GlycO and ProPreO: A case study of two deep domain life sciences ontologies

Satya S. Sahoo[1,2], Christopher Thomas[1], Amit P. Sheth[1], William S. York[2]

[1]Large Scale Distribution Information Systems (LSDIS) Lab, Department of Computer Science, University of Georgia, Athens, GA 30602, USA

[2]Complex Carbohydrate Research Center, University of Georgia, 315 Riverbend Road, Athens, GA 30602-4712, USA

Abstract:

The life sciences domain is evolving from a purely hypothesis oriented (hypothesis formulation/ hypothesis validation) discipline to a data driven field that is increasingly dependent on computational sciences for rapid and relevant progress. The vast amount of data generated by using high-throughput experiments begets a commensurate framework to enable storage, retrieval, tracking (provenance) and interpretation, of this data, in a 'meaningful' manner. Semantic technology is an apt candidate to realize this data management and usage framework.

We review development and use of two deep domain (*glycoproteomics*) ontologies, GlycO and ProPreO, as part of a real world experience in the application of semantic concepts, technology standards, tools and applications.

Keywords: *life science ontology, glycoproteomics, semantic-mediated experimental data annotation, ontology-mediated biological experimental workflow*

1. Introduction:

An ontology is a formal specification of set of concepts that enables an 'accepted' mechanism for interpretation of concepts and relationships between the concepts, in a given domain. Biological sciences, marked by a large spectrum of data and information representation formats, is an ideal field for the *application* of ontology based semantic tools and *realization* of integration, transformation and exchange goals of semantic technology. Ontologies are increasingly being accepted by the biological research community as a requisite tool, as exemplified by OBO♣. In this paper we describe the design and implementation of two domain ontologies developed for usage in proteomics and glycomics that focus on either domain knowledge or the process of scientific experiments that lead to large amount of raw data. The process of creating these ontologies exemplified a successful synergistic effort between glycoproteomics domain experts and semantic technology researchers.

♣ http://obo.sourceforge.net

2. Ontology Development:

In this section we detail the technological aspects of the development of GlycO and ProPreO. GlycO is a deep domain ontology to capture knowledge of the structure and function of Glycan structures and functions. ProPreO is a process ontology to capture comprehensive knowledge of the proteomics experimental lifecycle. It models a formal framework to encompass all processes, tools and data concepts and the attendant relationships to enable semantic annotation of experimental proteomics data, information and knowledge. We used the Protégé ontology editor for the schema design of these ontologies; Protégé is a valuable tool for the design of small ontologies. Envisioning an ontology with a very large number of instances, we had to go a different route for the population. Semagix Freedom [13] was used to extract potential instances from databases and the World Wide Web. Additional software developed at the LSDIS lab was used to transform the extracted textual information into highly expressive OWL-descriptions.

It is difficult to compare ontologies. Different ontologies focus on different domains, even different views of the same domain. Ontologies are also developed in light of different applications. For example, the CYC [1] ontology is developed with extreme logical rigor, whereas the TAP [2] ontology, SWETO [12] or the Gene Ontology GO [3] have a comparably simple logical model. Since both GlycO and ProPreO make extensive use of the expressiveness of OWL-DL, we can classify these two logically between CYC and the others with "lighter" models. An interesting reference point is the strategy used for population and the costs that arise. For CYC, every concept needs to be manually generated, which makes the development very expensive. The same holds, so far, for GO. Since TAP is populated by crawling and scraping websites, and SWETO is populated by commercial knowledge extraction and disambiguation technology that is part of Semagix Freedom, the costs beyond schema generation are relatively low. However, the TAP and SWETO schemas are not very expressive. In ProPreO and GlycO we went a different way. A very expressive schema was generated, but the crucial part of populating the schemas involved three different strategies and associated tools: Some of the instances were first inserted manually. These function as building blocks. This was followed by automated extraction of high quality and trusted knowledge sources. Finally we applied rule-based extraction techniques to generate highly expressive instance data automatically. This paper offers some insights into rationale and experiences in ontology schema development and population tasks.

2.1 Ontology structure:

An ontology may be classified along multiple dimensions, depending on the perspective of the author. The following sections discuss the structural aspects of the two ontologies, focusing on their level of granularity. The high degree of specialization (fine granularity) is a key quality of these ontologies that make them useful in the target domain.

2.1.1: GlycO:

Since no formalized domain description for the glycomics domain had previously existed, the ontology had to be built from scratch. Glycan research has an essential role in the understanding of proteins and cell metabolism. For this reason, a glycomics knowledge base must be aware of gene-, protein- and cell properties. Our goal was to

design an ontology schema that is expressive enough to model slight differences in glycans and in the components they are made of. In this context, GlycO is meant to be more than a controlled vocabulary; its intention is to be used for scientific analysis and discovery.

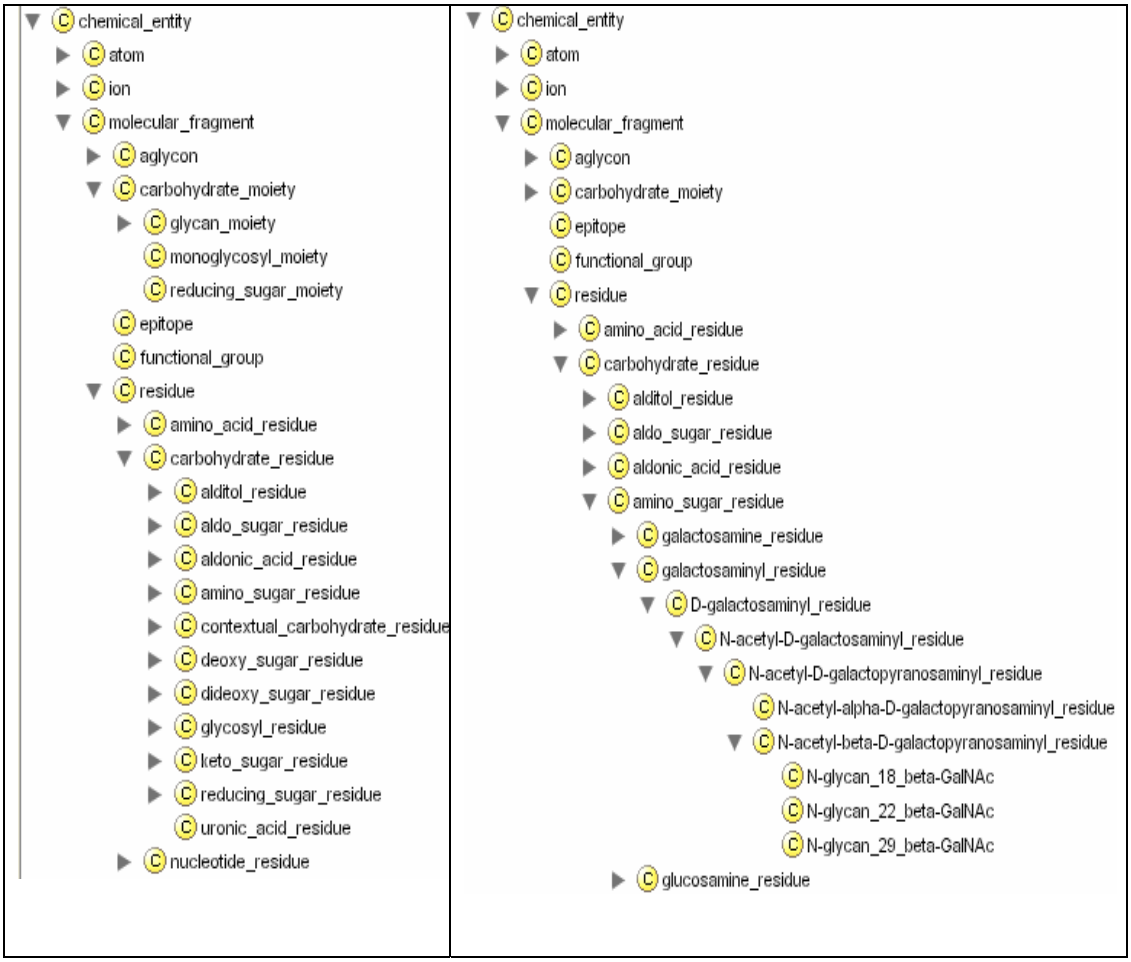

Figure 1: Selection of the first 3 levels in the hierarchy (left), the detailed classification for carbohydrate residues (right)

Initially, we analyzed the glycomics domain broadly, collected terms, and examined the way these terms are used by scientists. It turns out that the informal usage of the is_a relationship, as in "a glycan is a carbohydrate", constitutes a hierarchy of concepts with multiple inheritances. We wanted to keep the "colloquial" use of the glycomics terminology consistent with the ontology, while also adding more accurate descriptions. In addition, the is_a relationship between classes assures a very intuitive way of doing subsumption-based reasoning. There are many ways of classifying monosaccharide residues, which are the building blocks of glycans. For example, it is possible (and equally valid) to classify them according to the number of carbon atoms in the monosaccharide or as a structural variant. That is, a β-D-Glc*p* residue can be identified as both a hexosyl residue (with 6 carbons) and an aldosyl residue (embodying the aldo- structural variant). Other classifications are possible and the commonly used terminology suggests that a single monosaccharide residue can embody more than one

structural variation, (e.g., "keto" and "deoxy"), along with a ring form (e.g. pyranosyl), an overall configuration (e.g. *gluco*), an anomeric configuration (e.g. β) or an absolute configuration (*e.g.*, D). We account for all of these properties by allowing a particular monosaccharide residue inherit from several super classes.

We initially split the ontology artificially into three abstract classes: "Chemical Entity", "Chemical Property" and "Reaction". This is analogical to upper-level concepts "Entity", "Property" and "Event". At first it seems counterintuitive that a class can inherit from a property, but it should rather be read "abstract type of sugar, characterized by a named property". Hence all classes under "Chemical Property" constitute complex property descriptions that cannot be sufficiently described as proper properties within the first order paradigm of OWL.

With 11 levels, the hierarchy of GlycO is much deeper than that of many other domain ontologies. Every level in the class hierarchy has at least one more restriction than the previous. However, we relaxed Schulze-Kremer's [17] requirement, that a subclass should be distinguished from its super class by exactly one discriminating criterion, because it is not always practical. When designing a hierarchy of concepts that reflect a natural occurrence of objects, we are restricted to what actually exists. Depending on the properties modeled, in some cases no instance exists or is known that would fall in a subclass that adds exactly one restriction. This is important for the classification of new instances.

The classification scheme in GlycO is designed to extend this powerful concept to all of the monosaccharide residues within the glycan. For N-glycans, this is accomplished by defining a canonical tree that subsumes all possible N-glycans. That is, practically any known N-glycan can be completely specified by choosing a subset of the nodes of this canonical tree forming a connected (directed) graph that includes the root residue. Such a graph (known as GlycoTree) has been previously described [4], and we have formalized that structure as a collection of interconnected, canonical residue instances in GlycO.

2.1.2: ProPreO:

Glycoproteomics is the study of glycoproteins, produced as a result of post-translational modifications of proteins. Glycoproteins are being increasingly recognized as the key to the differentiation and final role of proteins in an organism. Glycoproteomics is a focused area of the proteomics research field and involves similar experimental techniques as used in proteomics. We developed the ProPreO ontology as an attempt to create a formal language underpinned knowledge base for capturing requisite information of proteomics lifecycle process and attendant data.

ProPreO structure dovetails its usage in the context of the high-throughput proteomics experimental lifecycle to create a foundation of a semantic-mediated data integration and management framework. The modeling of proteomics experiment data by the Pedro UML schema [5] in four stages namely *Sample Generation*, *Sample Processing, Mass Spectrometry and MS Results Analysis* provided a starting point in the development of ProPreO. But, mapping these stages as top-level concepts in ProPreO proved to be unsuitable and inadequate for the goals of ProPreO usage. We iteratively evolved the current top level concepts of ProPreO through multiple use cases to provide a flexible and adequate framework for reasonable current as well as future modifications.

ProPreO ontology top level concepts are:

data:

A proteomics experiment generates data and metadata at each stage of the experimental lifecycle. The various forms of proteomics data modeled includes data generated during a high performance liquid chromatography (HPLC) experiment, during mass spectrometry (including both raw and processed data) or gel electrophoresis data. The metadata required to provide the relevant experimental context includes parameters used in generating the data, accession numbers used in context of various databases.

The various concepts use a rich set of relationships to mirror existing implicit and explicit inter-dependencies in the domain. For example, an 'HPLC_gradient_time_point' is related to its 'component_portion' associating a component portion to a particular time point.

data_processing_tool:

There are many standard and intra-lab developed software applications used to process data at various stages of the experiment. Knowledge regarding these applications is required to form the correct interpretational and subsequent processing context of data. The various (proprietary) data processing applications for HPLC and mass spectrometry are modeled as subclasses of 'data_processing_tool'.

hardware:

The 'hardware' concept has two subclasses namely 'instrument' and 'instrument_component'. This enables ProPreO to capture information not only of the specific instruments used in an experiment but also critical metadata regarding the states of the various components of an instrument pertaining to an instance of the experiment that generated a given instance of data. For example, `hardware→instrument_component→HPLC_detector_type→HPLC_diode_array_detector` has two 'datatype properties' namely 'has_wavelength_detection_max' and 'has_detection_wavelength_min' to capture the settings of the wavelength detector of the HPLC for a particular experiment run.

This illustrates an example of ontology-mediated experimental data annotation, also providing the functionalities of a data provenance framework.

molecule:

This defines the broad classes of molecules relevant in a proteomics experiment; for example, glycans, proteins and peptides. These molecules have an extensive set of relationships to model required information; namely the accession number in the context of various public databases, experimental chemical mass or theoretical isoelectric point.

organism:

This class describes the taxonomic classification of a biological species. Using subclasses of this concept, the biological context of a given sample is captured, a critical provenance source that helps in providing the accurate context.

parameter_list:

A given instance of experiment involves a multitude of parameters that pertain to the instruments used in the experiment, the environmental settings in which the experiment was conducted or the user-defined values used in the various software applications for processing data at various stages. Hence, the values of these parameters form a vital source of knowledge used by scientist to compare, interpret and use experimental data.

ProPreO models parameters relating to denovo sequencing, database searching, HPLC or mass spectrometer runs and parameters of various components of a mass spectrometer for a specific sample analysis. These lists of parameters are related to the parameters defined in the top level concept 'data'.

task:

Various human-mediated or automated set of tasks are involved in a proteomics experiment. The concept 'task' has subclasses like 'filter components', 'identify components' or 'separate components' that are used in semantic modeling of a proteomics experiment instance.

This extensive set of critical concepts to capture the near-optimal amount of information defines the next stage of data provenance i.e. '**semantic experimental data**'. Using the inherent capabilities of an ontology modeling framework namely disambiguation and reasoning, it is possible not only to interpret data in a standard manner (goal of data provenance) but also enable applications to generate 'knowledge' from the vast quantities of data generated by high-throughput experiments.

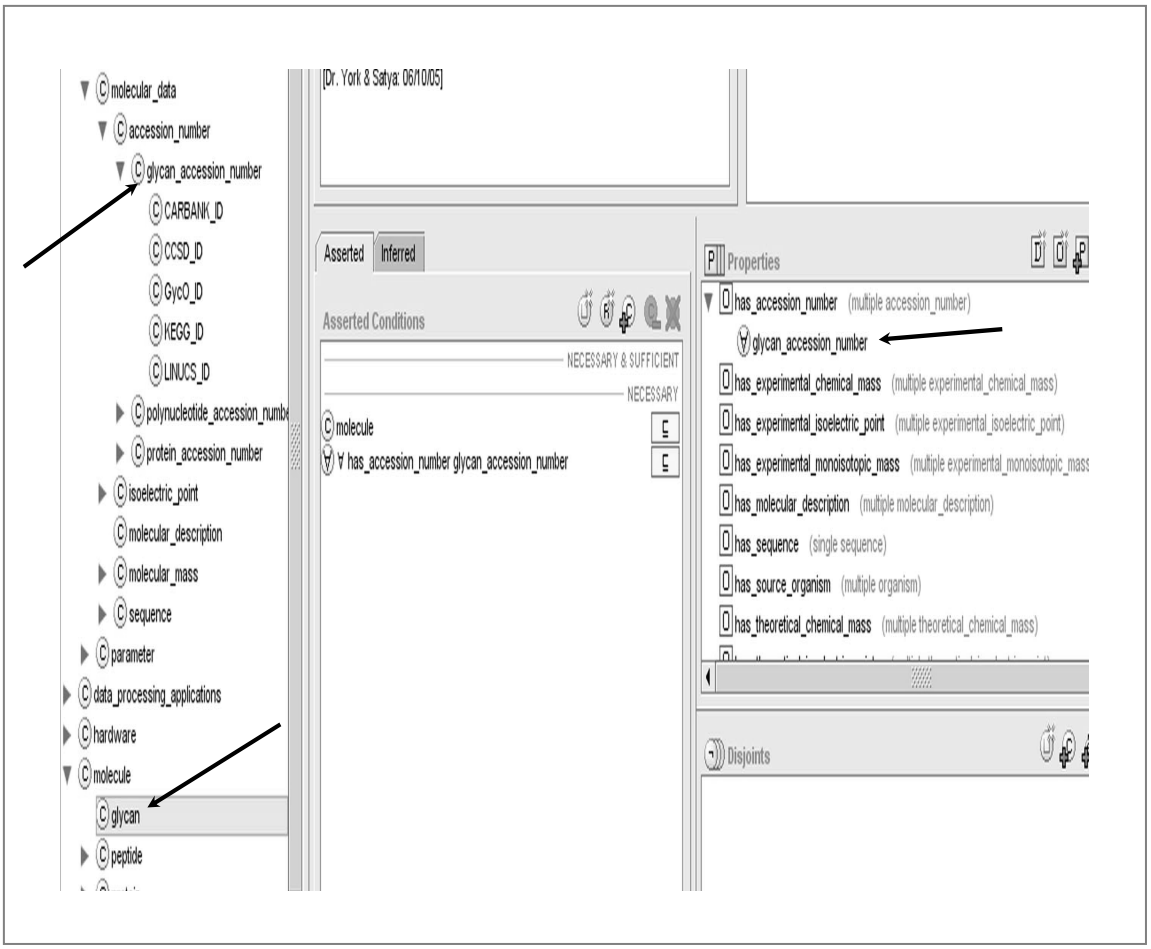

Figure 2: A snapshot of ProPreO: The molecule 'glycan' has multiple object properties like 'accession_number'. The figure highlights the usage of the constraint that defines that a given glycan molecule 'has_accession_number' values of the type 'glycan_accession_number'. Further, we define the list of existing glycan accession numbers namely CARBANK_ID, LINUCS_ID, KEGG_ID.

2.2 Ontology population:

The population of an ontology with instances, representing the concepts defined at the schema level forms a critical aspect of the knowledge captured by an ontology. In the following sections, we describe the various challenges faced when populating GlycO and ProPreO, and the approaches employed to overcome them.

2.2.1 GlycO:

Once a sufficient description of the domain was given by the developed schema, we started populating the ontology with instances. The population is done both manually and automatically. A small number of highly expressive concepts, such as monosaccharides, which function as building blocks of more complex carbohydrates, have been inserted manually by the domain expert to assure accuracy and comprehensive description at this important level. The number of monosaccharides is very limited; hence the manual population of this part of the ontology is also the most efficient way.

Other biological and biochemical structures are not as modest. Thousands of Glycans, Proteins and Genes make their virtual appearance in many different databases. In order to harvest this data, we used the Semagix Freedom toolkit that allows extraction of data from semi-structured web pages. Simply collecting this information is not enough, since database schemas are usually shallow and categorization is done by keywords rather than by a class hierarchy. Hence instances have to be classified after extraction from the source. If the class hierarchy is, amongst other restrictions, value restricted, keywords can be used to aid the classification of the instance data. In the case of GlycO, the classification is even finer and instances have to be classified according to their structure. The conversion of the glycan structure into the LINUCS [6] compatible GLYDE [7] format provides the initial step. The instance information is then analyzed according to GlycoTree [4]. The glycan is split into its residues and each residue is categorized as a contextual residue.

In order to have source data of highest quality, we chose to extract instances from different databases and compare them during the encoding phase. The databases used were KEGG [14] [18], SweetDB [15][19] and UGA's own CarbBank[16][20].

Populating an ontology automatically from several sources is both an opportunity and a challenge. In order to get the highest quality and quantity of knowledge, potentially more than one source has to be consulted for every instance put in the ontology. Each source might focus on different criteria and leave out others that we still want to insert in our knowledge base. Hence the knowledge extractor has to differentiate between new instances and those that have been inserted before and can be enriched with new information from a different database. For this, the extractor needs to have sophisticated entity disambiguation techniques. Most databases use unique proprietary accession numbers for their entries, so a disambiguation across databases by key is not possible. Different naming conventions prohibit disambiguation by name. Many different glycan structures have the same elemental composition (meaning the same number of each of its atoms and hence also the same molecular mass). Finally the IUPAC [8] notation for glycan structure is not unique, so it cannot be a discrimination criterion either.

The easiest way to disambiguate is to find a common link to a CarbBank accession ID for the particular glycan. CarbBank is still one of the most comprehensive and most referred-to collections of glycan structures and related publications. However,

since the curation of CarbBank was discontinued, not every glycan has a representation in CarbBank. For these new cases, the IUPAC structure of the glycan is sent to the SweetDB web based application[*http://www.glycosciences.de/tools/linucs/]* to convert it into the unique representation of the LINUCS format. This unique identifier allows a reliable disambiguation in the absence of other discriminating data. . This is an example of domain specific disambiguation approach where general techniques of disambiguation would most likely fail.

Another major obstacle that has to be overcome when a highly expressive schema is defined is that of *incomplete knowledge*. Some properties of a class might be set as required in the schema, because real-world entities that would belong to this class definitely have this property. However, these properties might not be explicitly stated in the knowledge source, but implicitly available in the glycan structure or otherwise deducible from known facts. Since rule base inference is not a part of the OWL framework, this deduction is best being done prior to adding the instance to the ontology with special tools, or, where appropriate, using SWRL [21] rules on top of OWL.

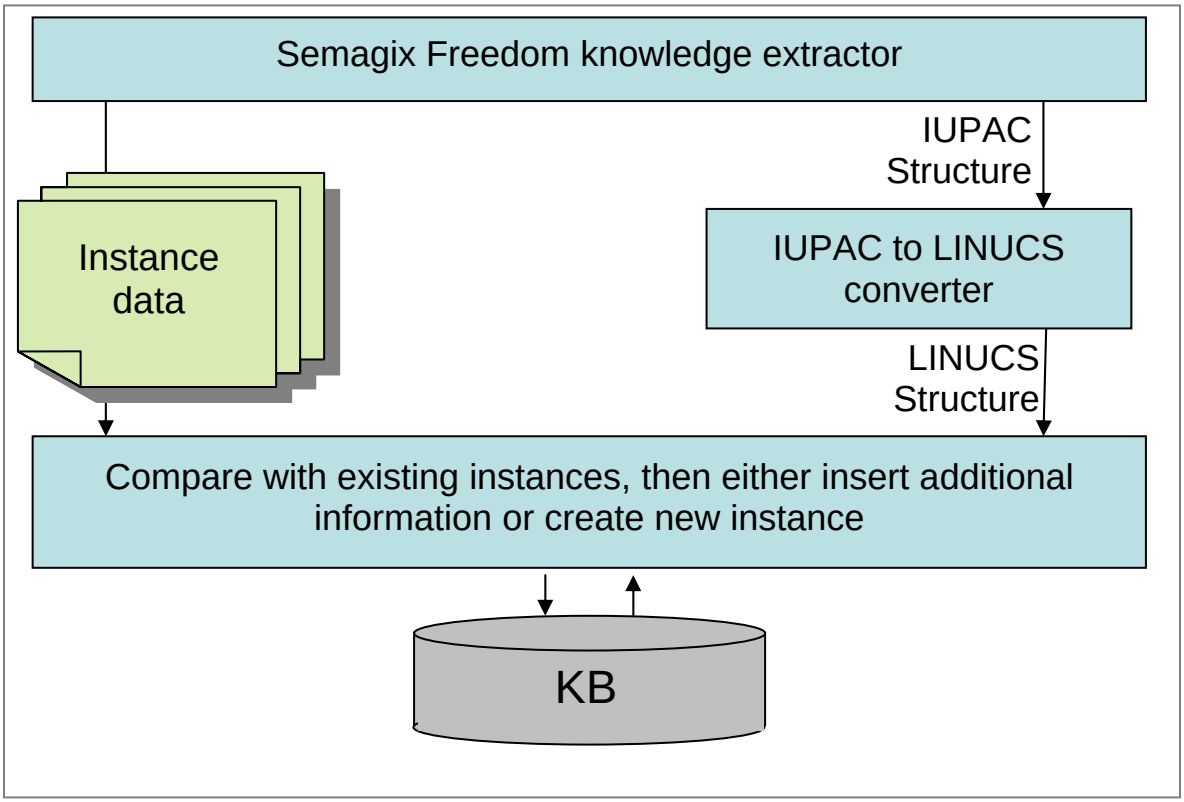

Figure 3: Ontology population workflow for GlycO

2.2.2: ProPreO:

The experimental data generated by high-throughput experimental protocols are of extremely large magnitude. For example, one mass spectrometry sample run generates 500 MB of data and in a typical research lab; hundreds of such samples are run in the course of one project. Moreover, geographical proximity and access to a common data repository, with an attendant high-speed network infrastructure, are necessary to maintain the instance base of ProPreO. But, to enable the desired usage of semantic experimental

data by reasoning or retrieval applications, it is required that the tool has access to all relevant data sets in a coherent manner.

Proteomics experimental data also involves a set of instances that is referred to recurrently by many other data sets. For example, the set of human tryptic peptide sequences are a limited set of instances of peptides that are generated from proteins by trypsin proteolytic enzyme. Hence, it is logical to create a central instance base of ProPreO with the set of human tryptic peptide sequences which may be referred to by other experimental data sets.

We have developed a novel approach to solve the challenges faced in populating ProPreO by having two levels of instance base:

Level 1: This is the typical instance base of an ontology consisting of extracted instances of the recurrent dataset, namely the human tryptic peptide sequences. An in-house database (at the Complex Carbohydrate Research Center, University of Georgia) was used to extract the relevant dataset.

Level 2: Using Universal Resource Identifiers (URIs), we populate the ontology with 'pseudo-instances' that are not datasets and hence not physically present instances; but point to repositories of relevant datasets. This enables us to present a logically unified view do the instance base in ProPreO to application.

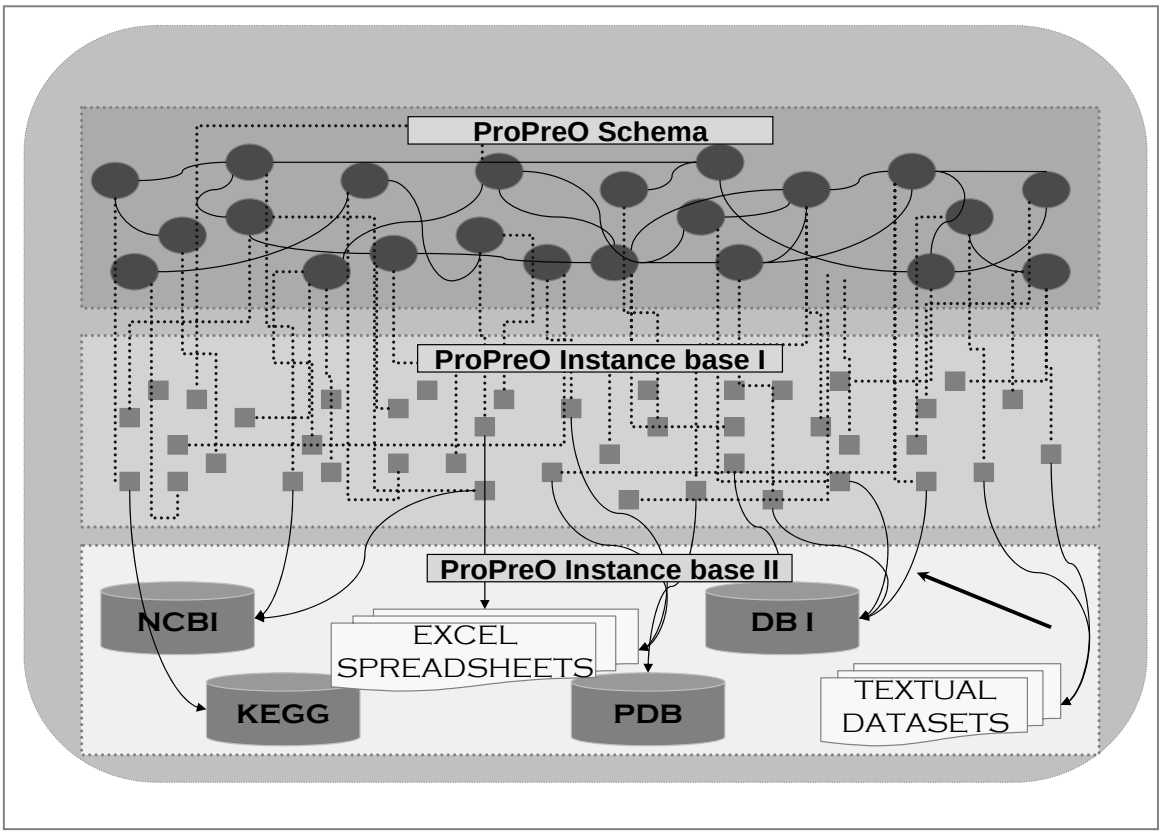

Figure 4: The figure illustrates the two levels of instance base of ProPreO. Instance base I corresponds to the Level 1 and Instance base II corresponds to the Level 2 described previously. (Note: The public databases like NCBI, KEGG and PDB are currently not linked to the ProPreO instance base)

Ontology metrics:

GlycO: 610 classes, 144 slots, 203 manually inserted instances and 24000 glycan instances extracted with Semagix Freedom out of which more than 500 have so far been structurally identified within the framework of the GlycoTree.

ProPreO: 340 classes, 200 slots and 40,000 instances (a majority of the instances are from database of human tryptic peptide sequences).

2.3 Ontology encoding:

The W3C recommended Ontology Web Language (OWL) comes in 3 different levels of expressiveness: Lite, DL and Full [9]. Our choice for the DL-flavor was based on carefully balancing the pros and cons of expressiveness and computability. OWL-Full is a syntactic and semantic extension of RDFS. Its expressiveness goes beyond that of First-Order Logic and is thus not decidable. For an Ontology that is used for reasoning tasks, consistency is mandatory. Using a language that is not decidable would not permit automatic consistency checking. The reasoner could produce wrong results even for simple queries.

OWL-Lite is based on the SHIF (D) and OWL-DL is based on the SHOIN (D) description logics. There are a number of differences between these two flavors. Most important for us were value restrictions and exact cardinality restrictions in OWL-DL. Whereas OWL-Lite allows only zero, one-to-one, and one-to-many relations, In OWL-DL we can specify an exact range of allowed relationships for a class. We can say, for example, that a student must have at least 3, but no more than 5 advisors on his committee. Just like OWL-Lite, OWL-DL is fully decidable and we can check the consistency of the ontology whenever it has changed. Hence we can guarantee correct reasoning results.

3. Ontology development rationale:

The key metrics for the success of ontology development effort is its acceptance and usage by the relevant research community. In the following sections, we describe two specific applications context of GlycO and ProPreO.

3.1: Semantic-mediated representation of glycan structures, description of glycan functions and annotation of proteomics experimental data:

The synthesis of glycans is a complex biochemical process, which is described as a set of metabolic pathways. A complex glycan is synthesized in several steps, each of which should be described in the ontology. The complex metabolic pathways and the single reactions that lead from one glycan to another are modeled to infer similar processes that might lead to the formation of similar glycans that have not yet been discovered or classified.

Glycans are represented as collections of interconnected monosaccharide residues, which are, in turn, classified according to their chemical context within the glycan structure. For example, a typical N-glycan contains a single β-D-Man*p* residue in its core. This residue is glycosidically linked to a specific site (oxygen-4) of the next residue, which is invariably a β-D-Glc*p*NAc residue. The identity of the β-D-Man*p* residue and its precise location in the core of the glycan allows it to be unambiguously

classified. In fact, glycobiologists often refer to this residue as "the core β-Man residue", with the implied assertion that this residue is in a particular molecular location and that its biosynthetic addition to the glycan was catalyzed by a specific class of enzymes, so called glycosyl transferases (i.e., a `GDP-mannose-dolichol diphosphochitobiose mannosyltransferase, EC 2.4.1.142`). This can be viewed as a colloquial classification of a glycosyl residue, as the different N-glycan structures all contain a glycosyl residue called "the core β-Man residue." The trained glycobiologist intuitively makes a large number of inferences when this colloquial name is invoked, such as correlations between N-glycan branching patterns and biosynthetic mechanisms. However, very few of the residues that make up N-glycans have a common name based on their identity and chemical context.

By modeling the GlycoTree structure, we built a mechanism by which glycans can be semantically classified (as suggested above) simply by checking their constituent (canonical) residues against residue lists, each of which corresponds to a specific type of glycan (e.g., high-mannose or complex N-glycan). Furthermore, the chemical and biological properties of each residue within the glycan, as well as the cellular machinery involved in its biosynthesis and degradation can be semantically inferred. That is, other biological objects (such as glycosyl transferases) and processes (such as metastasis) can be associated with canonical residues that they depend on or interact with. Some of these associations may be indirect (via other objects in the ontology), or inferred by analysis of quantitative information (e.g. correlation of the abundance of glycans containing a specific canonical residue and the observation of a cellular property like invasiveness) that is contained in a semantically annotated database. An example is the specification (within GlycO) that addition of "N-glycan_b-D-GlcNAc_9" is catalyzed by an instance of the GNT-V class of glycosyl transferases, and that structures elaborated when this residue is present are recognized by the lectin LPHA.

Semantic annotation of proteomics experimental data has to take into consideration the relevance of applying semantics to experimental data i.e. the semantic annotations must be relevant for:

a) Data provenance provides the details regarding the origins of a given instance of dataset. This may include, in context of a proteomics experiment, the organism from which the sample was created, temporal details like date, time or location. The semantic experiment data, enabled by ProPreO, makes data provenance information independent of syntactic constraints. As a specific instance, disambiguation of synonym words used in data provenance is possible through the ontology schema.
b) Automated data comparability - Applications may be built to exploit the semantic-mediated comparability of experimental data using ProPreO and GlycO, The rich and fine level granularity of concepts in ProPreO like parameters settings for a HPLC run, or the type of mass spectrometer analyzer used enable semantic applications to reason over experimental data and results.
c) Finding implicit relationship between data sets using relations in the ontology – leading to indirect but critical interactions perhaps leading to knowledge discovery. For example, semantic annotation of a database that includes information regarding of the binding of different lectins to various cancer cell lines and the physiological properties of these cell lines would reveal an association of GNT-V overexpression with elevated invasiveness of various types of cancer cells.

GlycO and ProPreO are a custom built to reflect a fine granularity critical in representing concepts and relations that are needed for standardization, reasoning and derive inference in a trusted and relevant manner.

3.2 Semantic-mediated implementation of Glycoproteomics workflow:

We are implementing an ontology-mediated glycoproteomics workflow using ProPreO and GlycO to annotate experimental data generated at multiple stages. We are taking advantage of the comprehensiveness and richness of GlycO and ProPreO to describe and represent datasets and processed information using ontologically defined concepts.

This is an on-going project; **figure 5** explains the context and some details of the work. We intend to use this implementation as a test bed for semantically-enabled storage, retrieval and reasoning over experimental data.

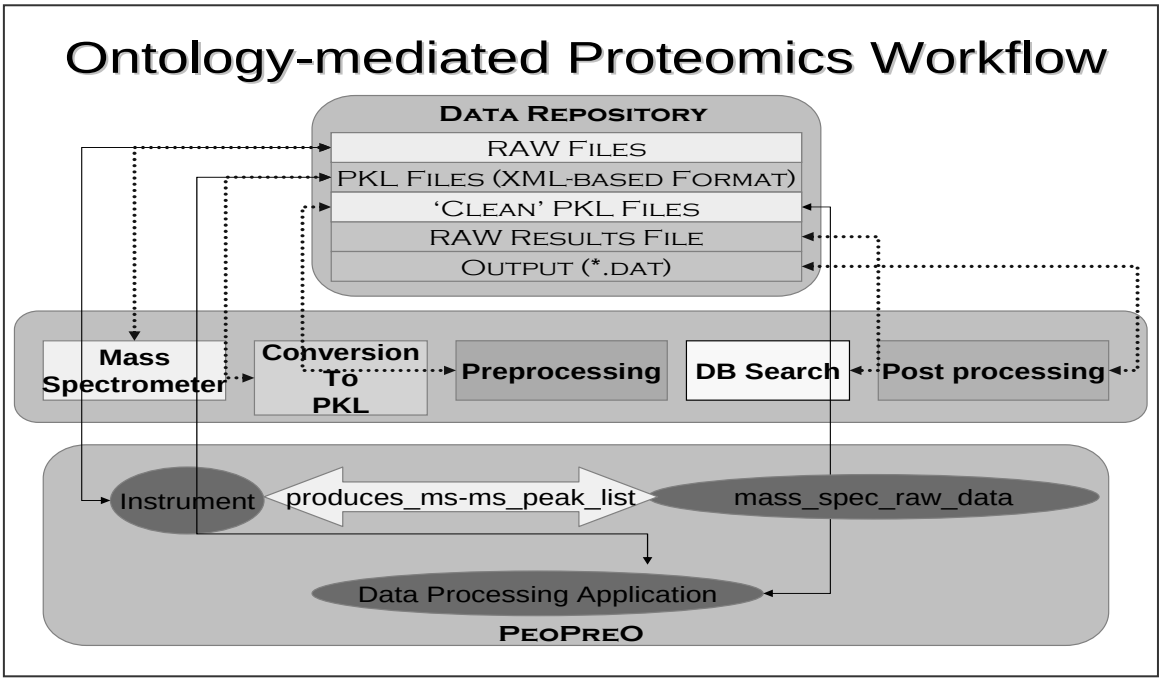

Figure 5: We describe a part of the proteomics experiment lifecycle (*analysis techniques*) and a representative sample of concepts from ProPreO to annotate the experimental data generated at various stages of the proteomics workflow.

4. Conclusion:

We have demonstrated in this paper the real-world application of semantic technologies to develop the deep domain ontologies GlycO and ProPreO. We faced multiple challenges in this endeavor including instance population of ProPreO, glycan structure modeling in GlycO.

It is particularly interesting to see how two ontologies for the same domain that overlaps in some of the modeled concepts, use different modeling paradigms. While for the development of GlycO the focus was on building a representation that expressively reflects actual glycan structure, the challenge faced in developing ProPreO was that of how to provide a unified interface to distributed heterogeneous data. Both ontologies

contain concepts such as "Glycan" or "Enzyme", but use very different representations of these concepts, due to the different natures of their application [10]. However, the real-world concepts denoted by these concept descriptions are the same and this knowledge is inherently available in the ontologies.

The logical rigor of our automatic population is, to our knowledge, unprecedented for large ontologies. The use of the Semagix Freedom© toolkit made the extraction of web resources an easy endeavor and the expressive schema description served as a rule base for the population with well described instance data. We believe that this method of designing and populating ontologies strikes a balance between the highly expressive but expensive design process for ontologies such as CYC and the affordable, but less descriptive models such as that of the TAP and SWETO ontologies.

Acknowledgements

This work is part of the Integrated Technology Resource for Biomedical Glycomics (5 P41 RR18502-02), funded by the National Institutes of Health National Center for Research Resources.

Special thanks to Dr. James A. Atwood III and Cory Henson for their contribution and participation in the development of ProPreO and GlycO respectively.

5. References:

1. Ramachandran, Deepak, P. Reagan, K. Goolsbey. First-Orderized ResearchCyc: Expressivity and Efficiency in a Common-Sense Ontology. In Papers from the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications. Pittsburgh, Pennsylvania, July 2005.
2. R.Guha and R. McCool. The tap knowledge base. http://tap.stanford.edu/
3. The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25:25-29, 2000.
4. N. Takahashi and K. Kato, *GlycoTree*, Trends in Glycoscience and Glycotechnology, 15**,** 2003**:** 235-251.
5. Taylor CF et. al. "A systematic approach to modeling, capturing, and disseminating proteomics experimental data" Nat Biotechnol. 2003 Mar; 21(3):247-54

6. Bohne-Lang et al 2001 Bohne-Lang A, Lang E, Forster T, von der Lieth CW. 2001. LINUCS: linear notation for unique description of carbohydrate sequences. Carbohydr Res. 336:1-11

7. Satya S. Sahoo, Christopher Thomas, Amit Sheth, Cory Henson, and William S. York, GLYDE – An expressive XML standard for the representation of glycan structure. Carbohydrate Research, 2005, In Press.

8. IUPAC IUPAC Commission on the Nomenclature of Organic Chemistry (CNOC) and IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Nomenclature of cyclitols. Recommendations, 1973. Biochem J. 1976 Jan 1;153(1):23-31

9. Horrocks et al, 2003 Ian Horrocks, Peter F. Patel-Schneider and Frank van Harmelen, From SHIQ and RDF to OWL: the making of a Web Ontology Language, Journal of Web Semantics 1(1): 7-26 (2003)

10. Sahoo, S. S.; Sheth, A. P.; York, W. S.; Miller, J. A. "Semantic Web Services for N-Glycosylation Process", International Symposium on Web Services for Computational Biology and Bioinformatics, VBI, Blacksburg, VA, May 26-27, 2005.

11. http://obo.sourceforge.net/

12. B. Aleman-Meza, C. Halaschek, A. Sheth, I. B. Arpinar, G. Sannapareddy, “SWETO: Large-Scale Semantic Web Test-bed,” Proc. of the16th Intl. Conf. on Software Engineering & Knowledge Engineering (SEKE2004): Intl. Workshop on Ontology in Action, Banff, Canada, June 21-24, 2004, pp. 490-493.

13. A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, Y. Warke, Managing Semantic Content for the Web, IEEE Internet Computing, July/August 2002, pp. 80-87.

14. www.genome.ad.jp/kegg/

15. http://www.glycosciences.de/sweetdb/index.php

16. http://ncbi.nlm.nih.gov subdirectory /repository/carbbank

17. Steffen Schulze-Kremer, Ontologies for molecular biology and bioinformatics, In Silico Biology 2, 0017 (2002)

18. Minoru Kanehisa and Susumu Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Nucleic Acids Research, 2000, Vol. 28, No. 1 27-30

19. Alexander Loß, Peter Bunsmann, Andreas Bohne, Annika Loß, Eberhard Schwarzer, Elke Lang, and Claus-W. Von der Lieth, *SWEET-DB: an attempt to create annotated data collections for carbohydrates*, Nucleic Acids Research, 2002 January 1; Vol 30, No. 1, 405–408.

20. Scott Doubet and Peter Albersheim, CarbBank. Glycobiology, 2, 1992, 505

21. http://www.daml.org/2003/11/swrl/