often very difficult, especially for field mycologists outside of academia. In contrast, semantic vernacular definitions are easily communicated in their entirety and should prove much easier to work with when biological species level distinctions are not critical. Further, since semantic vernacular definitions are expected at least initially on features that are easily observed, these terms have significant potential for widespread adoption among less scientifically oriented groups.

Another way in which scientific names and semantic vernacular definitions complement each other is through the contrasting emphasis of phylogeny versus observable properties. While it is valuable to know the relatedness among species, phylogeny is not an observable feature of an organism. It is an ongoing matter of research and debate, which often leads to instability

in scientific names. The semantic vernacular system favors stability over relatedness. While it is very likely that some level of ambiguity and uncertainty will remain around the application of semantic vernacular definitions, the intent is that this will be reduced to human error and the diversity of life rather than a necessary side effect of the naming system.

In addition, scientific names are important for the connection they provide to the historical literature (Patterson, 2010). Scientific names have been in use for over 250 years and provide a key index into much of the scientific literature. However, access to and accurate comprehension of that literature is a significant challenge to anyone interested in Fungi, but most especially those who are just getting started in the field. A system which provides useful, stable, complete, well-documented definitions for names will be of significant value to beginners. The proposed system will not only achieve this goal, but will also provide links to related scientific names and through them a solid connection to the historical literature.

Conclusion

The core features of the semantic vernacular system are:

- Each semantic vernacular name has an explicit, fixed definition.
- Semantic vernacular definitions are based on features that can be directly observed. At least initially these features will focus on features that can be observed by the naked eye in the field.
- Semantic vernacular definitions can be associated with multiple language-specific semantic vernacular names.
- Semantic vernacular definitions are amenable to peer review and codification.
- Semantic vernacular identifiers are inherently unique.

Such a naming system will be of great value to those studying groups of organisms where making species level distinctions is particularly challenging. The semantic vernacular system also provides a concrete path for developing useful, computable definitions for groups of organisms that can ultimately extend to the species level by incorporating microscopic, chemical or genetic properties as needed. Finally, they provide a framework in which individuals recording observations of species can more easily explicitly state the evidence based on which they have applied a name.

Appendix A

Below is an example semantic vernacular definition for "PineSpike" in Manchester OWL Syntax. Note that this example is incomplete for the sake of brevity. Specifically, the URL prefixes are not included, and neither are the definitions of the non-literal terms. Contact the author for the complete OWL file if you would like to see it.

```
Class: SV1234

    Annotations:
        rdfs:label "PineSpike"@en
        rdfs:label "KupferroterGelbfuß"@de

    EquivalentTo:
      Fungus
          and ((hasPileusDiscColor some Brown)
                  or (hasPileusDiscColor some DarkBrown)
```

```
            or (hasPileusDiscColor some LightBrown)
            or (hasPileusDiscColor some Orange))
      and ((hasPileusShapeFromSide some Conic)
            or (hasPileusShapeFromSide some Convex)
            or (hasPileusShapeFromSide some Plane))
      and (hasHymenophoreShape some Gilled)
      and (hasOverallShape some StipitateAgaric)
      and (hasPileusFleshColor some Orange)
      and (hasSporePrint some Obtainable)
      and (hasSporePrintColor some Black)
      and (hasStipeFleshColor some Orange)
      and (hasStipeSurfaceColor some Orange)
      and (hasSubstrateAttachment some Stipitate)
      and (hasTaste some Mild)
      and (hasUniversalVeil some Absent)

SubClassOf:
    FungusDescriptiveVernacular
```

References

D. Arora. 1986. Mushrooms demystified: A comprehensive guide to the fleshy fungi, 2nd Edition. Berkeley: Ten Speed Press. pp. 83–103.

R. Terry Chesser, Richard C. Banks, F. Keith Barker, Carla Cicero, Jon L. Dunn, Andrew W. Kratter, Irby J. Lovette, Pamela C. Rasmussen, J. V. Remsen, James D. Rising, Douglas F. Stotz, Kevin Winker. 2011. Fifty-Second Supplement to the American Ornithologists' Union check-list of North American Birds. *The Auk* 128(3) 600-613. doi: 10.1525/auk.2011.128.3.600. Complete list available at http://www.aou.org/checklist/north/

P. Hitzler, M. Krötzsch, B. Parsia, P.F. Patel-Schneider, S. Rudolph. 2009. OWL 2 Web Ontology Language Primer. http://www.w3.org/TR/2009/REC-owl2-primer-20091027/

E. M. Holden. 2003. Recommended English Names for Fungi in the UK. Report to the British Mycological Society, English Nature, Plantlife and Scottish Natural Heritage. http://www.plantlife.org.uk/uploads/documents/recommended-english-names-for-fungi.pdf

A.M. Hussey (as Mrs. T.J. Hussey). 1855. Illustrations of British Mycology Containing Figures and Descriptions of the Funguses of Interest and Novelty Indigenous to Britain. http://biodiversitylibrary.org/page/2976863

S. Knapp, J. McNeill, and N.J. Turland. 2011. Changes to publication requirements made at the XVIII International Botanical Congress in Melbourne - what does e-publication mean for you?. *PhytoKeys* 6: 5–11. doi: 10.3897/phytokeys.6.1960

D.L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation. February 10, 2004. Available from http://www.w3.org/TR/owl-features/ .

O.K. Miller, Jr. 2003. The Gomphidiaceae revisited: a worldwide perspective. *Mycologia* 95: 176-183. http://www.mycologia.org/content/95/1/176.full

S.L. Miller, T.M. McClean, J.F. Walker, B. Buyck. 2001. A Molecular Phylogeny of the Russulales Including Agaricoid, Gasteroid and Pleurotoid Taxa. *Mycologia* 93(2) 344-354.

N.F. Noy and D.L. McGuinness. "Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001. http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf

D.J. Patterson, J. Cooper, P.M. Kirk, D.P. Remsen. 2010. Names are key to the big new biology. Trends Ecol. Evol., 25:686-691. http://dx.doi.org/10.1016/j.tree.2010.09.004

U. Peintner, N. L. Bougher, M. A. Castellano, J.-M. Moncalvo, M. M. Moser, J. M. Trappe, and Rytas Vilgalys. 2001. Multiple origins of sequestrate fungi related to Cortinarius (Cortinariaceae). Am. J. Bot. 88: 2168-2179. http://www.biology.duke.edu/fungi/mycolab/publications/PeintnerAJB.pdf

D. Redecker, T.M. Szaro, R.J. Bowman, T.D. Bruns. 2001. Small genets of *Lactarius xanthogalactus*, *Russula cremoricolor*, and *Amanita franchetii* in late-stage ectomycorrhizal successions. Molecular Ecology 10: 1025-1034.

S. Redhead. 2003. WWW Voting on Vernacular Names for Mushrooms. *Innoculum* 54(5) 4. http://msafungi.org/wp-content/uploads/Inoculum/54(5).pdf

M. Trappe, F. Evans, J. Trappe. 2007. Field Guide to North American Truffles: Hunting, Identifying, and Enjoying the World's Most Prized Fungi. Natural History Series. Ten Speed Press.

W3C. 2009. "OWL 2 Web Ontology Language Document Overview". http://www.w3.org/TR/owl2-overview/