

## Chapter 14

# SEMBOWSER - SEMANTIC BIOLOGICAL WEB SERVICES REGISTRY

Satya S. Sahoo, Amit Sheth, Blake Hunter and William S. York

*Large Scale Distributed Information Systems (LSDIS) Lab, Computer Science Department and Complex Carbohydrate Research Center (CCRC), University of Georgia, USA*

**Abstract:** There are now more than a thousand Web Services [22] offering access to disparate biological resources namely data and computational tools. It is extremely difficult for biological researchers to search in a Web Services (WS) registry for a relevant WS using the standard (primarily computational) descriptions used to describe it. Semantic Biological Web Services Registry (SemBOWSER) is an ontology-based implementation of the UDDI specification, which enables, at present, glycoproteomics researchers to publish, search and discover WS using semantic, service-level, descriptive domain keywords. SemBOWSER classifies a WS along two dimensions-- the *task* they implement and the *domain* they are associated with. Each published WS is associated with the relevant ProPreO (comprehensive process ontology for glycoproteomics experimental lifecycle) ontology-based keywords (implemented as part of the registry). A researcher, in turn, can search for relevant WS using only the descriptive keywords, part of their everyday working lexicon. This intuitive search is underpinned by the ProPreO ontology, thereby making use of the inherent advantages of a semantic search, as compared to a purely syntactic search, namely disambiguation and use of named relationships between concepts. SemBOWSER is part of the glycoproteomics web portal 'Stargate'.

**Key words:** Semantic Web services, Web services registry, ProPreO ontology, SemBOWSER registry, WSDL-S, biomedical glycomics, service-level semantic annotation.

## 1. INTRODUCTION

*In silico* methods, involving the use of computational tools for conducting research, are now integral to many life sciences experimental protocols. Complementing *in-vivo* or *in-vitro* methods, *in-silico* methods have allowed scientists to leverage the rapidly increasing potential of Web accessible data repositories and software applications, which use these datasets, to gain valuable information that can be used to formulate new hypothesis or validate existing hypothesis.

*In silico* experimental methods are built around the notion of computational services that perform a well defined experimental task. These services may be used individually or chained together into multi-phase, complex processes to accomplish a more comprehensive objective. Atomic services, which are used individually, and composite services, which are constituted of multiple services, are relative concepts, as they are generally distinguished by the interface that the user interacts with to fulfill a task. However, even services that are accessible via a single interface, and thus considered to be an individual service, may be composed of multiple services. Thus, an important aspect of an atomic resource is its capacity to be seamlessly integrated into a multi-step process.

*In silico* life science research requires the collaboration of scientists with diverse technical backgrounds. For example, bioinformaticians develop computational tools and biologists use them to achieve a domain objective. These roles are not mutually exclusive and the ability of bioinformatics experts to grasp life sciences domain knowledge is of critical importance to enable them to develop relevant tools. e-Science is a term that comprises the role and characteristics of computational resources that are available to life sciences researchers. The variety of computational tools available include web accessible public databases, including NCBI databases [11], UniProt [36], and Pfam [5]; web based applications like BLAST search tools [2], structure and function prediction tools and visualization tools for biological pathways or structure of complex bioentities.

Ideally, a life science researcher should be able to navigate seamlessly across different applications with the relevant data. In reality, the large heterogeneity in terms of data representation formats, database storage schemas and the input and output data structures used by different applications make it extremely difficult to use all available computational resources in an optimal and integrated manner. In response to this complexity, there is an increase in the use of the Web services framework to wrap computational tools that process biological data and make them Web accessible. This adoption of the Service Oriented Architecture (SOA) in the life sciences domain reflects the prevalent practice in the business sector. A

growing list of biological Web services can be found in the <sup>my</sup>Grid project [22].

Semantic Web technology is being increasingly used to implement solutions that overcome many of the obstacles to the development and integration of Web services resources. This trend includes initiatives by the World Wide Web (W3C) consortium's Semantic Web Health Care and Life Sciences Interest Group (HCLSIG) [15]. One of the key efforts in this area has been the use of ontologies, which lie at the heart of the Semantic Web. Ontologies represent a consensus of the nomenclature used in a domain and capture domain knowledge in a form that can be consistently applied. This in turn, leads to better discovery, reuse and integration of both data and services. An ontology makes it possible to represent resources in a formal model that is 'understood' by software agents, thereby enabling the rapid automation of many processes in life sciences. This allows a reduction in the human intervention required in certain tasks of high-throughput experiment data management. In this way, informatics solutions can keep pace with the volume of data being generated.

Using ontologies to annotate services has been addressed by several initiatives, including WSDL-S [19] and its follow on SAWSDL [16] under the W3C is expected to lead to a recommendation in early 2007. This will provide a language that supports use of ontologies to improve reuse, discovery and composition of Web services.

In this chapter, we discuss the use of Semantic Web services (SWS) and focus on the importance of Web services registry and the use of semantic technology in a registry to enable researchers to search and discover relevant services easily and in a consistent manner. We focus on the Semantic Biological Web Services Registry (SemBROWSER) project to illustrate the use of semantics in a registry and briefly discuss the <sup>my</sup>Grid and BioMoby projects as other examples of projects using semantic technology in a registry.

## 1.1 Web Services in Biological Sciences

Services, available as computational tools, are increasingly being developed and implemented in conformity with the Web services framework. As discussed in other chapters in this book, Web services are platform neutral, highly interoperable and hold the promise of being seamlessly integrated into Web-based multi-step processes. Web services form a critical part of the Web based e-Science initiative due to their common characteristics, namely:

- a) Web based programmatic access: Web services are independent entities that may be invoked by other software applications, over the Web, using

well-defined interfaces. This allows Web services to be the ideal platform for developing high volume data processing or management tools with minimal human intervention

- b) A documented model based interaction: Web services describe their interface, their input and output and exchange data in XML schema documents. Thus, using the widely accepted XML platform during their complete lifecycle enables Web services to be compatible with a wide range of requirements.
- c) Availability for integration into complex Web processes: Individual Web services may be chained together into multi-step processes to form Web processes.

There are over 1000 Web services listed in the <sup>m</sup>Grid project [22]. This is an indication of the large number of Web services available in the life sciences domain ranging from genomics to biomedical glycomics [32]. As Web services are being rapidly adopted in a multitude of life sciences disciplines, there exist critical differences in terms of their functionality, input and output parameters, pre and post conditions, time to execute a particular task, reliability and other metrics that may also be loosely grouped into the Quality of Service (QoS) of Web services [6]. In this scenario of differing metrics, a life sciences researcher, not well-versed in the navigating XML schema based technical descriptions, which are used to describe Web services, has an extremely high initial barrier to adopt Web services. We believe life sciences researchers should not have to master technical aspects of Web service descriptions to allow them to use computational resources optimally.

Another component that must be present if Web services are to be incorporated as part of the standard suite of life science research tools is a middleware platform. This allows researchers to search for relevant Web services and if needed, combine individual Web services into Web processes that provide workflow process capabilities in SOA and Web-centric environments. Formally, the *in-silico* experimental phase requires the following:

- a) Establishment of infrastructure with a common meeting point where service providers can ‘publish’ their services and consumers can ‘discover’ relevant Web services.
- b) Standardization of the mode of interaction between service providers (bioinformatics professionals) and service consumers (life science researchers) – this is addressed via the SWS framework.
- c) Autonomous evolution of this ‘town market’ of SWSs into an established forum for providers and consumers. The concepts ‘process portals’ and ‘process vortex’ [34] are being developed for this purpose, allowing user supplied requirements and constraints to drive system

assisted semi-automated composition of multiple Web services into a Web process.

## 1.2 Registry of Bioinformatics Web Services

The business services domain has many established methods of soliciting required services from both known and unknown vendors. Request for proposal (RFPs), Request for quotations (RFQs) and electronic media based methods help customers and vendors to interact, negotiate and finalize a business transaction. In the life sciences domain, the increasing number of bioinformatics Web services requires a similar modality, allowing the researcher to search for Web services according to the required functionality, input and output, pre and post conditions and to combine these Web services into Web processes. A standard method for publishing, searching and discovering relevant Web services is critical in order to optimally leverage the increasingly complex computational resources available for life science research. This will provide a common meeting platform for service providers and service consumers.

Similar to a town market, a registry of Web services in life sciences provides a foundation for the following:

- a) A platform that allows Web service providers to offer their services. The services must be described in a standard manner, in terms of functionality, input and output, pre and post conditions.
- b) Standard interfaces to allow users to search and discover Web services in a standard and repeatable manner. Users may define a set of requirements and constraints to narrow down their search to candidate Web services.

To ensure that a Web services registry incorporates the above listed features, the publication of Web services is an important phase. Guidelines followed during the publication of a Web service should include:

- a) Description of interaction of the Web service, i.e. the functionality modeled as one or more operations
- b) Description of the input and output details for the specific Web service
- c) Description of the pre conditions that must be true before invoking the Web service and post conditions that will be true after the Web services ceases execution

The association of these multiple types of metadata with each Web service is necessary in order to facilitate its discovery and integration with other Web services to form Web processes. Specifically, these include semantic metadata incorporated directly within the SWS or stored in the registry (in addition to other attributes describing the SWS) as illustrated in Figure14-1.

Association of semantic metadata with Web services and registry allows software agents to use both in a complementary manner. This also allows the customization of the search interface according to user requirements. Since the search parameters and metadata associated with the Web services and registry are defined in a formal model, accuracy and relevance of the search results are higher compared to a purely syntactic search [10].

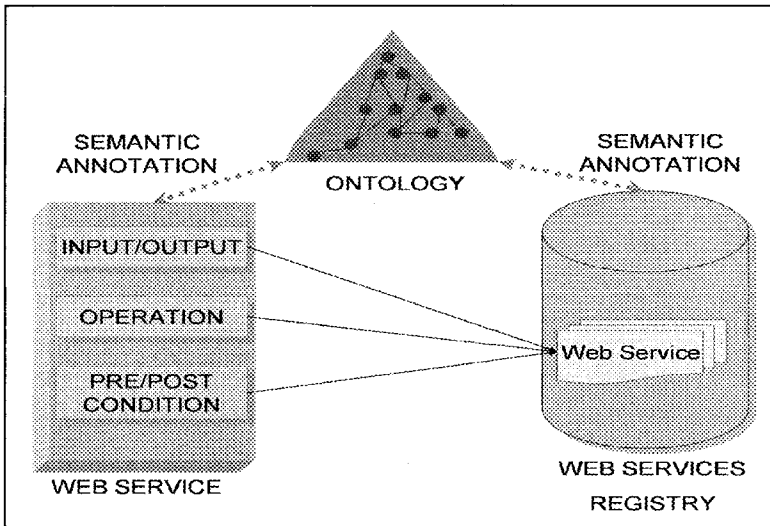


Figure 14-1. Semantic metadata for SWS may be incorporated into the SWS itself and also stored as part of the registry [30]

## 2. UDDI WEB SERVICES REGISTRY

### 2.1 Overview

A Web services registry or multiple (and communicating) registries are a critical component in the path towards widespread adoption of Web services oriented bioinformatics. Consistent with key characteristics and capabilities underlying the Web services stack, it is logical to apply a common approach or standard to the development and implementation of the Web services registry. The Universal Description and Discovery Interface (UDDI) [14], maintained by the Organization for the Advancement of Structured Information Standards (OASIS), is the standard which we describe and refer to in this chapter.

The UDDI standard allows for the publishing, search and discovery of Web services using standard and repeatable methods. These activities are made possible through the association of descriptive data and metadata with the Web services listed in the UDDI registry. The UDDI descriptions and metadata are used to:

- a) Categorize the Web service
- b) Define the modality to interact with the Web service
- c) Serve as a platform for integration of multiple, compatible Web services into a Web process

In the following section, we give further details of the UDDI model using roles of the providers, the Web services and the consumers as points of reference.

## 2.2 UDDI Data Models

The three data structures that we describe in the following sections model the information regarding the service providers, the domain functionality and the technical aspects of a SWS.

### 2.2.1 Bioinformatics Web services provider

Bioinformatics service providers are typically modeled in the UDDI standard using the *business entity* [14] data structure. The first step involves the correct interpretation of the problem being solved in context of the relevant domain. In order for Web services to have an equal footing with existing experimental research tools, they have to strictly adhere to the requirements and constraints of the problem domain. These may include the algorithm being used, the format and source of input data, and the assumptions made during the execution of the experimental method. Inclusion of these details (part of the provenance of the data), allows the user to evaluate the reliability of the results provided by the Web services.

The *business entity* data model in UDDI includes information regarding the name of the provider, contact details and the set of services offered by it. To facilitating its discovery and evaluation as a candidate Web service to perform a specific desired task these details can be semantically described in terms of concepts that are defined in a controlled vocabulary or ontology and associated with the Web service. For example, in certain e-commerce applications, the RosettaNet [12] standard can play a vital role in defining concepts regarding interactions between trading partners. Thus, in the business domain, semantic annotation using an ontology incorporating the RosettaNet nomenclature and protocols can enable search and discovery of services by software applications using descriptions or parameters specific to

service providers. These interactions can be formalized and lead to better automation if a domain specific specification such as RosettaNet is modeled as an ontology<sup>14</sup>. [29] shows the use of an ontology based on RosettaNet and additional ontologies for WS-agreement matching. Similarly, in the life sciences domain, assuming confidence in certain providers, users may search for SWSs based on the criteria specific to service providers, whereby these SWSs are annotated with respect to relevant ontologies.

### 2.2.2 Bioinformatics Web services

A Web service can be modeled in the UDDI specification using domain-specific descriptors for its functionality along with its technical and programmatic features. The *business service* data model in the UDDI standard [14] is used to describe the Web service's domain functionality.

The domain specific description of the Web service specifies the task that it executes. This includes the categorization of the Web service in accordance with a classification framework. In the business domain there are many widely accepted classification frameworks. The North American Industry Classification System (NAICS) taxonomy is a popular example. In the life sciences, various controlled vocabularies and domain ontologies exist, such as SNOMED-CT [13], Gene Ontology (GO) [4], ProPreO (for proteomics experiments) [33] and many others that can be found at OBO [9]. The use of the ProPreO ontology in the annotation of Web services is illustrated in the following section on Semantic Biological Web Services Registry (SemBOWSER).

The technical details of the Web service's programming interface, commonly referred to as the Application Program Interface (API), are modeled according to the *binding template* in the UDDI standard [14]. These technical details include the input and output data models used by the Web service, the methods or functions available as part of the Web service (the *operations* of the Web service). The Web Services Description Language (WSDL) [18] is used to describe the Web service technical interface. There has been a considerable focus on the semantic annotation of the interaction interface of Web services, particularly the .data types used in the input and output and the operations exposed by a Web service. The W3C has received four submissions for semantic annotation of Web Services, including WSDL-S [19] and OWL-S [17], that now have a wide research following. A W3C recommendation fashioned after WSDL-S called SAWSDL is anticipated in early 2007. These specifications support enriched description

<sup>14</sup> A partial RosettaNet ontology can be found at <http://lsdis.cs.uga.edu/projects/meteor-s/index.php?page=6>.

of Web services by associating metadata with respect to ontologies or conceptual models. Research initiatives like the METEOR-S [10], BioMoby [24] and <sup>my</sup>Grid [7] have used semantic technologies to add semantic annotations to Web services in various domains, including life sciences. Figure 14-2 is an excerpt of a WSDL-S file annotated using the ProPreO ontology.

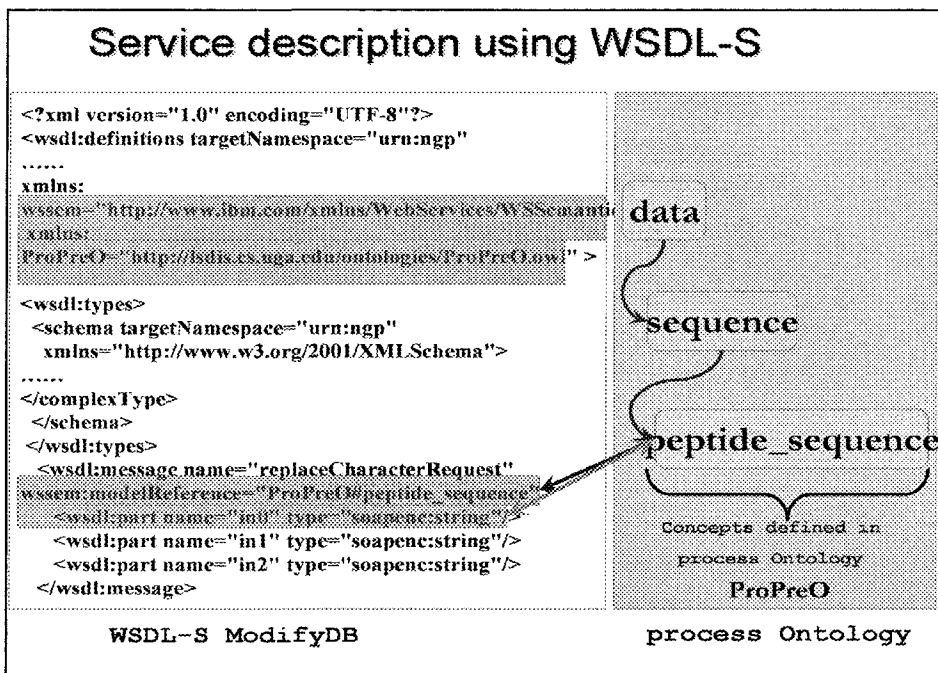


Figure 14-2. Excerpt from WSDL-S file of a Web service to identify N-glycosylated peptides from a list of identified peptides

The final UDDI data structure we introduce is the *tModel*. The *tModel* models both the *business service* and *binding template* information. *tModels* are a precise model of reference that may be used to search, discover and integrate Web services listed in a Web services registry.

### 2.2.3 Users

There are multiple ways for users to discover relevant Web services. A user may search for a Web service according to the functionality, the input and output data, the constraints related to performance or quality and service providers. However there are no data models for users in a Web services registry using the UDDI standard.

Hence, an application that seeks to implement customized search features for users unfamiliar with the XML-schema based search interfaces needs to store the requisite data in native data models of UDDI. In SemBROWSER we store such metadata about SWS in the existing data models of the UDDI framework.

### **3. SEMANTIC BIOLOGICAL WEB SERVICES REGISTRY (SEMBROWSER)**

As introduced in preceding sections, there have been several initiatives using semantic technologies to ease the adoption of Web services as an integral experimental tool in the repertoire of a life sciences researcher. The different projects have focused on use of semantics on different aspects of Web services and their registry. One approach involves the use of semantics to describe the data types used by the Web services and the subsequent search, discovery and integration of Web Services using the compatibility of input and output data types [10]. Another approach involves the association of semantics with the operations exposed by the Web services as well as the input and output data models [26]. In contrast the SemBROWSER approach considers the Web services as single functional entities that perform a given task and hence associates the semantics to this feature of the Web service. SemBROWSER also associates semantics to the operations and data types using the WSDL-S specification.

In the following sections we describe in detail the SemBROWSER project.

#### **3.1 Implementation of SemBROWSER**

As part of the Integrated Technology Resource for Biomedical Glycomics funded by the National Center for Research Resources (NCRR), we are using SWSs to allow the seamless sharing and use of computing resources and data by researchers in their routine work. The suite of glycoproteomics services uses the inherent advantages of SWSs namely, to be used in a platform-independent manner, the use of XML-based representation formats for exchange of data and ultimately the possibility to form multi-step, complex Web processes leveraging associated semantic metadata. The SWSs, developed as part of the biomedical glycomics project, include tasks such as data format transformations, filtering, categorization based on given sets of constraints, as well as search and identification of patterns in datasets. As discussed in the previous sections, the rapid increase in the number of available services in a registry makes it difficult for a

researcher, unfamiliar with the XML-schema based service descriptions, to search and discover Web services. Using the unique SemBROWSER approach to leverage Semantic Web technology, we aim to make the search and discovery of Web services more intuitive for researchers.

### 3.1.1 Semantic annotation of Web services

Software applications used in automated search and discovery of Web services cannot distinguish between Web services purely on syntactic definition of input, output, pre or post conditions and operations. For example, two Web services with similar pre and post conditions may perform significantly different functions. Moreover, two Web services performing similar functions may use different protocols or assume different experimental conditions. In the life sciences domain, these variations in protocols or experimental conditions assume significance and it is not viable to integrate or compare datasets obtained from two Web services with these differences. Hence, it is extremely important to use semantic descriptions of Web services to decide on the compatibility of two Web services for integration into a Web process or subsequent processing of their datasets.

The WSDL-S specification defines WSDL based elements (using the extensible elements of WSDL) that can be used to semantically annotate a Web service, allowing publishers to unambiguously describe its characteristics. The WSDL-S specification is agnostic to the ontology specification language, unlike OWL-S or WSMO [31] which require use of specific conceptual models. The WSDL-S specification [19] also recommends the following guiding principles for semantic annotation of Web services:

- a) Use of existing standards in Web service and emerging standards in SWS to minimize disruption of existing Web services by newer implementations.
- b) Freedom of Web services publishers to choose the specific language for annotation, including OWL, Unified Modeling Language (UML) or Web Services Modeling Language (WSML) [21].
- c) An accommodation for multiple annotations of each instance of a Web service that can be described using more than one classification term.
- d) An accommodation for semantic annotation of data types described in Web services using XML schema. This encourages reuse of interfaces described for Web services.
- e) Implementation of mappings between XML complex types and concepts defined in formal semantic models such as ontologies.

SemBROWSER lists SWSs using the WSDL-S specification and the ProPreO ontology to describe the WSDL based elements. This enables

software applications to search and discover Web services using the WSDL-S based semantic descriptions. As part of the METEOR-S project [10], tools have been developed to generate WSDL-S files of Web services using relevant ontologies [8]. Since the anticipated W3C recommendation SAWSDL is largely based on WSDL-S, we also expect this work will be very easily adapted to support this emerging standard.

### 3.1.2 ProPreO ontology

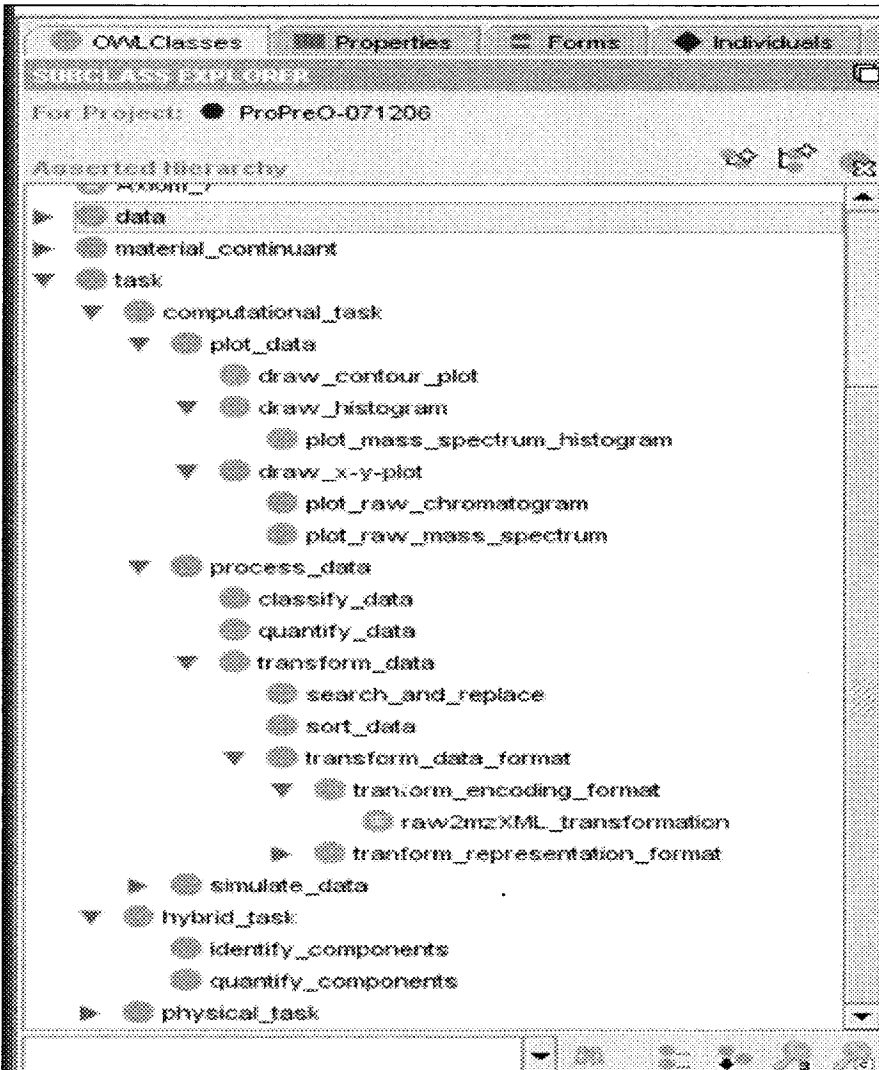


Figure 14-3. ProPreO - process ontology with concepts used in the annotation of SWSs used in the Web process described in Figure 14-4 (Protégé toolkit visualization)

ProPreO [33] is a process ontology designed to model the complete lifecycle of a glycoproteomics experiment. ProPreO models the different stages in a glycoproteomics experiment including the biological source of the sample (with its associated metadata regarding the source organism and the conditions that existed during growth of the tissue or cells from which the sample was extracted), the separation techniques (such as high-performance liquid chromatography) used to isolate molecules of interest in the sample, the analytical techniques (such as mass spectrometry) used to identify and quantitate molecules in the sample, and finally the computational resources used to process the resulting datasets.

ProPreO was modeled using the OWL-DL (Web Ontology Language) language [20]. There are 400 concepts and 32 properties with 200 restrictions in ProPreO. The ProPreO ontology is populated with real world instances of tryptic peptides, parent proteins, theoretical chemical and monoisotopic mass concepts. The size of ProPreO ontology knowledge base is 3.2 million instances and 18.6 million triples or assertions. ProPreO is listed on the Open Biomedical Ontologies (OBO) repository and freely available for download. Figure 14-3 shows a section of the concept hierarchy of ProPreO.

The top level concepts defined in ProPreO are:

- a) Data - which constitute the basic units of information. Data can include collections of information or individual units of information. Data can be experimental (measured) or theoretical (calculated)
- b) Material continuant - which is a real-world object
- c) Task - which is a process that is initiated or implemented by an agent

These top level classes loosely follow the Basic Formal Ontology (BFO) approach [35]. This allows ProPreO to be used in conjunction with other biological ontologies which also conform to the BFO approach. ProPreO was developed for application in the semantic annotation of various resources, namely experimental data and Web services, and to formally model provenance data.

### 3.1.3 Semantic annotation in SemBOWSER

In addition to the WSDL-S specification based semantic annotation of WSDL elements of Web services SemBOWSER uses concepts in the ProPreO ontology for service-level annotation of Web services that it lists. In existing approaches using semantic techniques in Web services registry, the focus is on annotation with respect to the data types used in the input/output and the operations exposed by the Web services.

To accomplish a task, the domain experts in glycobiology have to execute a number of sub-tasks, some sequentially and others in parallel.

Some of these sub-tasks form constituents of the process to accomplish many other tasks. For example, the isotopic distribution of a peptide depends on its elemental composition. The amino-acid sequence of a peptide is used to calculate the elemental composition of an ion observed in the mass spectrum, we name this sub-task as *Calculate\_ion\_elemental\_composition*. This sub-task is used in many other scenarios namely identification of phosphate-related post translational modifications in proteins. Hence, we model this sub-task as single a SWS which is integrated in multiple Web processes.

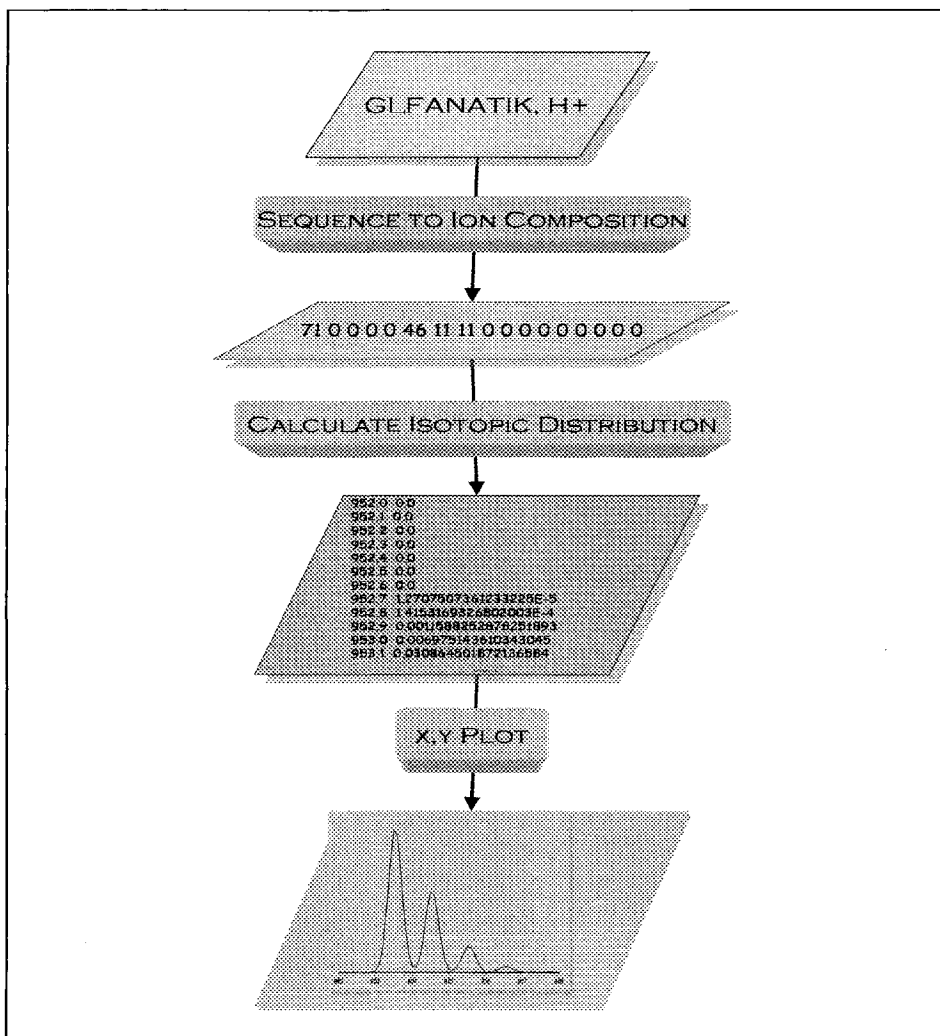


Figure 14-4. Isotomer distribution calculation as a Web process

We name such sub-tasks as one *unit of task*. The granularity of a *unit of task* was decided in consultation with domain users. The domain users, based on their experience, identified the sub-tasks of a process as *unit of task* based on its potential to be reused as component SWS in multiple Web processes. Each *unit of task*, modeled as a single Web service, usually has one publicly accessible operation.

We use a process in glycoproteomics as an example to further explain the concepts introduced above. The process to calculate the isotopic distribution for a given molecule and generation of the corresponding mass spectral pattern can be combined to form a Web process (illustrated in Figure 14-4).

Each atom of a particular chemical element (e.g., hydrogen, carbon, etc.) can exist as a different *isotope* whose mass depends on the number of neutrons in its nucleus. As atoms are combined to make a molecule, the presence of different isotopes leads to a distribution of masses for the molecule. This mass distribution is not random, but depends on the natural abundance of the different isotopes of each constituent element. This mass distribution is a distinctive feature of the molecule, and can be used to help identify it in mass spectral data. The isotopic distribution, which depends on the elemental composition of the molecule, can be calculated by a recursive algorithm based on probability theory, and the results can be plotted as a graph such that it approximates the observed mass spectral pattern.

The approach to implement a *unit of task* as a Web service required suitably relevant semantic annotations to describe the Web service as one functional unit. In addition to applying WSDL element based annotations, SemBOWSER associates semantic keywords with each Web service that describes the functionality of the service as one logical unit. Thus, users not conversant with the technical details of a WSDL file can search for relevant Web services by using familiar, domain keywords to describe the task that they require the Web service to accomplish.

We use three related Web services (illustrated in Figure 14-4) as examples to detail the SemBOWSER approach:

- a) *Calculate\_ion\_elemental\_composition* SWS- the amino-acid sequence of a peptide is used to calculate the elemental composition of an ion observed in the mass spectrum.
  - Input: a string specifying the amino-acid sequence and the adduct (e.g.,  $H^+$ ) that results in ionization
  - Output: a string specifying the elemental composition of the ion, with the number of atoms of each element separated by spaces.
  - Operation: *AA2Ele()*
  - Pre-condition: amino-acid composition known
  - Post-condition: ion composition known

- b) *Calculate\_isotopic\_distribution* SWS – simulate the isotopic distribution envelope for the given ion composition
- Input: The elemental composition (calculated by AA2Ele) along with the ionic charge, required digitization, and mass spectral resolution
  - Output: a table specifying a list of mass to charge ( $m/z$ ) values and the corresponding signal intensity for each
  - Operation: *simulate\_MS()*
  - Pre-condition: ion-composition known
  - Post-condition: spectrum simulated
- c) *x-y\_Plot* SWS – create an x,y-plot
- Input: a two-column table containing x,y-data to plot
  - Output: an image file illustrating the x,y-plot
  - Operation: *xy\_plot()*
  - Pre-condition: x,y data available
  - Post-condition: x,y-plot generated

At the time of publication of these Web services, the provider is prompted to associate semantic keywords with them, categorizing them along two axes of categorization:

- a) *Domain*: The broad life sciences sub-disciplines that are related to the Web service. The keywords for each of the relevant disciplines are associated with the given Web service, namely *glycoproteomics*, *proteomics*, *ms-ms\_data\_analysis*. These keywords are representative, and depending on the granularity used in modeling of the referred formal model, the associated semantic keywords may be extremely specific. Figure 14-5 shows the process publishers use to associate domain keywords with SWS published in SemBOWSER.
- b) *Task*: The *unit of task* executed by the Web service may be described by keyword(s), namely *Calculate\_ion\_elemental\_composition*, *Calculate\_isotopic\_distribution*, *x-y\_plot*.

The service providers visually browse through the available categories and associate one or multiple, relevant keywords with a Web service. The two SWSs listed above as examples would be annotated in the following manner:

- a) *Calculate\_ion\_elemental\_composition* SWS: The semantic domain keywords associated are *ion* and *elemental\_composition* and the semantic task keyword is *calculate\_elemental\_composition*.
- b) *Calculate\_isotopic\_distribution* SWS: The associated domain keywords are *theoretical\_mass\_to\_charge\_ratio*, *theoretical\_ion\_abundance*, and task keywords are *simulate\_ms\_data*, and *calculate\_isotopic\_distribution*.

- c) *x-y\_Plot* SWS: : The associated domain keywords are *graphical\_data\_representation* and *x-y\_plot* and task keywords are *plot\_data* and *draw\_x-y-plot*

The keywords that are available to be associated with a Web service are concepts defined in the ProPreO ontology, which is the formal model used for reference. The use of concepts from an ontology ensures that the keywords used in the annotation process are not only clearly defined but also allows the well defined named relations connecting these concepts to be used to discover related Web services. Moreover, using ProPreO enables us to apply disambiguation and mapping techniques to search for keywords used by users.

For example, the users may use the synonyms of keywords associated with the SWSs instead of the exact keywords used by the publisher to describe a SWS. Since SemBROWSER uses ontology concepts as semantic keywords associated with SWSs, it can still find the relevant SWS by mapping the search keyword (input by the user) to the original keyword (used by the publisher). The synonyms of a concept, defined in an ontology, are also part of the ontology and are leveraged by SemBROWSER.

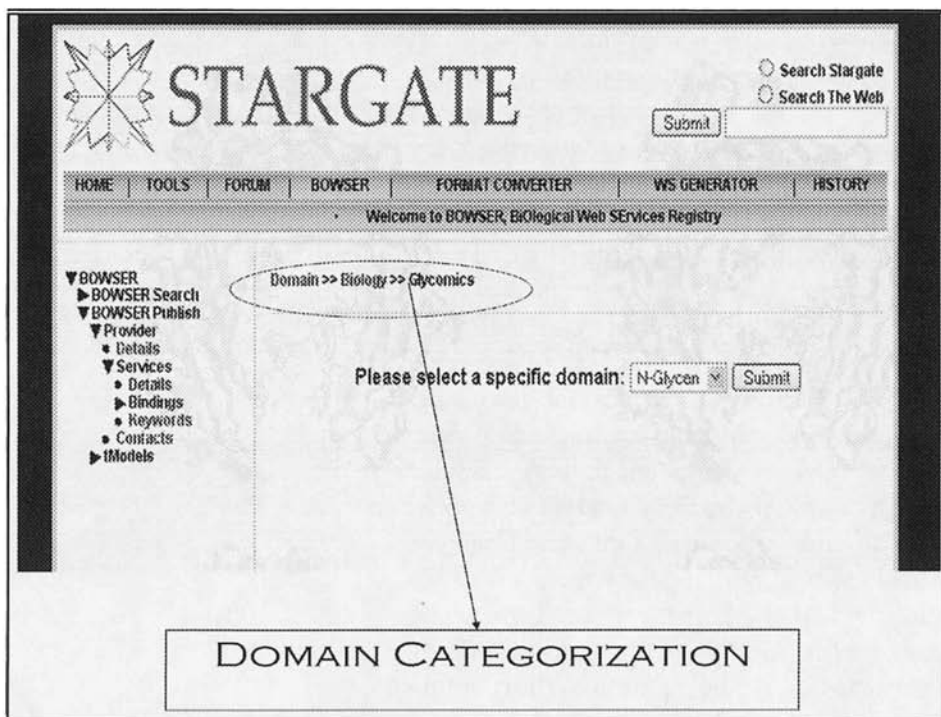


Figure 14-5. Classification of Web service according to 'domain' in SemBROWSER

In addition, if a user uses a keyword that is a *subclass* of the keyword associated by the publisher, SemBOWSER will return the correct SWS to the user. The search algorithm will compare the keyword input by the user to the keywords associated by the publisher of the SWS. If the user input keyword is not an exact match with publisher input keyword, but does match with its *subclass*, then the given SWS is the 'nearest' and most relevant result for the user. This uses the notion that a *subclass* of an entity is more refined concept than its *superclass*; hence the *superclass* is more general thereby encompassing more similarity with the concept than its *siblings* or *child* concepts.

These features allows SemBOWSER to present a more domain oriented and user friendly interface for users without compromising on the level of accuracy and relevance in retrieval of Web services.

Using the example of the SWS *calculate\_isotopic\_distribution*, if a user uses the synonym of the domain keyword *theoretical\_mass\_to\_charge\_ratio* namely *theoretical\_m-z\_ratio*, a concept listed as a synonym of *theoretical\_mass\_to\_charge\_ratio* in ProPreO, SemBOWSER will still return the *Calculate\_isotopic\_distribution* SWS as a result.

Another important advantage of using an ontology for the semantic annotation of Web services in a registry is the use of named relations between concepts in an ontology. Using the relations defined in the ontology, a registry can return logically related Web services that may be integrated together to form a Web process to achieve a broader goal. By using semantic relationships, the list of Web services returned to the user, even in the absence of exact matches, would be more relevant to the context of the user's search compared to a purely syntactic search of the services registry.

In the presence of multiple relationships between concepts, the framework for ranking these relations is important. The ranking of semantic relationships [1], [3] between concepts, according to context or relevance, modeled in an ontology is an exciting new area of research in the Semantic Web field. Ranking of relationships between entities is conceptually similar to ranking Web pages by different search engines namely Google or Yahoo. [1] discusses the implementation of an application that ranks relationships using both the Semantic and statistical metrics. Semantic metrics include the 'context' of the query, subsumption and, trust. Statistical metrics include 'rarity' of occurrence of the relation, the 'popularity' of the relation and, the length of the 'associations' between entities in the relation. While highly relevant the use of ranking relationships between entities to return most relevant SWSs, is part of future work in SemBOWSER.

At present, the ProPreO ontology is parsed to create data structures (Figure 14-6 gives a schematic representation of the data structures) to store

the set of keywords to define the *domain* and the *task*. These data structures are used as a reference to generate the graphical browsing interface for the users of the specified *domain* and *task*. The listing hierarchy created in the graphical interface uses the class hierarchy (*is-a* relationship) defined in the ontology. As stated earlier, these keywords associated by the service providers are in addition to the annotation using the WSDL-S specification. The keywords selected by the service provider are stored as part of the UDDI data structures.

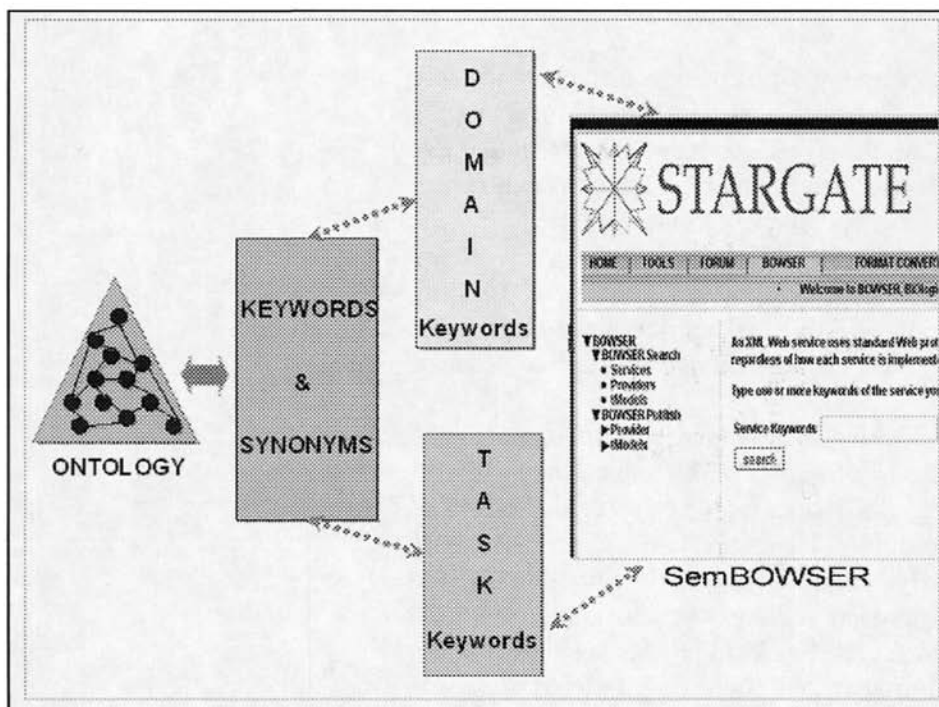


Figure 14-6. Data structures in SemBROWSER to store the concepts from ProPreO ontology

The user searching SemBROWSER for relevant Web services may use multiple search methods depending on the requirement. The user may search for Web services based on type of WSDL elements. The WSDL-S specification allows for the discovery of relevant Web services using semantic search parameters. Users may also search for Web services using domain keywords, part of the users' everyday working lexicon, that were associated with each Web service during their publication. SemBROWSER would allow users to graphically browse through the *domain* and *task* taxonomy of keywords and select relevant search terms.

## 4. DISCUSSION

### 4.1 Related Work

The <sup>my</sup>Grid project [7] seeks to create relevant middle layer services to facilitate and ultimately enable e-Science to incorporate provenance and standard data management practices. The <sup>my</sup>Grid project has developed tools to develop and execute *in silico* experiments. One such tool is the Taverna workbench, discussed in [28].

The <sup>my</sup>Grid project has also implemented a semantic discovery tool for searching Web services using the function, input and output parameters called Feta [25]. Feta assumes that the Web services it deals with have multiple operations that are functionally independent. This is in contrast to the notion of a single functionality Web service adopted in the SemBROWSER model. Feta also differentiates between the *service* and *operation* concepts as logically separate entities. The *service* entity encapsulates the information relating to published service namely, provider organization name, author of the service description and free text to describe the functionality of service.

However, the Soaplab services in <sup>my</sup>Grid share similarity to the glycoproteomics Web services listed in SemBROWSER, since they are also implemented with the notion of a single operation per service. Feta itself describes the operations using multiple attributes namely *task*, *method*, *application*, and *resource* described in detail in [25]. For a more detailed and thorough description of the Feta approach, architecture and implementation, [25] is a good reference. The approach used by Feta project is suitable to the type of SWSs listed in the <sup>my</sup>Grid project, but could be improved along the lines of SemBROWSER to provide an intuitive user interface for naïve life sciences researchers.

The BioMoby project is another significant project using semantic technology for publication, search and discovery of services [24]. Specifically, as part of the Moby-Services (Moby-S) project, the Moby central acts as a centralized registry that allows search by specification of input or output types augmented by graph crawling [24].

Though BioMoby features the most comprehensive attempt to use semantic technology to define data types used in SWSs, a user interface implementation, to allow life sciences researchers to look up relevant SWSs using only keywords they are already familiar with is also missing. The BioMoby Dashboard (an interface to help service providers for developing and deploying their services) and [27], which is an implementation using the BioMoby API, currently do not allow biologists to look up SWSs using

descriptive domain keywords underpinned by a domain ontology. We consider the incorporation of data type definitions using semantic annotation, according to WSDL-S specification, and service descriptions is a compelling solution to cater to both life sciences researchers and automated service discovery and composition using software applications.

## 4.2 Future Work

The use of Semantic Web technologies to allow life sciences researchers to effectively leverage available computational resources (mainly as SWSs) is still in its early stages of development and adoption. The projects discussed in this chapter have yet to employ one of the most potent advantages of the Semantic Web, i.e. relationships. Well defined relationships between concepts used to annotate Web services and registries (where the services are listed), should be used in the next step of SWSs registries development. In SemBROWSER, we plan to augment the retrieval process of SWSs using the underlying relationships between concepts associated with each Web service. Assuming that relationships in an ontology relate functionally close concepts, it is possible that related Web services can be retrieved to form a Web process. Thus, users searching for relevant SWSs to chain together to accomplish a complex task through the implementation of a Web process can be returned a list of semantically related Web services that have a greater probability of being successfully integrated into a Web process. This potentially means that the Web services will have semantically compatible input and output, their operations would form a logical chain of successive stages in a broad goal and their pre/post conditions will also be a series of compatible values.

It is also important to note the exact behavior for identification of relevant SWS, based on inputs, may be implemented in multiple ways. The SemBROWSER implementation, using subsumption rules, reflects one such approach. Other approaches may involve solicitation of more search parameter details, from the user, to return relevant SWS.

We are also working on the use of multiple ontologies, in addition to ProPreO, to semantically annotate the Web services listed in SemBROWSER. Consistent with one of the design principles of the WSDL-S specification, we plan to use relevant ontologies to describe the *business entity* data structure of UDDI associated with a Web service. We are also leveraging the potential of using relationships between concepts to retrieve related Web services that may be chained together to form a Web process to accomplish a broader objective.

UDDI version 3.0 introduced the notion of an ‘association’ of Web services registries. Though there are no current implementations of this

notion in life sciences domain, we believe that major SWSs registries should collaborate to complement their strengths and minimize shortcomings. Since it is commonly accepted that multiple ontologies are needed to successfully model any given domain, we believe that multiple, cooperating Web services registries hold the key to the successful adoption of SWSs by researchers in the life sciences domain.

## **5. CONCLUSIONS**

In this chapter, we have discussed the importance of a Web services registry that utilizes the Semantic Web technology to associate semantic metadata with listed Web services at both services and registry levels. With the rapid increase in adoption of the Web services framework to share computational resources in life sciences community, the Web services registry as platform to search, discover and integrate services into Web processes is critical.

The heterogeneity in data representation formats, the input, output, operations and other Web services interface descriptions hamper the search for relevant Web services and integration into complex Web processes. Semantic Web technology in the form of annotations associated with Web services and stored as part of the Web services themselves or the registries offer a solution to these obstacles. These semantic metadata are referred from a formal model, namely an ontology. The formal models define concepts clearly and comprehensively to allow software applications to disambiguate between similar entities and use the well-defined relationships defined between these concepts to retrieve related resources (SWSs).

Using SemBROWSER as a case study, we have described one approach to associate semantic metadata with Web services and a registry to enable users to easily find relevant Web services and integrate them into Web processes using familiar domain keywords. In this context we have briefly discussed the contending approaches used in the Feta registry (part of the <sup>my</sup>Grid project) and the BioMoby projects.

## **ACKNOWLEDGMENTS**

SemBROWSER was developed as part of the Integrated Technology Resource for Biomedical Glycomics (5 P41 RR18502), funded by the National Institutes of Health National Center for Research Resources.

## REFERENCES

- [1] Aleman-Meza B., Halaschek-Wiener C., Arpinar I.A., Ramakrishnan C., and Sheth,A., Ranking Complex Relationships on the Semantic Web, *IEEE Internet Computing*, 9(3), pp. 37-44, May/June 2005
- [2] Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. "Basic local alignment search tool. *J. Mol. Biol.* 215:403-410, 1990
- [3] Anyanwu K., Maduko A., and Sheth. A. SemRank: Ranking Complex Relationship Search Results on the Semantic Web. *In the Proceedings of the 14th International World Wide Web Conference (WWW2005)*, May 2005. Chiba Japan, 117-127.
- [4] Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis, A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium *Nature Genet.* 25: 25-29, 2000.
- [5] Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L.L., Studholme D.J., Yeats C., and Eddy S.R., "The Pfam Protein Families Database", *Nucleic Acids ResearchDatabase Issue* 32:D138-D141, 2004 (<http://www.sanger.ac.uk/Software/Pfam/>)
- [6] Cardoso J., Sheth A., Miller J., Arnold J., and Kochut K., Quality of Service for Workflows and Web Service Processes, *Journal of Web Semantics*, Elsevier, 1 (3), 281-308, 2004.
- [7] Goble C., Using the Semantic Web for e-Science: Inspiration, Incubation, Irritation *Lecture Notes in Computer Science* 3729:1-3
- [8] Gomadam K., Verma K., Brewer D., Sheth A.P., and Miller, J.A. Radiant: A tool for semantic annotation of Web Services, *4th International Semantic Web Conference (ISWC 2005)* Galway, Ireland
- [9] <http://bioontology.org/resources-obo.html>, Open Biomedical Ontologies - OBO
- [10] <http://lsdis.cs.uga.edu/projects/meteor-s/>, METEOR-S: Semantic Web services and processes
- [11] <http://www.ncbi.nlm.nih.gov/>, NCBI homepage
- [12] <http://www.rosettanet.org/Rosettanet/Public/PublicHomePage>, RosettaNet Standard
- [13] <http://www.snomed.org/snomedct/>, SNOMED-CT
- [14] <http://www.uddi.org/>, UDDI
- [15] <http://www.w3.org/2001/sw/hcls/>, W3C Semantic Web Health Care and Life Sciences Interest Group
- [16] <http://www.w3.org/2002/ws/sawSDL/#Charter>, Semantic Annotations for WSDL Working Group charter
- [17] <http://www.w3.org/Submission/OWL-S/>, OWL-S
- [18] <http://www.w3.org/TR/wsdl>, Web Services Description Language – WSDL
- [19] <http://www.w3.org/Submission/WSDL-S/>, WSDL-S
- [20] <http://www.w3.org/TR/owl-features/>, Web Ontology Language – OWL
- [21] <http://www.wsmo.org/TR/d16/d16.1/v0.21/>, Web Service Modeling Language (WSML)
- [22] Hull D., Stevens R., and Lord P. Describing Web Services for user-oriented retrieval *W3C Workshop on Frameworks for Semantics in Web Services*, Digital Enterprise Research Institute (DERI), Innsbruck, Austria. 2005.
- [23] Li K., Verma K., Miller J., Gomadam K., and Sheth A., Semantic Web Process Design, *in Semantic Web Processes and Their Applications*. J. Cardoso, A. Sheth, Editors. Springer, 2006 (in print)

- [24] Lord P., Bechhofer S., Wilkinson M.D., Schiltz G., Gessler D.,
- [25] Hull D., Goble C., Stein L. Applying semantic web services to Bioinformatics: Experiences gained, lessons learnt in ISWC 2004. Springer-Verlag Berlin Heidelberg, p350-364.
- [26] Lord P., Alper P., Wroe C., and Goble C., Feta: A light-weight architecture for user oriented semantic service discovery in *Proc of 2<sup>nd</sup> European Semantic Web Conference*, Crete, June 2005.
- [27] Nagarajan M., Verma K., Sheth A.P., Miller J., and Lathem J., Semantic Interoperability of Web Services - Challenges and Experiences, *ICWS 2006*
- [28] Navas-Delgado I., Rojano-Muñoz M., Ramírez S., Pérez A.J., Andrés León E., Aldana-Montes J.F., and Trelles O. Intelligent client for integrating bioinformatics services, *Bioinformatics Advance Access* published on January 1, 2006, DOI 10.1093/bioinformatics/bti740. *Bioinformatics* 22: 106-111., 2006.
- [29] Oinn T., Greenwood M., Addis M., Alpdemir M.N., Ferris J., Glover K., Goble C., Goderis A., Hull D., Marvin D., Li P., Lord P., Pocock M.R., Senger M., Stevens R., Wipat A., and Wroe C. Taverna: Lessons in creating a workflow environment for the life sciences *Concurrency Computat. Pract. Exper.* 00:1-7, 2000.
- [30] Oldham N., Verma K., Sheth A.P., Hakimpour F., Semantic WS-Agreement Partner Selection, *Proceedings of the 15th International World Wide Web (WWW) Conference*, Edinburgh, Scotland, May, 2006, 697 - 706 , 2006.
- [31] Patil A., Oundhakar S., Sheth A., and Verma K., METEOR-S Web service Annotation Framework, *Proceeding of the World Wide Web Conference*, New York, NY, May 2004, 553-562, 2004.
- [32] Roman D., Keller U., Lausen H., de Bruijn J., Lara R., Stollberg M., Polleres A., Feier C., Bussler C., and Fensel D., Web Service Modeling Ontology, *Applied Ontology*, 1(1): 77 - 106, 2005.
- [33] Sahoo S.S., Sheth A., York W.S., and Miller J.S., Semantic Web Services for N-glycosylation Process, *International Symposium on Web Services for Computational Biology and Bioinformatics*, VBI, Blacksburg, VA, USA May 26-27, 2005
- [34] Sahoo S.S., Thomas C.J., Sheth A.P., York W.S., and Tartir S., Knowledge Modeling and its Application in Life Sciences: A Tale of two Ontologies, *Proceedings of the 15th International World Wide Web (WWW) Conference*, Edinburgh, Scotland, May, 2006, pp. 317 - 326, 2006.
- [35] Sheth A., van der Alst W., and Arpinar I. B. Processes Driving the Networked Economy. *IEEE Concurrency* 7, 3, 18—31, July-September 1999.
- [36] Smith B., Kumar A., and Bittner T. Basic Formal Ontology for Bioinformatics, <http://www.uni-leipzig.de/~akumar/JAIS.pdf>
- [37] Wu C.H., Apweiler R., Bairoch A., Natale D.A., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Mazumder R., O'Donovan C., Redaschi N., and Suzek B. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34:D187-D19, 2006. <http://www.pir.uniprot.org/>