Experiences Collecting and Representing Information about Antibodies for the Semantic Web.

Alan Ruttenberg

A goal of the HCLSIG is to gain experience with gathering and representing biological information to be made available on the semantic web. Here we present work and results of an effort to gather and represent information about antibodies used in biology research. For the most part, this information is not available in curated databases, so doing this involves more than just choices about how to map an existing representation to RDF/OWL. Instead we gather information from a variety of sources, including the Alzforum Antibody compendium and by scraping vendor sites.

This project illustrates many typical steps in the process of gathering and publishing content for the semantics web.

- Identifying and acquiring the content
- Parsing semistructured information such as the contents of relational databases and web pages
- Identification of biological entities such as proteins and species
- Use of existing ontologies, where possible, and development of new terms where not
- Choosing a model that clearly represents what we know
- Using OWL to represent the information
- Using an OWL reasoner (Pellet) to verify the consistency of and query the knowledge base

A primary obstacle in this project was that most antibodies are often identified by the name of the protein which they target. Proteins have many, sometimes ambiguous, synonyms. We make use of several sources of protein synonyms - Entrez Gene, OMIM, and the Enzyme database - and a number of heuristics to translate names into known identifiers.

Another issue is the relation of the antibody to the protein that it is annotated to target. Antibodies are constructed in a number of ways, and the method and details of the construction determines how specific the antibody is. Antibodies may also target other proteins, proteins in other species, and may or may not target post-translationally modified forms of a protein. Sometimes we know these details, sometimes not. We discuss the implications of this for the choice of properties and thus the intended meaning of statements in the model.

As the nature of information gathered from these site is nonuniform and of different levels of details, representing it in OWL is quite helpful given it's ability to represent partial information and different levels of abstraction.

The OWL representation of the data, as well as the full source code implementing this project will be made openly available.