ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Modeling Issues and Solutions: Building a Taxonomy from a Biology Textbook

A. Patrice Seyed[1], John Pacheco[2], Andrew Goldenfranz[3], Vinay Chaudhri[2]

[1]Department of Computer Science and Engineering, University at Buffalo, NY, USA
[2]SRI International, Menlo Park, CA, USA; [3]Fremont Union High School District, CA, USA

1 Introduction

Our task is to create a taxonomy from an AP Biology textbook's glossary terms [1] for Project Halo [2]. Project Halo's goal is to build a reasoning system capable of answering novel questions and solving advanced problems in a broad range of scientific disciplines. In support of this goal, the resulting taxonomy is to be used as a foundation for translating passages within the biology textbook into logical formulas on which a reasoning system will operate.

In order to bootstrap our task, we imported ~2400 glossary terms and definition strings from the textbook's electronic glossary into Collaborative Protégé in OWL format, as classes and comment strings. Our team consisted of biologists and KR specialists. We took an iterative approach, where the biologists of our team attempted initial classifications, restricted to the *subclass-of* relation, and were encouraged to add additional classes when they deemed it appropriate. As they proceeded, modeling issues were identified and discussed during workgroup sessions. The issues and the solutions that were implemented are as follows.

2 Results

2.1 Entity/Role Dichotomy

Initially the biologists of our team encoded classes for organic molecules both on the basis of their structure and on the basis of their function (see Figure 1). For instance, proteins and steroids are defined by their chemical composition. In contrast, hormones are defined by the function they perform, and there is overlap between **Steroid** and **Hormone**, i.e. some hormones are steroids while others are proteins.

Figure 1. Naïve Classification of Steroids and Hormones

As a solution, we define hormones as roles that certain chemicals play. However, **Steroid-Hormone** remains a class in the taxonomy, which represents a useful class of biologists' intuitive thinking (see Figures 2 and 3).

Figure 2. Classification of Hormone Roles

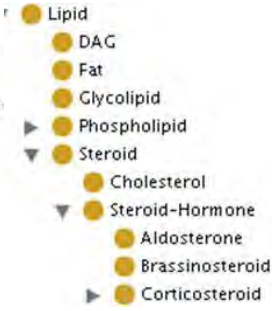

Figure 3. Classification of Steroids Based on Molecular Structure

2.2 Linnaean Classification

The biologists of our team wanted to classify the different kingdoms under the class **Kingdom** (see Figure 4). However, there are 5 instances of **Kingdom** (in the context of a U.S.-based textbook).

Figure 4. Naïve Classification of Kingdoms

As a solution, we represent Linnaean taxonomy under organism, and used common English names for simplicity (see Figure 5). For example, "Cow is an Animal" is clearer than "Cow is an Animalia":

Figure 5. Classification of Organisms

Figure 6. Treatment of Classification Units

As potential refinements, we can add the Latin-named classes as instances of their classification unit (see Figure 6). For example, **Animalia** is an instance of **Kingdom**, and **Chordata** is an instance of **Phylum**. As yet another approach (not pictured), we can treat classification units as meta-classes. For example, **Chordate** is an instance of the meta-class **Phylum**, and **Animal** is an instance of the meta-class **Kingdom**.

2.3 Entity/Process Dichotomy

The biologists of our team wanted to classify **Light-Microscope** under the subclass **Technology** (see Figure 7). They also wanted to classify **Technology** under the subclass Inquiry. These two uses of the term 'Technology' refer to two different senses. The glossary definition for Technology is "The application of scientific knowledge for a specific purpose, often involving industry or commerce but also including uses in basic research."

Figure 7. Naïve Classification of Technology

Our solution in this case was to refactor the taxonomy (see Figures 8 and 9). We noted that some terms of the glossary are polysemous. Definitions including "; also" were a clear indicator of this. For example, Wild Type is "An individual with the phenotype most commonly observed in natural populations; **also** refers to the phenotype itself."

Figure 8. Classification of Processes

Figure 9. Classification of Artifacts

2.4 Classifying Areas of Research

The initial tendency was to classify areas of research (e.g. **Genetics**, **Anatomy**, **Ecology**) under **Inquiry**. Areas of research are complex social entities, involving research activities and educational institutions constituted of departments, faculty members, programs and curricula. However, the definitions of the terms for each area of research is prefixed by "the scientific study of". Given the commitment to processes, their classification under **Inquiry** is appropriate.

2.5 Subclass/Subprocess Dichotomy

There was a strong initial tendency to use the hierarchy to organize sub-parts or sub-processes (see Figure 10). For example **Telophas**e is a subclass of **Mitosis**, instead of a sub-process.

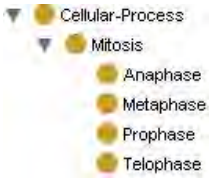

Figure 10. Naïve Classification of Processes

Our solution was to move parts or sub-processes to appropriate locations whenever they are found (see Figure 11). During workgroup sessions, we reinforced how to use the subclass of relationship consistently.

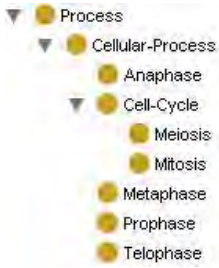

Figure 11. Refactoring of Processes

3 Conclusions

Initially, the biologists of our team relied on prior knowledge and definitions for organizing the class hierarchy, and the classes were treated as organizational "buckets". Ontological principles were iteratively applied to identify modeling issues and provide a foundation for the taxonomy building process.

After following this process for several workgroup sessions, the biologists had a much better sense for these types of modeling issues, and hence were more effective in continuing the taxonomy building process. These lessons and resulting taxonomy can help AURA better answer "What is" questions. Furthermore, these lessons can be applied to other ontologies, although it may depend on what formalism is used (e.g., for dealing with meta-classes).

Acknowledgements

This work was funded by Vulcan Inc.

References

1. Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, Robert B. Jackson. (2010) Campbell Biology, Pearson Publishing.
2. Gunning D. et. al. (2010) Project Halo Update – Progress Toward Digital Aristotle, AI Magazine, Fall 2010, 33-58.