Formal Ontology in Information Systems
P. Garbacz and O. Kutz (Eds.)
IOS Press, 2014

doi:10.3233/978-1-61499-438-1-415

KB_Bio_101: Content and Challenges

Vinay K. CHAUDHRI [a], Daniel ELENIUS [b], Sue HINOJOZA [a], and Michael WESSEL [a]
[a] *Artificial Intelligence Center, SRI International, Menlo Park, CA 94025*
[b] *Computer Science Laboratory, SRI International, Menlo Park, CA 94025*

Abstract. KB_Bio_101 contains knowledge about processes and mechanisms, and was created from an introductory textbook in biology. We give an overview of its content, summarize the key concepts represented, and give some examples of problems requiring further ontology research.

Introduction

The knowledge representation used in KB_Bio_101 contains many of the standard features such as classes, individuals, class-subclass hierarchy, disjointness, slots, slot hierarchy, necessary and sufficient properties, and Horn rules. By expressing this representation in first-order logic with equality [10], we have translated it into multiple formats including, OWL2 functional and an answer set program. As explained in [5], the translation of KB_Bio_101 into OWL is lossy. These translations are available through our website [10], and an OWL version is available through Bio Portal [4]. To facilitate an inspection of its content, especially its graph structured existential rules, we have also made available the graphical view of the concepts as seen through our system on a public website [10]. Because these graphs were generated through an automatic screen-capture process, in some cases, their layout and cropping is not optimal. We hope their availability will help the community to better understand the content of KB_Bio_101.

KB_Bio_101 has several innovative features, and here we highlight a few that make it valuable to the research community. First, biologists directly authored KB_Bio_101 ensuring that the knowledge is correct and understandable. Second, KB_Bio_101 underwent extensive testing during its development. The biologists assembled a test suite of more than 2000 competency questions and executed them against the KB to ensure that it gave correct answers to more than 90% of the queries. Third, KB_Bio_101 underwent substantial end-user testing by students as students accessed it through an electronic textbook called Inquire Biology [2]. This testing showed that KB-enabled Inquire Biology improved student learning demonstrating that the knowledge represented in it is practically useful for a learning application. Fourth, in relation to other ontologies such as Gene Ontology [7] and the Foundational Model of Anatomy [9], KB_Bio_101 offers unprecedented richness and complexity in many of its represented concepts. This richness and complexity results both from the expressiveness of the representation language and the number of semantic relationships used. Further, KB_Bio_101 covers the

full range of biological concepts instead of limiting itself to only genes or anatomy. Finally, KB_Bio_101 was created through at least 12 person years of effort by biologists and 5 person years of effort by knowledge engineers. As most academic researchers cannot readily undertake such a large scale knowledge engineering effort, KB_Bio_101 offers a platform for a range of research topics such as ontology evaluation, ontology modularization, ontology mapping, and most importantly, novel ontology design.

In this paper, our objectives are to (1) give an overview of the KB's content; and (2) highlight some open research challenges in creating this KB. The content of the KB builds on an upper ontology called Component Library (CLIB) [1]. The current KB contains knowledge from chapters 2-22, 36, 41 and 55 of the introductory biology textbook [8], but only chapters 2-12 have been subjected to an educational utility evaluation. Therefore, our description will be limited only to the content from chapters 2-12.

1. Chapter-Specific Content of KB_Bio_101

KB_Bio_101 contains more than 5500 classes and more than 100,000 axioms. Therefore, covering all the axioms in a single paper is impractical. For each textbook chapter, the biologists identify *key* concepts that are central to that chapter. Each chapter typically includes 100-150 key concepts, thus, giving more than 1100 key concepts for chapters 2-12. Because viewing 1100 concepts is a large task, we have identified a few focal concepts for each chapter below. When reading the electronic version of the paper, you may click on each concept name to navigate to an online visualization of that concept on our website [10].

1.1. Chapter 2: Chemical Context of Life

This chapter contains information about the Atom, Molecule, atomic units (e.g., Orbital, Atomic-Nucleus, Valence-Shell, and Electron-Shell); subatomic particles (Electron, Neutron, and Proton); isotopes of atoms; and basic types of chemical reactions (Forward-Reaction and Reverse-Reaction). The chapter also covers bonds such as Ionic-Bond, Covalent-Bond (both polar and non-polar) and Hydrogen-Bond. Chemical bonds provide foundational representation for many of the later chapters such as chapters 4, 5 and 9.

1.2. Chapter 3: Water and Life

The main concepts for this chapter are Water and Water-Molecule. The representation of Water-Molecule reuses the representation of bonds introduced in chapter 2: each Water-Molecule *possesses* two Polar-Covalent-Bonds that holds together the Hydrogen atoms; each Water substance contains a Hydrogen-Bond between Water-Molecules. The typical features of water that are identified in this chapter are: (a) cohesive behavior; (b) ability to moderate temperature; (c) expansion upon freezing; and (d) versatility as a solvent. We represented cohesive behavior by asserting that different Water-Molecules in Water attract each other. We modeled ability to moderate temperature by asserting that the attraction between

Water-Molecules *inhibit*s the Increase or Decrease in temperature. We modeled its expansion upon freezing by creating a process Freezing-Of-Water in which its density reduces. Water's versatility as a solvent was modeled to the extent that the KB contains a variety of Aqueous-Solutions.

1.3. Chapter 4: Carbon and Molecular Diversity of Life

Key concepts in this chapter include Carbon, Hydrocarbon-Molecule, Organic-Molecule, Carbon-Skeleton, Functional-Group and Isomer. These concepts also significantly rely on the representation of chemical bonds introduced in chapter 2.

1.4. Chapter 5: The Structure and Function of Large Biological Molecules

The key concepts for chapter 5 are different types of macromolecules essential for life: Protein (related concepts: Polypeptide, Amino-Acid), Carbohydrate (related concepts: Polysaccharide, Monosaccharide), Lipid (related concepts: Fatty-Acid, Fat, Steroid, Phospholipid), and Nucleic-Acid (related concepts: Nucleotide, DNA, RNA). The representations of these concepts build on the modeling decisions about Functional-Group and Carbon-Skeleton from chapter 4, and also reuse representations of Peptide-Bond, Glycosidic-Bond, and Phosphodiester-Bond from chapter 3. This chapter also covers some processes such as Polymer-Synthesis, Polymer-Breakdown and Protein-Denaturation.

1.5. Chapter 6: A Tour of the Cell

The key concepts in this chapter are many of the subclasses of Cell (e.g., Eukaryotic-Cell, Plant-Cell), and the different organelles that define its structure (e.g., Nucleus, Ribosome, Lysosome, Endoplasmic-Reticulum, etc.).

1.6. Chapter 7: Membrane Structure and Function

The key concepts in this chapter are Biomembrane, its constituents (e.g., Phospholipid-Bilayer, Membrane-Protein, etc.), membrane properties (e.g., *permeability*, *fluidity*, etc.), and different processes involving membranes (e.g., Active-Transport, Passive-Transport, Osmosis Diffusion, etc.). The representations in this chapter rely on Lipids and Proteins from chapter 5, and the cell structure from chapter 6.

1.7. Chapter 8: An Introduction to Metabolism

The key concepts in this chapter include ATP-Cycle, Metabolism, Exergonic-Reaction, Endergonic-Reaction, Spontaneous-Change, Catabolic-Pathway, Anabolic-Pathway, Cellular-Work, Enzyme and Enzymatic-Reaction. We modeled Cellular-Work and its subclasses. We also modeled Enzyme-Regulators and their involvement in processes such as Competitive-Inhibition. A key to representing these concepts was to using roles such as Competitive-Inhibitor and Noncompetitive-Inhibitor.

1.8. Chapter 9: Cellular Respiration and Fermentation

The key concepts in this chapter are Redox-Reaction, Cellular-Respiration and its steps (i.e., Glycolysis, Pyruvate-Oxidation, Citric-Acid-Cycle and Oxidative-Phosphorylation), Anaerobic-Respiration and Fermentation. This chapter heavily relies on the representations of Metabolism introduced in chapter 9. Processes such as Glycolysis are highly complex with numerous steps and participants. Especially for Glycolysis, we factored its representation into smaller chunks by separating it out into Energy-Investment-Phase-Of-Glycolysis and Energy-Payoff-Phase-Of-Glycolysis each of which was quite complex. The chapter also contains information on how Glycolysis is regulated which is not modeled in the current KB.

1.9. Chapter 10: Photosynthesis

We modeled in detail the steps of Photosynthesis: Light-Reaction and Calvin-Cycle and various entities that play key role in Photosynthesis such as Light, Chloroplast, Chlorophyll, Photosystems, and Electron-Transport-Chain. The representations in this chapter rely on representation of pathways from chapter 8.

1.10. Chapter 11: Cell Communication

The key concepts for this chapter are Cell-Communication and its subclasses (Communication-By-Direct-Contact, Cell-Communication-With-Mating-Factor-A, Cell-Communication-Leading-To-Apoptosis, etc.), Cell-Signaling and its subclasses (Cell-Signaling-With-Mating-Factor, Cell-Signaling-With-Receptor-Tyrosine-Kinase), the steps of Cell-Signaling (Signal-Reception, and Signal-Transduction); and the biological processes related to Cell-Communication (e.g., Mating-of-Saccharomyces-Cerevisiae and Limb-Development). Modeling of these concepts relies on the relationships between Membrane-Protein and Plasma-membrane from chapter 7, Cell-Cell-Junction from chapter 6 and Glycogen-Breakdown from chapter 4. This chapter is one of the most complex chapters among the first twelve chapters presenting various examples of the cell communication processes and knowledge about interactions between different communication agents.

1.11. Chapter 12: The Cell Cycle

The key concepts in this chapter include Cell-Division, Binary-Fission, Animal-Mitotic-Cycle, Plant-Mitotic-Cycle, Mitosis, Interphase. The knowledge in this chapter relies on numerous aspects of the structure of Cell such as Chromosome, Nucleus, Microtubule, Centrosome and Centriole.

2. Conceptual Modeling Challenges

For many kinds of the knowledge found in the textbook, our upper ontology provided the needed background knowledge in terms of general classes and relationships. For some sentences, however, the necessary background knowledge was unavailable. We did not want to allow proliferation of new relations to keep

the vocabulary small and simple so that the biologists could use it with minimal difficulty and training. The resulting challenges and gaps in the background knowledge can be described from two perspectives: from a top down perspective of core themes of biological knowledge, and a bottom up perspective of specific conceptual modeling problems.

2.1. Challenges Driven by Core Themes in Biology

In the United States, the College Board is responsible for standardizing introductory college level curriculum. It defines eight core themes in biology each signifying a major area of biological knowledge. These core themes include: structure and function, energy transfer, regulation, continuity and change, science as a process, evolution, interdependence in nature, and finally, science, technology, and society [6]. Each core theme requires novel ontology design and research of the sort that we have undertaken for energy transfer and regulation [3]. We consider here two of the core themes that we have not yet covered in our work.

The core theme of science technology and society is devoted to providing a broader context of science, and describing how various concepts introduced in the textbook have social relevance. This core theme overlaps with recent interest on modeling socio-technical systems. An example of a textbook sentence that highlights the knowledge that needs to be modeled is: *Many ecologists believe that this effort suffered a major setback in 2001, when the United States pulled out of the Kyoto Protocol, a 1997 pledge by the industrialized nations to reduce their CO2 output by about 5% over a ten-year period.* This sentence covers knowledge about the social system of countries and how they make agreements and pledges to deal with the technological problem of pollution.

The core theme of science as a process concerns itself with describing experiments, linking evidence, data, and studies with a conclusion or theory. It presents alternative models for various phenomena, for example, a sandwich model for membranes, a multi-step model of Cancer development, etc. Some example sentences that illustrate such information are: *Researchers wondered whether a cell's progression through the cell cycle is controlled by cytoplasmic molecules; The researchers concluded that molecules present in the cytoplasm during the S or M phase control the progress to those phases.* These sentences require representing hypotheses and conclusions.

2.2. Challenges Driven by Specific Modeling Problems

Processes, defaults, negation, causality, etc. are well-known and actively researched problems in both knowledge representation and ontology research. Although these issues arose when we created `KB_Bio_101`, the real challenge resided in reducing a piece of biological knowledge to a representational technique such that we could either properly apply it or formulate a necessary specific extension. We consider a few examples below.

Most common forms of causal relationships involve two events, and considerable prior work exists on representing them. Many example sentences, however, introduce causal relationships between a Property and an Event, between an Entity

and an Event, and between a structural arrangement and an Event. Some example sentences are as follows: *Because of the high specific heat of water relative to other materials, water will change its temperature less when it absorbs or loses a given amount of heat; Because electrons have a negative charge, the unequal sharing of electrons in water causes the oxygen atom to have a partial negative charge and each hydrogen atom a partial positive charge; The electrons of an atom also have potential energy because of their position in relation to the nucleus.* The literature is lacking on how to appropriately capture such knowledge in a conceptual model.

In the textbook, many sentences address variation within a species/entity, but give no specifics about it. Our discussions with biology teachers revealed that capturing the concepts of diversity and variation is important (i.e., these concepts capture important abstractions about the knowledge). Some example sentences follow: *Energy exists in various forms; In interphase, the relative durations of G1, S, and G2 may vary; The cell division frequency varies with the type of cell.* This challenge potentially overlaps with the recent interest in capturing biodiversity.

2.3. Summary

KB_Bio_101 represents a major advance in the construction of large and complex conceptual representations. Its translations in standard formats make it a valuable data set for ontology management research, and an excellent starting point for developing novel conceptual representations.

Acknowledgment

This work has been funded by Vulcan Inc. and SRI International.

References

[1] K. Barker, B. Porter, and P. Clark. A library of generic concepts for composing knowledge bases. In *First International Conference on Knowledge Capture*, 2001.

[2] Vinay K Chaudhri, Britte Cheng, Adam Overholtzer, Jeremy Roschelle, Aaron Spaulding, Peter Clark, Mark Greaves, and Dave Gunning. *Inquire Biology*: A textbook that answers questions. *AI Magazine*, 34(3):55–72, 2013.

[3] Vinay K. Chaudhri, Nikhil Dinesh, and Stijn Heymans. Conceptual models of energy transfer and regulation. In *Proceedings of International Conference on Formal Ontologies in Information Systems*, 2014.

[4] Vinay K. Chaudhri, Stijn Heymans, and Michael A. Wessel. The KB_Bio_101 Page in BioPortal, 2014. See http://bioportal.bioontology.org/ontologies/AURA.

[5] Vinay K. Chaudhri, Michael A. Wessel, and Stijn Heymans. *KB_Bio_*101: A challenge for OWL reasoners. In *The OWL Reasoner Evaluation Workshop*, 2013.

[6] College Board. Biology: Course description. http://apcentral.collegeboard.com/apc/public/repository/ap-biology-course-description.pdf, 2010.

[7] The Gene Ontology Consortium. Gene Ontology: Tool for the unification of biology. *Nat Genet*, 25:25–29, 2000.

[8] Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Robert B. Jackson. *Campbell biology.* Benjamin Cummings imprint of Pearson, Boston, 2011.

[9] Cornelius Rosse and José LV Mejino Jr. A reference ontology for biomedical informatics: The Foundational Model of Anatomy. *Journal of biomedical informatics*, 36(6):478–500, 2003.

[10] Michael Wessel. The AURA KB Translations — FOPL, TPTP, ASP, SILK, and OWL2, 2013. See http://www.ai.sri.com/~halo/public/exported-kb/.