Bottom-Up Ontology Building for Domains without Existing Standards

Steven S. Chong, Thomas Rodenhausen, Dongfang Xu, Hong Cui, P. Bryan Heidorn
University of Arizona

Abstract
Phenotypes are physical characteristics of organisms and are expressed with entities and qualities. Phenotypic characters are important for describing species, studying organismal function and understanding organismal evolution. However, existing phenotype descriptions are not amenable to computation and terms should be mapped to ontologies to disambiguate similar and dissimilar concepts. We discuss issues we encountered when building an ontology with a bottom-up approach. We also describe an upcoming user study to compare expert user performance on existing ontology building software with another system enhanced with features informed by our prior work. By combining a bottom-up ontology building approach with user studies guiding the design of low-barrier ontology building environments, ontology construction will become easier for domain experts and will encourage additional participation in the process, leading to more complete ontologies and greater use of ontologies in research data management.

Keywords: user-centered design; ontology; phenotypic character; user interface; user experiment


Copyright: Copyright is held by the authors.

Contact: stevenchong@email.arizona.edu, rodenhausen@email.arizona.edu, dongfangxu9@email.arizona.edu, hongcui@email.arizona.edu, heidorn@email.arizona.edu

1 Background
Phenotypes are the physical characteristics of organisms and expressed with entities (e.g. organs or parts that possess certain characters) and qualities (characters and/or character states possessed by entities). For example, leaves rounded is a phenotypic character, whereas leaves is an entity and rounded is a character state; the character, shape, is implied. Phenotypic characters are important for describing species, studying organismal function and understanding organismal evolution. Having computable and comparable phenotype data enables large-scale, data-driven research, such as studying trait evolution and its relationships to phylogeny and environmental/habitat changes (e.g. Zambe et al., 2014) and resolving taxon names through concept analysis with character-based evidence (e.g. Franz et al., 2015).

However, existing phenotype descriptions are not amenable to computation. Textual phenotype descriptions are continuously being published and when added to the massive number of past descriptions, they impose a major obstacle for phenotype data integration across life, inhibiting large-scale biological research. Due to the large terminology and expression variations in these natural language descriptions and lack of standards in describing taxa, the same character can be described differently in various sources while characters described similarly may not be homologous, making these descriptions unsuitable for computerized analyses. Transforming various natural language expressions into computable data requires a process, called ontologizing, where the semantics of varied expressions are made explicit and mapped to terms in an ontology (Mabee et al., 2007). Ontologizing ensures “apples are compared to apples” and lays the foundation for meaningful data integration and machine inference and reasoning.
2 Software Tools and Bottom-Up Ontology Building

Existing phenotype ontologies are incomplete and their coverage remain inadequate for curation tasks (e.g. Cui, et al. 2015). The development of software tools is needed to enhance the utility of taxon descriptions and support the construction of ontologies. Our prior research suggests that a bottom-up approach coupled with low-barrier ontology building environments will encourage scientists to participate in domain ontology construction, leading to more collaboration, less term disagreements and more complete ontologies.

We have developed the Exploring Taxon Concepts (ETC) Toolkit\(^1\), a suite of web-based software tools that converts published morphological descriptions into phenotypic character data that can be reused and repurposed. The Text Capture tool parses descriptions and marks up entities and their associated qualities with XML tags. One type of output from Text Capture is entity and quality phrases mined from organism morphological descriptions. These are good candidate terms for building a domain phenotype ontology.

The current Ontology Building tool provides a user-friendly interface that allows users to create taxon specific phenotype ontologies from the mined candidate terms. It displays the source and textual context for terms, a feature not found in other ontology editors (Figure 1). Phenotypic characters appear in natural language textual descriptions and because of a lack of standardization in the usage of phenotypic terms, access to the source materials (e.g. taxonomic description text) is needed to disambiguate the meanings of terms. This feature facilitates bottom-up ontology construction, an approach that aligns well with the literary warrant principle for controlled vocabulary construction (National Information Standards Organization, 2005). We have used the Ontology Building tool to construct an ontology for the plant genus *Rubus*, and this work allowed us to identify a set of issues that require consideration when building a high quality phenotype ontology:

a) **Non-specific structure terms**: These generic terms (e.g. apex, surface, and base) can be part of many different structures. Considering apex as an example, the pattern will create terms such as *leaf apex* and *leaflet apex*, while also making these specific apices subclasses of the general concept of apex.

b) **Modifiers**: Modifiers, including “degree modifiers” (such as weakly, greatly, slightly, and moderately), “frequency modifiers” (such as often, frequently, and rarely), and “coverage modifiers” (such as generally and mostly) need to be controlled. Modifiers are important in determining whether a family level quality can be inherited at a species level, but they have not received attention in existing ontologies.

c) **Terms with implied numerical meanings**: These terms are qualities that imply a numerical value on associated terms (e.g. *doubly serrate leaf* is a leaf that has two levels of serration). The numerical relationships have to be made explicit.

d) **Terms with multiple definitions**: Semanticizing requires ontologies to make the meaning of phenotypic terms explicit and “understandable” to computers. In an ontology, the meaning of a term is described in logical form. For example, *stout* can be defined as *strong* and *increased size*; that is, *stout* means *both strong* and *increased size*. *Strong* can be defined as a state (subclass) of *fragility*, which may also have other states such as *brittle*. Ontologies allow us to define *strong* and *brittle* as disjoint – meaning anything that is *strong* cannot be *brittle* at the same time. When these kinds of relationships among the terms are defined, ontologies can be used as an infrastructure to clarify the meanings for entity and quality terms, to support semantic similarity computation, and to automate logical reasoning. It is also clear to the description authors that if they need a term to describe *increased size* only, they should not use *stout*.

---

\(^1\) Exploring Taxon Concepts Toolkit: http://etc.cs.umb.edu/etcsite/start.html
e) **Unspecified relationships between related terms:** Some terms have relationships that are difficult to express in an ontology. For example, *cyme* and *cymiform* are clearly related; something that is *cymiform* has a shape that resembles a *cyme*. However, existing ontologies cannot handle this relationship.

![Figure 1: Screen capture displaying description context (bottom panel) in ETC Ontology Builder](image)

3 **User Interface Comparisons**

Based on our experience with using the Ontology Building tool, we have developed a second prototype version with a user interface (UI) that supports dealing with the ontology building issues described above. For example, the new Ontology Building tool can automatically detect non-specific structure patterns and offers suggestions for how to handle non-specific terms (Figure 2). We are not aware of other ontology editors having these features.

![Figure 2: The prototype system detects the user is attempting to add a substructure (apex) to multiple parent structures (leaf and leaflet), triggering the non-specific structure pattern.](image)

863
We will conduct a pilot experiment with six domain experts (e.g. 1 botanist, 2 botany post-docs, 3 botany graduate/undergraduate students) to compare the effectiveness of the current and prototype UIs. Research provides evidence that domain experts can effectively construct an ontology when the core structure is established (Jupp et al., 2012). The experts will build upon an incomplete plant ontology by following the bottom-up approach. Participants will be provided with a brief training period on ontology building (10 minutes). This pilot study will reveal which user interface is most helpful with assisting experts to create quality ontologies (ontology terms and relationships). This study will identify if level of user expertise affects user performance with ontology building (e.g. are experienced botanists better than students at defining relationships or will students more acquainted with digital technology perform better?). Qualitative data will be collected through questionnaires and interviews, while quantitative data will be obtained by several methods, including automated system logs to track user keystrokes and button clicks and eye-tracking techniques to ascertain cognitive effort, word familiarity and the conceptual difficulty of text passages and sections in the ontology software.

4 Conclusions

Our ultimate goal is to develop software that will improve the ontology building process for phenotypic data by domain experts. Not only will this result in more complete and consensus-based ontologies, it will also increase the semantic awareness of domain experts (especially in students) and influence their academic writing practice – for example, they will use more terms from ontologies in their phenotype descriptions. Key to achieving that goal is conducting usability studies that identify features needed and/or desired by domain experts. Employing a bottom-up ontology building strategy also allows users to take advantage of description source context and their own knowledge for disambiguating terms. By incorporating features that make the ontology building easier, scientists will better enjoy building domain ontologies, be motivated to participate in the construction process, and be encouraged to use ontologies when managing their own data.

5 References