Protégé, Ontology and Knowledge Acquisition: Knowledge Representation, The Foundation of Intelligent Systems

By Terry Hengl

Introduction

Knowledge is a decisive competitive-advantage for today's corporations. Knowledge of schedules, raw materials, labor, manufacturing and distribution is essential to the supply chain while knowledge of customer interests and buying habits, latest technologies, budget constraints, marketing plans are crucial to product development. It offers a powerful tool for gaining market share and preserving a competitive edge, but it is costly to capture and control. Methodologies and technologies that assist in the acquisition, maintenance, and distribution of knowledge are essential to an organization's success. Today's society, and the world in general, have contributed to this growth in importance of knowledge management and knowledge based systems. Some examples include (*www.worldedreform.com/intercon2/f15a.pdf*):

* Accelerating rate of change in every aspect of technology and society
* Staff migration and attrition (downsizing and reengineering)
* Geographic dispersion associated with globalization of markets
* Global integration of cultures, companies and markets
* Increase in networked organizations
* Increased level of education and training of the population
* Growing knowledge-intensity of goods and services
* Revolution in information technology

It is not easy to efficiently and cost-effectively identify, acquire and maintain this knowledge. At a minimum, organizations must be able to:

* agree to an organization-wide vocabulary to ensure knowledge is consistently communicated and understood;
* identify, explicitly represent and model this knowledge;
* share and reuse this knowledge across independent applications and domains.

This article looks at these aspects of knowledge acquisition by examining Protégé (*http://Protégé.stanford.edu*), a free open-source Java tool with an extensible architecture for creating customized knowledge-based applications - based on Ontologies (*www.ontology.org*). It also reviews the concept of Ontologies (see side bar), and associated support methodologies, which establish the vocabulary and model the concepts along with their inter-relationships. This concept also includes processing of the associated attributes for a particular field of knowledge. By reviewing the evolution of Protégé, an ontology modeling and knowledge acquisition environment (developed by Stanford Medical Informatics (*http://camis.stanford.edu*) at the Stanford University School of Medicine), an understanding of fundamental concepts that underpin knowledge acquisition emerges. This environment creates and modifies the ontologies and knowledge bases it generates to enable developers and domain experts to build knowledge-based systems.

Other associated technologies and standards mentioned in this article include the Open Knowledge Base Connectivity (OKBC), Generic Frame Protocol (GFP), Resource Description Framework (RDF), OWL and OWL-S.

The Evolution of Protégé

Opal (*http://smi-web.stanford.edu/pubs/SMI_Abstracts/SMI-86-0137.html*) was an expert system shell-based application developed as part of the medical domain Oncocin (*http://citeseer.ist.psu.edu/context/1419258/0*). Oncocin developed this knowledge acquisition and advice system for protocol-based cancer therapy. Opal enabled patient history entry by the

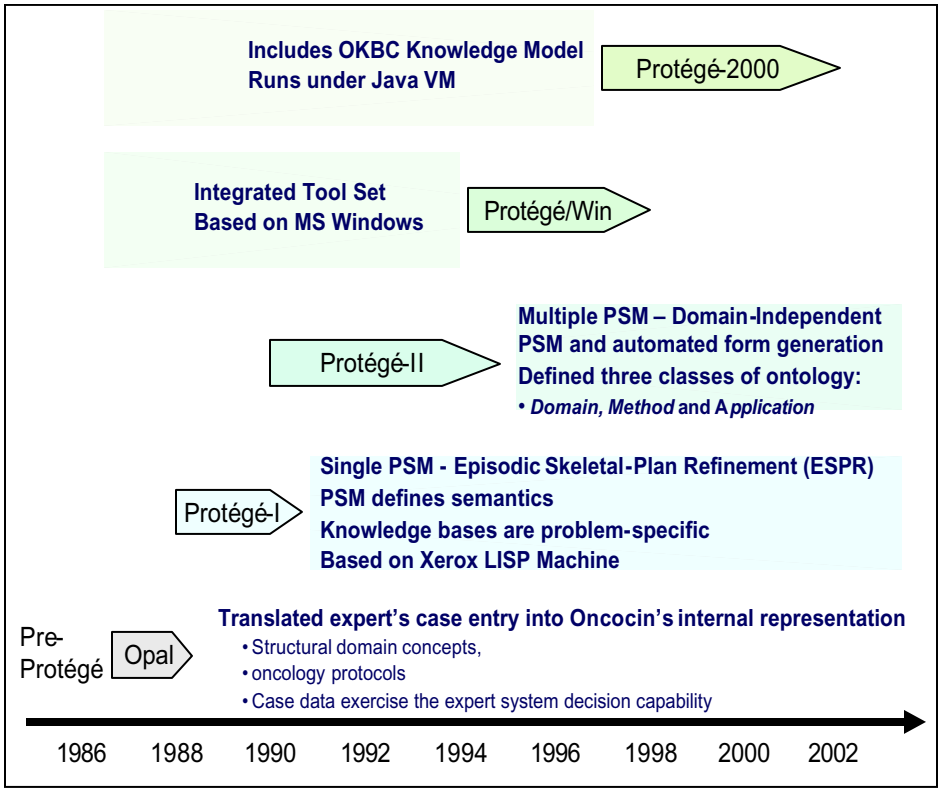

Figure 1: Evolution and enhancements of Protégé from 1986 to present.

physician or nurse (the domain experts), resulting in a suggested treatment or test. The knowledge base, a collection of if-then rules and other data, captured the clinical protocols. Opal translated the expert's input, via graphical forms, into an internal representation specifically tailored for Oncocin. This project identified three different levels of knowledge for this particular information:

1. Structural domain concepts used by the knowledge engineer to create the Opal knowledge-acquisition application;
2. The domain expert (oncologist) knowledge - oncology protocols;
3. Case data entered by the user to exercise the expert system decision capability.

Since Opal was inference engine based, it enabled reuse by knowledge engineers to create different knowledge bases - ultimately creating domain specific expert systems. The knowledge engineer was responsible for the knowledge acquisition; a tedious and time-intensive task. Unfortunately, the concept of a knowledge engineer separated the domain experts from the domain knowledge bases and this separation introduced a potentially large source of incorrect knowledge.

In 1987, Mark Musen built an application for knowledge-based systems with a goal of building knowledge-acquisition application as part of Oncocin. Based on these three different levels of knowledge, he believed that knowledge acquisition occurs in phases with knowledge obtained in one phase defining the structural knowledge for subsequent phases. Musen's goal was to reduce the work engineers did to construct knowledge bases during knowledge-acquisition. He noticed that knowledge obtained during a specific phase influences the knowledge related application required for later stages.

From this early concept, Protégé evolved through four phases, resulting in a rich development environment available for both research and knowledge-management. (see http://smi-web.stanford.edu/pubs/SMI_Abstracts/SMI-2002-0943.html for more information on the evolution of Protégé.

Protégé-I - Knowledge Bases

This early phase, where Protégé-I simplified the knowledge acquisition process for building medical expert systems, learned from the earlier OPAL-based system. The intent of the tool was to simplify the knowledge acquisition process for the knowledge engineer, already overloaded with complex tasks to perform - a key issue with early expert system development. One goal of Protégé was to provide an application that created Knowledge Acquisition tools (KA) from a formally defined collection of concepts. This enabled the domain expert to create the knowledge base, eliminating the time consuming process of the knowledge engineer learning the domain. Early assumptions for Protégé 1 included: