The Cyc Knowledge Server
CMSC828D Report 1

Cody Buntain
Department of Computer Science
University of Maryland

September 16, 2012

1 Cyc/OpenCyc Overview

In 1984, Dr. Douglas Lenat set out to construct an intelligent system capable of automated, common-sense reasoning and natural language understanding through the capture and formalization of general, real-world information. This effort would eventually become the Cyc Knowledge Server™ [2], which has since grown into an extensive knowledge base, formal language, and inference system supported and marketed by the Cycorp corporation. Cycorp has since spun off the OpenCyc initiative, which enables free and open access to a limited version of the Cyc knowledge base and the use of its related tools, to foster market growth and third party contribution to its knowledge base.

To give a better idea of Cyc's purpose, one need only consider the origin of its name: en*cyc*lopedia. Cyc's goal is to capture enough broadly known, or common-sense, information to *understand* the contents of a given encyclopedia. That is, Cyc should know the information an encyclopedia's author considers too well-known to be worth including in an entry but is still essential in understanding what that entry means. Cyc's authors consider this widely accepted and well-known information as the foundations of "common-sense" knowledge. One should note, however, that this goal is significantly different from the goal of knowing all of the information within an encyclopedia (though Cyc is outfitted with a significant amount of in-depth knowledge in a variety of areas like biology, materials science, and others).

If one is wondering why someone would dedicate a significant effort to *understanding* the content of an encyclopedia instead of *memorizing* its content, one should note that Cyc's value follows from its extremely broad knowledge of common-sense reality. Armed with this extensive knowledge, Cyc-powered systems are capable of learning and inferring new information with minimal additional explanatory effort. In fact, one could consider the Cyc system to be a realization of McCarthy's *advice taker* program in that Cyc leverages its common-sense knowledge to ingest and reason over statements without requiring an in-depth explanation of what is meant by those statements.

As an example of this sort of interaction, a user could tell a Cyc-powered system of the recent acquisition of a pet name "Penny." Based on the information *already* encapsulated within Cyc's core, the system would then be able to infer a great deal about Penny in much that same way that a human would. For instance, the system would infer that Penny is a living thing, a domesticated animal, and not a human as well as several other characteristics about Penny's size, biology, and agency.

To accomplish such common-sense inference, Cyc is composed of three primary components: the core Cyc knowledge base (or ontology), the CycL language specification in which the knowledge base is written, and the Cyc Inference Engine. The Cyc knowledge base is the ontological database that embodies Cyc's knowledge about the general world, the tangible and intangible things that populate it, and how these things are related. To formalize these relations and ensure machine readability, the knowledge base is written in

the custom CycL formal language. This language is then parsed and reasoned over by the inference engine to identify and assert new information. By its very nature, however, the information in Cyc's knowledge base (either taught or inferred) must continue to grow if it is to continue to be useful on a broad scale.

2 Information Content and Maturity

As mentioned, Cyc's need for growth ensures it will never be a complete representation of common-sense knowledge. The OpenCyc initiative exists to address this exact need by enabling a crowd source-like approach to expanding its knowledge. This growth is not limited to purely general knowledge and in fact contains much more thorough knowledge on a broad number of topics. For instance, Cyc's core contains information on biology, geography, geo-political entities, literature, materials, food, weather, belief, attitude, and emotion among others [4].

Since this core contains the essence of the encapsulated common-sense information, it is worth understanding its form and content. One can visualize Cyc's knowledge base as a pyramid in which generality increases with height. At its apex, the knowledge base contains the "upper ontology," a formalization of very general concepts and their relations (e.g., there exists a thing called an "Event" that takes place at some specific instance in time, so it is also a "Temporal-Thing"). At this level, one is introduced to Cyc's most fundamental representation: the "Thing." Every item in the entire knowledge base is either a specialization or instance of a "Thing."

Below this upper ontology is the set of Cyc's "core theories" about space, time, and causal relationships. This level begins to build the rules and inferences necessary for common-sense reasoning. One such example of these core theories at work is as follows: if a was caused by b then b must precede a. Following these core theories are the many domain-specific theories that have been built in to Cyc over the years. While these rules are not necessary for common-sense reasoning, they do greatly increase the utility of the system in that they cover a broad number of domains (finance, epidemiology, network topology, and other interactions) and allow the system to reason about specific areas of interest without significant exposition.

Finally, at the base of the pyramid lies the ground truth set of facts. Up until this point, the previous levels dealt exclusively with rules (i.e., theories), but here are the actual facts that populate the world. These facts consist of assertions like "George W. Bush is an individual and a previous president of the United States but is not the current president of the United States" or "Moby Dick is a white sperm whale from the context of the American novel." It is this vast collection of "general knowledge" about the world at large that provides Cyc with the context necessary to process and understand additional knowledge imparted to it on a larger scale.

2.1 Microtheories – Focus, Brevity, and Inconsistency

Cyc comes complete with a large collection of constructs called "microtheories" that allow the user to focus the knowledge base into a specific domain, knowledge area, or time [3]. Microtheories are basically collections of possibly overlapping assertions with the caveat that the assertions present within a single microtheory be mutually consistent. In this way, the microtheory construct enables more compact assertions without having to worry so much about side effects with other assertions. Furthermore, by separating knowledge data into these encapsulated microtheories, Cyc helps avoid inconsistencies in term usage across domains (e.g., "domain" in the computer-specific microtheory versus "domain" in the geo-political sense).

2.2 OpenCyc, Cyc, Size, and the Semantic Web

It is true that the line between the free OpenCyc implementation and the proprietary Cyc knowledge base is rather nebulous. It seems that OpenCyc contains a very large percentage of the data from Cycorp's Cyc commercial version save for a small number of proprietary symbols. Be that as it may, OpenCyc's current version (release 4.0) of the Cyc core knowledge base contains on the order of hundreds of thousands of terms

and millions of relations for defining and relating tangible and intangible things (events, animals, concepts, etc.) [7].

Beyond simply growing the Cyc knowledge base, OpenCyc is pushing to keep the Cyc system relevant in the context of the Semantic Web. As one of Berners-Lee's goals of the Semantic Web is to enable understanding within and across automated systems, Cyc fits very well into the Semantic Web paradigm. As such, the current OpenCyc release contains more than 60,000 links to other non-Cyc knowledge bases [7]. These other resources include WordNet, the CIA World Factbook, and other ontology structures like DBPedia. This Semantic Web crossover is further enabled by OpenCyc's new Ontology Web Language (OWL) support, the W3C standard for knowledge transfer and alignment between Semantic Web resources.

3 Interacting with the Knowledge Base

One may now be asking one's self how one really makes use of the Cyc system. Fundamentally, Cyc supports only two forms of interaction: imparting information into the system (via the assertion of facts or inclusion of inference rules), and requesting information from the system (in the form of queries). These interactions can be accomplished through one of two ways: either direct interaction with the Cyc knowledge base through the Cyc Knowledge Browser and the CycL formal language, or through the Cyc developer APIs.

With the CycL language, a user can ask questions like, "What types of things are cats?" with the query `(#$genls #$Cat ?X)`, which basically requests all the collections of which `Cat` is a specialization (or what collection `?X` generalizes `#$Cat`). The CycL language is relatively large and encompasses a lot of expressive power, so one should refer to [1] for more information. If one does not have a mind to learn the CycL language, however, one can still interact with the system using the Java-based OpenCyc API [10].

4 Additional Resources

For further reading on the Cyc and OpenCyc systems, see these resources: [1, 7, 5, 9, 6, 8].

References

[1] *Cyc 101 Tutorial*. Cycorp Corporation, http://opencyc.org/doc/tut, 2002.

[2] About cycorp. Webpage, Cycorp Corporation, http://cyc.com/cyc/company/about, September 16 2012.

[3] Cycorp. Foundations of knowledge representation in cyc – microtheories. In *Cyc 101 Tutorial*. Cycorp Corporation, http://www.cyc.com/doc/tut/ppoint/Microtheories_files/v3_document.htm, 2002.

[4] Cycorp. Survey of knowledge base content. In *Cyc 101 Tutorial*. Cycorp Corporation, http://www.cyc.com/doc/tut/ppoint/MoreContentAreas_files/v3_document.htm, 2002.

[5] Cycorp. Technical report, Cyc.com, http://www.cyc.com, 2012.

[6] OpenCyc. Opencyc faq. Technical report, OpenCyc.org, http://www.opencyc.org/faq/opencyc_faq, 2002.

[7] OpenCyc. Webpage, OpenCyc.org, http://www.opencyc.org, 2012.

[8] OpenCyc. Opencyc documentation. Technical report, OpenCyc.org, http://opencyc.org/doc, 2012.

[9] OpenCyc. Opencyc introduction. Technical report, OpenCyc.org, http://www.opencyc.org/cb/welcome, 2012.

[10] OpenCyc. Opencyc java api. Technical report, OpenCyc.org, http://www.cyc.com/doc/opencyc_api/java_api/, 2012.