VIVO: Case Study of an Ontology-Based Web Site

Brian Caruso, Brian J. Lowe, Jon Corson-Rikert, Medha Devare

Albert R. Mann Library
Cornell University
Ithaca, NY 14853
{bdc34,bjl23,jc55,mhd6}@cornell.edu

Overview

VIVO (http://vivo.library.cornell.edu/), a Virtual Life Sciences Library at Cornell University, is the result of an initiative to make information about the wide variety of life sciences activities taking place across Cornell's colleges and campuses publicly discoverable through an intuitive and extensible Web interface. While not intended to replace any individual unit's Web presence, VIVO provides a new and comprehensive view of the life sciences at Cornell with the goals of showcasing current research, fostering interdisciplinary collaboration, and demystifying the university's structure for outsiders.

VIVO's persistence layer is a relational database set up to store ontological structures and instance relationships. The software is built with free and open-source tools and is actively being extended to new Web portals and to accept Semantic Web data encoded according to the emerging W3C standards.

Background

Cornell University identified genomics as a major research focus area in 1998 with the establishment of a Genomics Initiative, which was broadened in 2001 to encompass a range of activities referred to as the New Life Sciences. Cornell University Library responded in 2002 by creating a Life Sciences Working Group to plan for new library services and to design a website consolidating information about existing services for the life sciences. This working group soon identified a broader need for a sense of community among the diverse disciplines included in the life sciences, especially at an institution as complex and academically diverse as Cornell. What developed in early 2004 is VIVO, an ontology-driven website organized to display the relationships among the people, departments, research and service labs, courses, publications, and online resources active in the life sciences at Cornell.

System Description

The VIVO system combines a Web-based ontology editor with a simple content management system, allowing the underlying ontology and the publicly visible content instances populating the ontology to be edited using the same tool. The data reside in a custom relational database persistence layer with common datatype properties (display name, free text description, dates) currently stored as columns on the instance table for efficient access. Object properties are represented in a manner akin to a typical triple store. The VIVO system is currently implemented with Java 1.5, Apache Tomcat 5.5, MySQL 4.1, and Lucene 1.9.

VIVO's ontology structure was loosely based on the Advanced Knowledge Technologies (AKT) Support and Portal ontologies. These provided classes useful for describing the research activities of a university, such as Educational-Employee, Government-Organization, and Publication-Reference.

Rather than attempting a comprehensive domain ontology for the life sciences, the system primarily models affiliations and other common relationships. For example, professors are related to their departments, to papers they have published, and to projects in which they participate. These relationships can be readily identified by the librarians and student assistants who edit the site, and effectively mirror human affiliations and interests across multiple subject domains, allowing users to browse across the university from any starting point. Shared research interests, facilities, and equipment can also be discovered via full-text search capability.

The initial system launch required a great deal of human effort to create specific subclasses and properties, and to populate instances for all faculty members involved in the life sciences and their affiliations. Text-rich paragraphs were gathered from public websites to aid the site's search engine in returning relevant instances. Recent development, however, has focused more on the automated integration of data from external sources into this existing hand-edited knowledgebase.

Data Integration

Several import tools have been written—largely on an ad-hoc basis—to accommodate multiple sources of new and updated information, including content licensed from commercial publishers, annual faculty reporting data, and the university's grants and contracts data warehouse. We are in the process of extending our tools to accommodate more diverse input sources via a generic intermediate XML input format.

Work is also underway to enhance the system's support for maintaining multiple ontologies in separate namespaces. As new instances of the software are deployed to drive different Web applications, it will be important to reuse upper ontologies not only for semantic interoperability but to minimize the human effort required to set up an application. An import utility currently leverages Stanford's Protégé-OWL API to convert RDFS and OWL ontologies to VIVO's native relational database format (Knublauch et al. 2004). This custom schema is also evolving to support more of the semantics of OWL, which will allow the library to take advantage of the ontology work being done in the relevant research disciplines.

Filtering and Public Display

By ingesting data from multiple sources into a simple and extensible data structure, the library is able to offer unique sources of integrated content filtered and aggregated to best suit the needs of both administrative and Web applications. These filtered views normally appear as portals within one application, but we have recently added support for independent CSS stylesheets and separate domain names to allow multiple independent websites to mine the same underlying content. By delivering content as XML via SOAP and REST-style Web services, VIVO offers even more flexible re-use of content in convenient aggregations with display entirely under the control of the consuming website.

Inferencing

To be sustainable, VIVO will need to leverage a minimum amount of ongoing human intervention to best advantage. While automated and semi-automated data sources will provide new content on a continuing basis, we need to maximize our ability to pull desired content from the database with a minimum amount of manual specification. Certain relationships such as co-authorship and shared grants indicate active collaborators, and faculty affiliations with multiple disciplinary graduate fields can be aggregated to indicate the breadth of scholarship within any department. Leveraging one or more basic relationships to infer more complex and dynamic affiliations uses the ontological structure to best advantage.

Acknowledgements

VIVO's development has been informed by the work of, among others, the Cornell University Library Life Sciences Steering Committee (Janet McCue, Jean Poland, Carolyn Reid) and the Cornell University Library Life Sciences Working Group: Medha Devare (chair); Ithaca: Kathy Chiang, Phil Davis, Zsuzsa Koltay, Marty Schlabach, Leah Solla, Susanne Whitaker, and Jill Powell; Weill Cornell Medical Center: helen ann brown and Kris Alpi. Student curators have included Siwing (Ivy) Tsoi, Ajay Gupta, and Emily Clausen.

References

Knublauch, H.; Fergerson, R. W.; Noy, N. F.; and Musen, M. A. 2004. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In 3rd International Semantic Web Conference (ISWC 2004), Hiroshima, Japan, 2004.