Chapter X

AN ONTOLOGY BASED APPROACH FOR DATA INTEGRATION

An application in Biomedical Research

Vipul Kashyap[1], Kei-Hoi Cheung[2], Donald Doherty[3], Matthias Samwald[4], M. Scott Marshall[5], Joanne Luciano[6], Susie Stephens[7], Ivan Herman[8] and Raymond Hookway[9]

[1]*Partners Healthcare System, Clinical Informatics R&D, 93 Worcester St, Suite 201, Welleley, MA 02481, USA, +1(781)416-9254, E-mail: vkashyap1@partners.org;*

[2]*Yale Center for Medical Informatics, Yale University School of Medicine, 300 George Street, Suite 501, New Haven, CT 06511, USA, +1(203)737-5783, kei.cheung@yale.edu;*

[3]*Brainstage Research, 5001 Baum Blvd, Suite 725, Pittsburgh, PA 15213,USA +1(412)683-1410, donald.doherty@brainstage.com;*

[4]*Medical Expert and Knowledge Based Systems, Medical University of Vienna,Spitalgasse 23 A-1090, Vienna Austria, +43-69981168028 matthias.samwald@meduniwien.ac.at;*

[5]*University of Amsterdam, Gersthove 41, 1112 HN Diemen, The Netherlands, +31-6-2044-0929, marshall@science.uva.nl;*

[6]*Department of Genetics, Harvard Medical School, NRB Room 238, 77 Louis Pasteur Ave, Boston, MA 02115, USA, +1.(617)993-9994, jluciano@gmail.com;*

[7]*Oracle Corporation, 10 Van de Graaf Drive, Burlington, MA, USA,+1(781)744-037, susie.stephens@oracle.com;*

[8]*World Wide Web Consortium (W3C),c/o Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ, Amsterdam, The Netherlands, +31-641044153, ivan@w3.org;*

[9]*Hewlett Packard, Marlborough, USA, +1(508)467-4921*

Abstract: In this chapter, we explore the area of translational medicine that aims to improve communication between the basic and clinical sciences so that more diagnostic and therapeutic insights may be derived. Translation research goes from bench to bedside, where the effectiveness of results from preclinical research are explored with patients, and from bedside to bench, where information obtained from patients can be used to refine our understanding of the biological principles underpinning the heterogeneity of human disease and polymorphism(s). Translational medicine requires integration, aggregation and analysis of data and information across multiple sub-fields of biomedical research such as: systems physiology, anatomy, molecular and cell biology, pharmacology and clinical studies. Ontologies and semantic web technologies are expected to have a significant role to play in making this a reality. We present an example use case for biomedical data integration and illustrate how

a scientific question can be answered by open source data available on the web today. We discuss research issues such as the use of a centralized data warehouse approach versus a federated approach for data integration; and modeling alternatives for ontologies and mappings of ontological concepts to underlying data sources. Pragmatic issues on how ontologies and semantic web technologies can be deployed in a cost effective and efficient manner are also discussed.

Key words: Semantic Web technologies, Translational Medicine, Data Integration, Resource Description Framework (RDF), Web Ontology Language (OWL), OWL reasoners, Ontologies, Query Processing, Semantic Inference, Knowledge, Data and Process Models

1. INTRODUCTION

The healthcare and life sciences sector is playing host to a battery of innovations triggered by the sequencing of the human genome as well as genomes of other organisms. A significant area of innovative activity is that of translational medicine which aims to improve the communication between basic and clinical science so that more diagnostic and therapeutic insights may be derived. Translational research [1] goes from bench to bedside, where the effectiveness of results from preclinical research are tested on patients, and from bedside to bench, where information obtained from patients can be used to refine our understanding of the biological principles underpinning the heterogeneity of human disease.

A large extent of the ability for biomedical researchers and healthcare practitioners to work together - exchanging ideas, information and knowledge across organizational, governance, socio-cultural, political and national boundaries - is currently mediated by the internet and its exponentially-increasing digital resources. These digital resources embody scientific literature, experimental data, and curated annotation (metadata) whether human- or machine-generated. This is the digital part of the scientific "information ecosystem" [2]. Its structure, despite the revolution of the web, continues to reflect a degree of domain hyper-specialization, lack of schematization, and schema mismatch, which works against information transfer.

The key requirement is to enable organization of knowledge on the web by its meaning, purpose and context of use; and to effectively bridge and map meanings across specialist domains of discourse. We want the expression of this meaning to be digital, machine readable, capable of being filtered, aggregated and transformed automatically. We want it to be seamlessly embedded in the structure of web documents. And we would like to provide built-in visibility of information change - provenance - and

explanation. In sum, we would like to make the context of information - which is established by both use and meaning - available with information content.

Modern biomedical science produces vast amounts of data that is produced by practitioners and researchers in finely subdivided sub-specialties. It is common for biologists in different sub-specialties to be completely unaware of the key literature in each other's domain. Yet, particularly in applying research to curing and preventing diseases - the bench to bedside transition - an integrated understanding across sub-specialties becomes essential. In complex diseases this is a difficult task, inadequately supported by the current information ecology of science. This difficulty applies with the most force to highly controversial and rapidly evolving areas of research, such as the understanding and cure of neurodegenerative diseases (Parkinson's Disease (PD), Alzheimer's Disease (AD), Huntington's disease, Amyotrophic Lateral Sclerosis (ALS) or Lou Gehrig's disease, and others).

As an example, AD affects four million people in the United States and causes both great suffering and enormous costs to society. Yet there is still no agreement on exactly how it is caused, or where best to intervene to treat it or prevent it. The Alzheimer Research Forum records fifty significant hypotheses related to aspects of the etiology of AD, most of them combining supporting data and interpretations from multiple specialist areas of biomedicine. One typical recent hypothesis on the etiology of AD [3] combines data from research in mouse genetics, cell biology, animal neuropsychology, protein biochemistry, neuropathology, and other areas.

Many areas of biomedical research including drug discovery, systems biology, and personalized medicine rely heavily on integrating and interpreting data sets produced by different experimental methods, in different groups, with heterogeneous data formats, and at different levels of granularity. Research in other neurodegenerative disorders such as PD, the second most common neurodegenerative disorder, is also quite frequently multimodal, and like AD research often includes interpretations based on clinical phenotype data collected from different patient populations.

There is a need for a synthesis of understandings across disciplines, and across the continuum from basic research to clinical applications. This applies to most complex diseases and too many health care issues. A useful synthesis must combine not only data, but also interpretations of the data. It must support both, the well-structured standardized presentation of data as well as the discovery and fusion of convergent and divergent interpretations of data. With advances in hardware instrumentation and data acquisition technologies (e.g., high-throughput genotyping, DNA micro arrays, and mass spectrometry), there is an exponential growth of healthcare as well as

life science data. In addition, the results of these experimental approaches are typically stored in heterogeneous formats in disparate data repositories. Over time, it has become increasingly difficult to identify and integrate these data sources. Even if we assume that the data is stored in the same format, the complex nature of healthcare and life science data makes deep domain knowledge a prerequisite to understand and integrate the data in a meaningful way. The problem is becoming more acute, both due to continuing increases in data volumes and the growing diversity in types and formats of data that need to be interpreted and integrated.

In this chapter, we illustrate by the means of a real world use case based on PD, how ontologies can be used to enable data integration and information synthesis across various data repositories belonging to various biomedical research areas. We begin with a brief description of the disease and an illustrative query example in Section 2. This is followed by a discussion on development of domain ontologies to characterize information and knowledge about the disease in Section 3. A discussion of the data sources is presented in Section 4 including a discussion of pragmatic issues related to data integration in Section 4.3. Conclusions and future work are presented in Section 5.

2. USE CASE: PARKINSON'S DISEASE

The neuroscience domain provides a rich and diverse set of scientific studies (involving both biomedical and clinical research) with associated datasets. In this section, we present a use case pertaining to PD, which is the focus of research and activity of a broad collection of neuroscience researchers, practitioners and neurologists. The use case provides an example to illustrate how semantic web technologies can potentially be used to enable the bench-to-bedside vision and support the ability to cross-link, aggregate and interpret the information across various perspectives [4]. In this chapter, we will focus on the systems physiology, cell and molecular biology perspectives on PD.

2.1 Systems Physiology Perspective

A scientist researching PD from a systems physiology perspective wants to know the structures (anatomy) involved in the disease and the ways those structures interact (physiology) or fail to interact resulting in the disease state. For instance, it is well known that brain cells that cluster together to form a brain structure known as the substantia nigra degenerate and die in PD patients. These particular brain cells, or neurons, contain a chemical

substance called dopamine. Neurons in the substantia nigra communicate with other neurons in other brain structures through the transmission of dopamine. Neurons that are able to listen and respond to signals from dopamine containing neurons must have dopamine receptors. A scientist interested in a systems physiology perspective knows that there are fewer dopamine containing neurons in a PD patient's brain than normal so they may reasonably ask if the neurons in the brain that receive dopamine may have something to do with disease related behaviors. The scientist may ask a semantic web enabled search engine "what neurons in the brain have dopamine receptors and what anatomical structure do they belong to?"

A part of the substantia nigra is composed of neurons that transmit the chemical dopamine. It's when these dopamine transmitting neurons in the substantia nigra die, that the amount of dopamine released into another part of the basal ganglia known as the striatum decreases and PD symptoms appear. The striatum connects to a third part of the basal ganglia known as the pallidum. In fact, there are two distinct connections that form parallel anatomical pathways from the striatum to the pallidum. One connection is known as the direct pathway and the other the indirect pathway. Each pathway is activated in a different way by the dopamine released from the substantia nigra. Activating the direct pathway facilitates movement, for instance moving your arm or legs. In contrast, activating the indirect pathway inhibits movement. At the systems physiology level it is the balance of activity between the indirect and direct pathways from the striatum to the pallidum that at one extreme leads to PD (over-activity in the indirect pathway results in over-inhibited movement in the D1 receptor) and at the other extreme leads to Huntington's disease (over-activity in the direct pathway results in too much movement in the D2 receptor).

2.2 Cellular and Molecular Biology Perspective

Studies identifying genes involved with PD are rapidly outpacing the cell biological studies that would reveal how these gene products are part of the disease process. The alpha synuclein and Parkin genes are two such examples. The discovery that genetic mutations in the alpha synuclein gene could cause PD has opened new avenues of research in PD. When it was also discovered that synuclein was a major component of Lewy bodies, the pathological hallmark of PD in the brain, it became clear that synuclein may be important in the pathogenesis of sporadic as well as rare cases of PD. More recently, further evidence for the intrinsic involvement of synuclein in PD pathogenesis was shown by the finding that the synuclein gene may be duplicated or triplicated in familiar PD, suggesting that simple over expression of the wild type protein is sufficient to cause disease. Since the

discovery of synuclein, studies of genetic linkages, specific genes, and their associated coded proteins are ongoing for PD research - transforming what had once been thought of as a purely environmental disease into one of the most complex multigenetic diseases of the brain. Studies of genetic linkages, specific genes, and their associated coded proteins are ongoing for PD research. Mutations in the Parkin gene cause early onset PD, and the Parkin protein has been identified as an E3 ligase, suggesting a role for the proteasomal pathway of protein degradation in PD. DJ-1 and PINK-1 are proteins related to mitochondrial function in neurons, providing an interesting genetic parallel to mitochondrial toxin studies that suggest disruptions in cellular energetics and oxidative metabolism are primarily responsible for PD. Other genes, such as UCHL-1, tau, and the glucocerebrosidase gene, may be genetic risk factors, and their potential role in the sporadic PD population remains unknown. Mutations in LRRK2, which encodes for a protein called dardarin, is the most recently discovered genetic cause of PD, and LRRK2 mutations are likely to be the largest cause of familial PD identified thus far. Dardarin is a large complex protein, which has a variety of structural moieties that could be participating in more than a dozen different cellular pathways in neurons. As the cellular pathways that lead to PD is not fully understood, it is currently unknown, how, or if, any of these pathways intersect in Parkinson's disease pathogenesis.

2.3 Example Query

Based on the current research in PD, a neuroscience researcher or practitioner might be interested in asking a query, such as the following:

> *Show me the neuronal components of receptors that bind to a ligand which is a therapeutic agent in {Parkinson's, Huntington's} disease in reach of the dopaminergic neurons in the {pars compacta, pars reticularis} substantia nigra.*

The above query can be visualized in the context of a biomedical researcher trying to propose and validate (or invalidate) research hypotheses. Hypothesis validation involves either trying to conduct bench experiments or re-use existing scientific results available on the web. These results are typically available in the form of scientific publications through a widely available repository such as PubMed. In the context of this project, we seek to demonstrate the value of structured scientific facts describing experiments, hypotheses, etc. Consider the following use case scenarios where the above query might be submitted to the semantic web:

- A researcher interested in anatomy/physiology trying to hypothesize the location of various components on a neuron might want to look at

scientific data or hypotheses on the presence of certain type of receptors on a neuron.

- A researcher interested in molecular biology trying to hypothesize the location of some receptors on neurons might want to look at scientific data or hypotheses on the anatomical structure of a neuron or the pharmacology of chemical compounds that bind to a receptor.
- A clinical researcher interested in developing new therapies for a disease may be interested in understanding the mechanisms of how chemical compounds or ligands bind to receptors.

Common to these scenarios is the ability of a researcher to access information from different knowledge domains and correlate it with data and information from his or her information domain. This can form the basis for various decision making activities such as forming new hypotheses or validating old ones.

The information query discussed above, requires the synthesis and integration of data and data interpretations across the following specializations in biomedical research:

- Anatomy/Physiology relating to the compartments of neurons
- Molecular Biology relating to the receptors on neurons
- Pharmacology relating to chemical compounds that bind to receptors.
- Clinical information relating to ligands that are associated with a disease.

Ontologies and semantic web technologies play a critical role in enabling the synthesis and integration of data and data interpretations. The key role played by ontologies is that they provide a common language to express the shared semantics and consensus knowledge developed in a domain. This shared semantics is typically captured in the form of various domain specific ontologies and classifications such as MeSH, GO and the Enzyme Commission. Ontological concepts provide the shared semantics to which various data objects and data interpretations can be mapped to enabling integration across multiple biomedical data sources and domains. We now discuss issues related to creation and modeling of ontologies.

3. ONTOLOGIES

In this section, we present our approach to creating a domain ontology that provides the basis for data integration. The design and creation of the domain ontology is based on the description of the use case and the sample query presented in the previous section. Different modeling alternatives and choices that emerged when creating the ontology are presented; and issues relating to re-use of pre-existing ontologies are also discussed.

3.1 Ontology Design and Construction

Various ontology creation methodologies have been proposed in [21,22]. Most of them have primarily focused on the manual processes of interaction between subject matter experts and ontologists. We propose a methodology for iterative creation of ontologies based on available resources such as textual descriptions of subject matter knowledge; information needs and queries of the users; and cross-linking to pre-existing ontologies whenever there is a need for more extensive coverage. Our approach complements the approach adopted in [23], where the ontology creation process is driven by the underlying database schemas. It may be noted that in the semantic web scenario; a large number of web repositories contain semi-structured (e.g., XML-based) data and typically their underlying schemas are not available. A brief illustration of the ontology design process is illustrated below in Figure X-1.

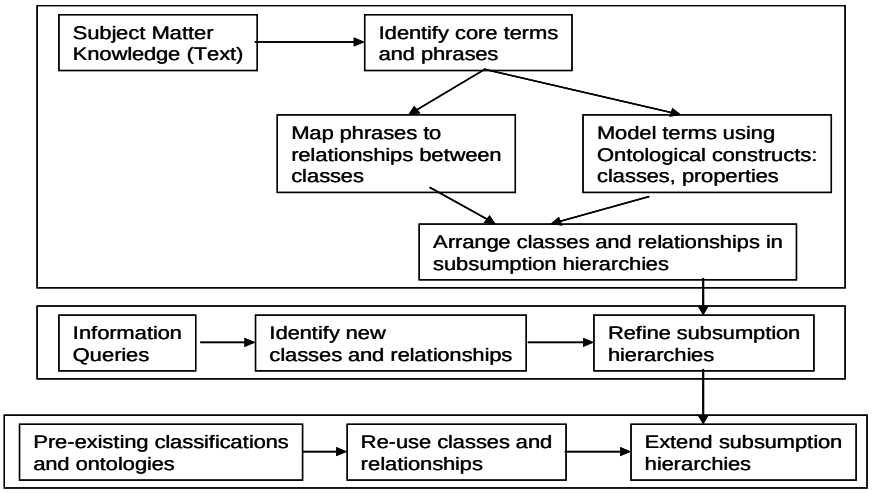

Figure X-1. Ontology Design and Creation Approach

We begin by looking at the textual description of the PD use case descriptions and (manually) extracting the key concepts that relate to the disease. For pragmatic reasons, we decided to focus on creating ontology specifically for the use case and then expanding it later as information needs and usage becomes clearer. We will also cross-link to pre-existing ontologies wherever there is a need for more extensive coverage instead of "re-inventing the wheel".

3.1.1 Identifying Concepts and Subsumption Hierarchies

The first step in designing the PD ontology was to characterize the core vocabulary in terms of the classes and the subsumption hierarchy that describes the information related to the use case. Consider the following textual description:

*Studies identifying **genes** involved with **PD** are rapidly outpacing the cell biological studies which would reveal how these gene products are part of the disease process in PD. The **alpha synuclein** and **Parkin** genes are two examples.*

*The discovery that **genetic mutations** in the **alpha synuclein gene** could cause PD has opened new avenues of research in the PD field. Since the discovery of **synuclein**, studies of genetic linkages, specific genes, and their associated coded proteins are ongoing in PD research field - transforming what had once been thought of as a purely environmental disease into one of the most complex multigenetic diseases of the brain.*

*Mutations in the Parkin gene cause early onset Parkinson's disease, and the **parkin** protein has been identified as an **E3 ligase**, suggesting a role for the proteasomal pathway of protein degradation in PD. **DJ-1** and **PINK-1** are proteins related to mitochondrial function in neurons, providing an interesting genetic parallel to mitochondrial toxin studies that suggest disruptions in cellular energetics and oxidative metabolism are primarily responsible for PD. Other genes, such as **UCHL-1, tau**, and the **glucocerebrosidase** gene, may be genetic risk factors, and their potential role in the sporadic PD population remains unknown. **Mutations in LRRK2**, which encodes for a protein called **dardarin**, is the most recently discovered genetic cause of PD, and LRRK2 mutations are likely to be the largest cause of familial PD identified thus far. **Dardarin** is a large complex protein, which has a variety of structural moieties that could be participating in more than a dozen different cellular pathways in neurons.*

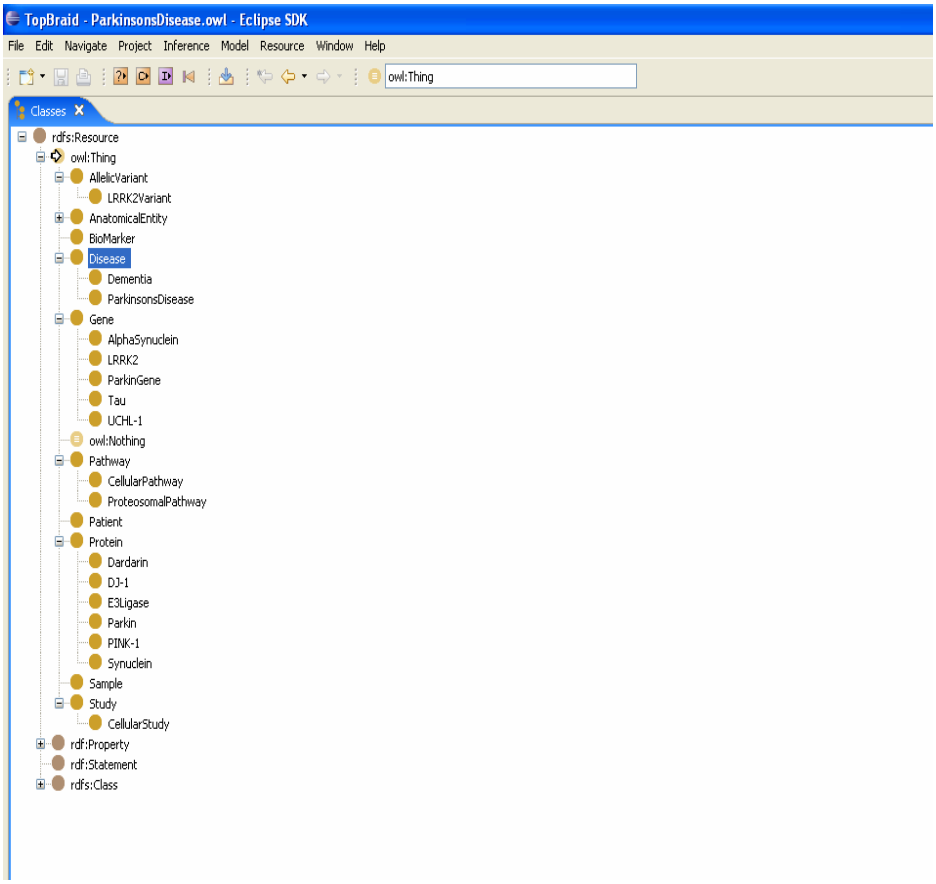

Figure X-2. Parkinson ’s disease Ontology: Concepts and Subsumption Hierarchy

Based on the above analysis, one can begin to identify some important ontological concepts relevant to PD:

- Genes, such as alpha synuclein, Parkin, UCHL-1, tau and glucocerebrosidase
- Proteins, such as synuclein, Parkin, Dardarin, DJ-1 and PINK-1
- Allelic Variations or Genetic mutations such as in the LRRK2 gene
- Diseases, such as PD.

These and some of the other identified concepts are illustrated in Figure X-2 above.

3.1.2 Identifying and extracting relationships

The next step was to identify and represent relationships between the concepts identified in the ontology above. Consider the following textual description:

The discovery that genetic mutations in the alpha synuclein gene could cause PD has opened new avenues of research in PD. When it was also discovered that synuclein was a major component of ***Lewy bodies, the pathological hallmark of PD in the brain****, it became clear that synuclein may be important in the pathogenesis of sporadic as well as rare cases of PD.*

DJ-1 and PINK-1 are proteins related to mitochondrial function in neurons, providing an interesting genetic parallel to mitochondrial toxin studies that suggest disruptions in cellular energetics and oxidative metabolism are primarily responsible for PD. ***Other genes, such as UCHL-1, tau, and the glucocerebrosidase gene, may be genetic risk factors,*** *and their potential role in the sporadic PD population…*

Based on the above analysis, the following relationships may be added to the PD Ontology:

- Lewy Bodies is_pathological_hallmark_of PD.
- UCHL-1 is_risk_factor_of PD.

The representation of these relationships and their inverses is illustrated in Figure X-3 below.

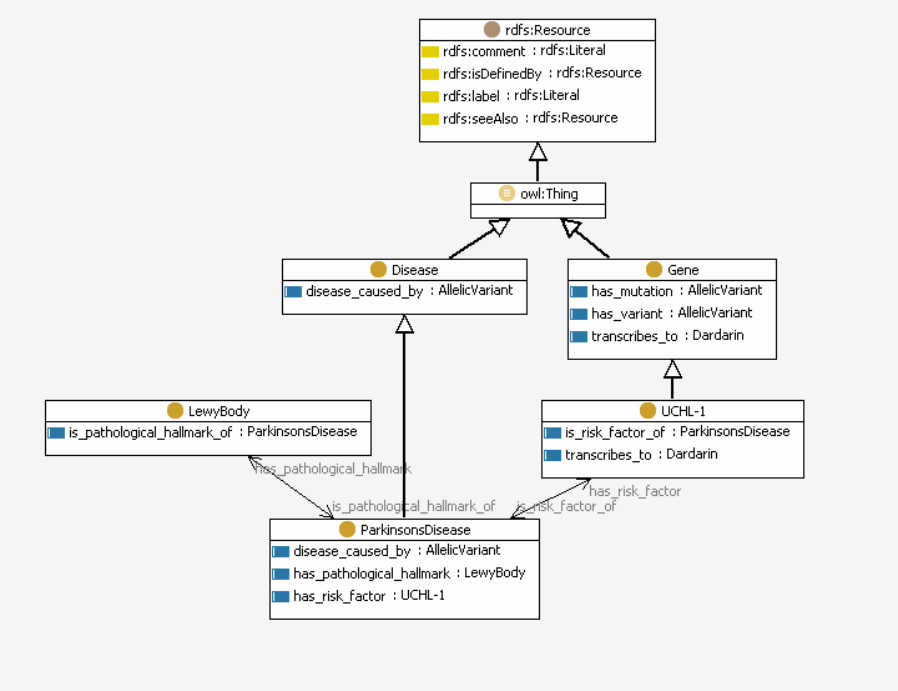

Figure X-3. Parkinson's disease Ontology: Relationships between Concepts

3.1.3 Extending the ontology based on information queries

We next considered various information queries and identified concepts and relationships that needed to be part of the Parkinson ’s disease ontology. This was important as, otherwise, without these concepts and relationships it would not have been possible for a biomedical researcher to specify these queries for retrieving information and knowledge from the system. Consider the following queries:

- What ***cell signaling pathways*** are implicated in the pathogenesis of Parkinson ’s disease? In what cells?
- What proteins ***are involved*** and in which pathways?

The above queries lead to addition of new concepts and relationships illustrated in Figure X-4 below.

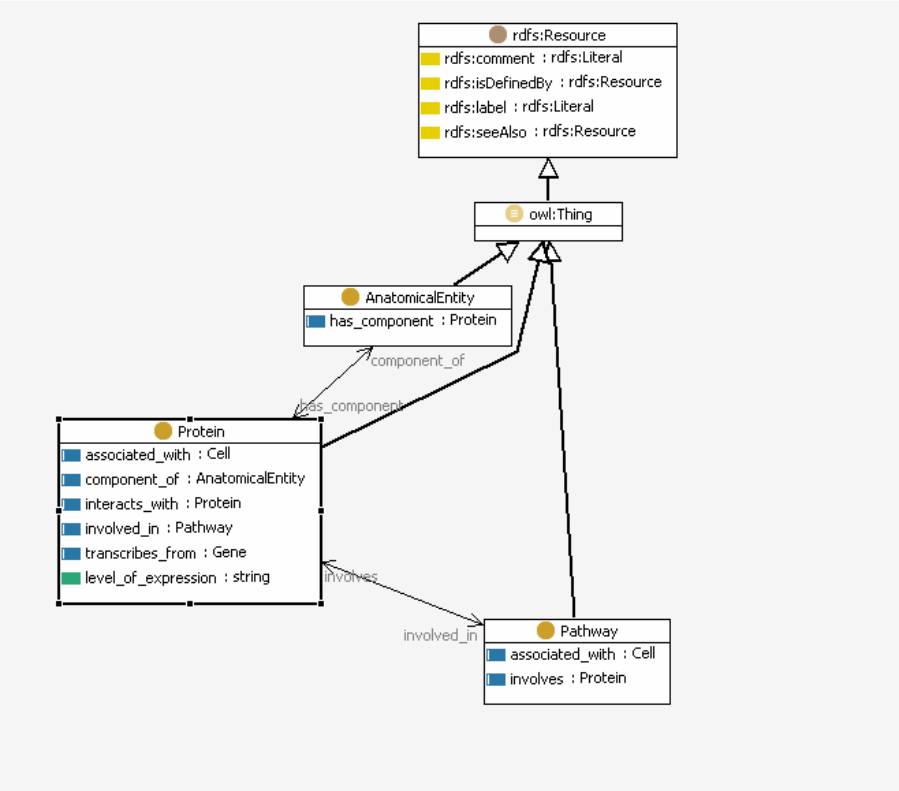

Figure X-4. Parkinson's disease Ontology: Adding Concepts and Relationships to support Information Queries

3.1.4 Ontology Re-use

As we expand and refine our use case, it became clear to us that we needed a methodical way of extending our ontology. Clearly our intention is to re-use various ontologies and vocabularies being developed in the healthcare and life science fields. Some of the dimensions along which the ontology could be extended and the associated re-usable ontologies are as follows:

- **Diseases:** Over time, we could anticipate addressing information needs related to other related diseases such as Huntington ’s disease or AD. Some ontologies/vocabularies we are considering reuse of are the International Classification of Diseases [8] and a subset of Snomed [7].
- **Genes:** Over time, we may want to add more genes and other genomic concepts such as proteins, pathways, etc. to the ontologies. We are considering linking to Gene Ontology [9].

- **Neurological Concepts**: We may need to add more concepts related to brain structures and parts to extend our ontology if required. We are considering the use of NeuroNames [10], a nomenclature of primate brain structures and related terms that describe the superficial features of these structures.
- **Enzymes:** Concepts for various enzymes and other chemicals may be required, for which purpose, we will consider linking to the Enzyme Nomenclature [11].

Some criteria that need to be investigated for re-use of ontologies are as follows:

- How specific or general should the ontology be that is being re-used? For instance, do we choose ontology of neurological diseases or do we choose a generalized ontology of diseases and include an appropriate subset?
- At what level of granularity do we include concepts from ontology? For instance, we can include ontologies at a very shallow level, with just the concepts being included. A deeper level of inclusion may entail inclusion of associated properties, relationships and axioms as well.

Depending on the level of inclusion, ontology re-use could lead to circularities and inconsistencies. We propose to use pragmatic approaches to handle these situations. For instance, we may choose to include certain concepts at a "shallow" level to avoid potential inconsistencies or give priority to a concept from one ontology over another

3.2 Ontology Design Choices

We now discuss various design choices that came up in the process of ontology design and discuss possible criteria that might help choose one representation over another.

- **Use of Relationships versus Classes for Modeling Knowledge:** Suppose we want to represent the knowledge of transcription of a particular gene, such as UCHL-1 into a protein such as Dardarin. Two possible choices for representing this are:
  - UCHL-1 transcribes_into Dardarin
  - Represent a class called transcription with properties has_transcription = Dardarin, has_gene = UCHL-1.

  The former could be chosen if usage scenarios are just interested in gene products, whereas the latter would be preferable if a researcher is interested in the laboratory conditions under which the transcription takes place. Information about these lab conditions could be modeled as object properties associated with the transcription class. It should be noted that

the former alternative is preferable from the point of view of storage and querying.

- **Modeling of Diseases:** Diseases could be modeled in an ontology as static classes or as dynamic processes. The former is probably more suitable in the context where a researcher is interested in the association of diseases and genes, but on the other hand, if someone is interested in understanding the change in physiological states due to administration of a drug.

- **Use of Instances versus SubClasses:** A generic/specific relationship can be modeled by using instances or subclasses. Consider the following examples:
  - Parkinson's Disease may be viewed as a subclass or an instance of the class Disease
  - UCHL-1 may be viewed as subclass or an instance of the class Gene.

  The appropriate modeling choice depends on the usage scenario. In general if one were interested in counting the occurrence of diseases or alleles in a population, the subclass modeling choice would be needed. However, if the usage scenario just requires annotation for the sake of search and query expansion, then the instances choice might suffice.

- **Granularity of representation of relationships:** Relationships could be represented at different levels of genericity or specificity. For example, one could represent the relationship between an allelic variant and disease at the following levels of granularity:
  - AllelicVariant causes Disease
  - LRR2KVariant causes Parkinson's Disease

  The latter representation might be preferred if the usage scenarios are focused on only causes for Parkinson 's disease or if only very few and specific allelic variants are known to cause Parkinson 's disease.

- **Representation of uncertainty:** Current semantic web specifications are silent when it comes to representation and reasoning with uncertain information. For instance, consider the following statement from the use case in Section 2.

  The discovery that genetic mutations in the alpha synuclein gene **could cause** *Parkinson's disease in families has opened new avenues of research in the Parkinson's disease field.*

  Even though the underlying meta-models and tools do not support representation and manipulation of uncertainties, pragmatic approaches

such as using reification and using the query language operators to retrieve information satisfying certain threshold constraints.

- **Domain/Range polymorphism:** Relationships can have multiple domain and range classes and the current RDF/OWL semantics of combining these classes may not accurately reflect the intended meaning. For example:
  associated_with(Pathway, Cell)
  associated_with(Protein, Biomarker)
  The associated_with relationship has two domain classes, Pathway and Protein; and two range classes, Cell and Biomarker. It may be necessary to represent specific OWL constraints [6] to make sure that the proper semantics are identified and represented.

- **Default Values:** In some cases, it is important to represent default values of object properties, for e.g., consider the statement:
  The default function of proteasomal pathway is protein degradation
  Currently, SW tools do not support the representation of and reasoning with default values. In the context of data integration, it can help us retrieve information or identify mappings in the absence of availability of data from the underlying data sources. In certain cases, certain implicit assumptions can be made explicit using default values.

- **Ternary and other Higher Order Relationships:** Typically, a biomedical discovery, such as an association between a particular gene and a disease could be at the same time be established to be true in the context of one study and could be established to be false in another. This requires the representation of a ternary relationship between genes, diseases and studies as follows:
  established_in(Disease, Gene, Study)
  There is a need for representation of ternary and other higher order relationships, which is typically implemented by representing them as a class [5].

3.3 Summary

In this section, we presented our approach for creating a domain ontology for data integration and high lighted some modeling and representation choices that need to be made. From a pragmatic perspective, the semantic web is not at a stage whether standardized ontologies for various domains are freely above. So, there is a need for creating ontologies based on well defined use cases and functional requirements on one hand; and also to link

to pre-existing classifications and ontologies available for more complete coverage. Multiple modeling and representation choices emerge when designing these ontologies and there is no one good answer for all situations. A pragmatic approach should be adopted and modeling choices should be based on the use cases and functional requirements at hand.

4. DATA SOURCES

In Section 3, we discussed our approach for modeling and representing the domain ontology. In particular, we used the collection of information queries identified as being useful to our user group, i.e., biomedical researchers. We now discuss with the help of the example query discussed above in Section 2.3, how the query can be answered based on retrieval and integration of data from different biomedical data sources. We first describe a set of data sources that are relevant to the use case. We then describe how the use case query illustrated in Section 2 can be answered using these data sources. Some of the elements of our solution such as conversion of the data sources into RDF, the ability to map RDF graphs to ontological concepts and the ability to merge RDF graphs based on declaratively specified mapping rules are then highlighted.

4.1 Relevant Data Sources to the Query

A list of data sources that are being integrated using semantic web technologies are as follows:

- **Neuron Database:** Neuron DB [**Error! Reference source not found.**] provides a dynamically searchable database of three types of neuronal properties: voltage gated conductances, neurotransmitter receptors, and neurotransmitter substances. It contains tools that provide for integration of these properties in a given type of neuron and compartment, and for comparison of properties across different types of neurons and compartments.

- **PDSP KI Database:** The PDSP KI Database [13] is a unique resource in the public domain which provides information on the abilities of drugs to interact with an expanding number of molecular targets. The KI database serves as a data warehouse for published and internally-derived Ki, or affinity, values for a large number of drugs and drug candidates at an expanding number of G-protein coupled receptors, ion channels, transporters and enzymes.

- **PubChem:** PubChem [14] is organized as three linked databases within the NCBI's Entrez [15] information retrieval system. These are PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem also provides a fast chemical structure similarity search tool. Links from PubChem's chemical structure records to other Entrez databases provide information on biological properties. These include links to PubMed [16] scientific literature and concepts from the MeSH taxonomy [17].

Now consider the use case query presented in Section 2. The answer to this query can be constructed by correlating data retrieved from underlying data sources as follows:

- Data that identifies *Distal Dendrite* as a compartment on the *dopaminergic neuron* and the receptor *D1* belonging to the *Distal Dendrite* can be retrieved from the Neuron Database.
- Data that identifies *5-Hydroxy Tryptamine* as a ligand that can bind to the *D1* receptor can be retrieved from the PDSP KI database
- Data that identifies that the ligand *5-Hydroxy Tryptamine* is associated with *Parkinson 's disease* can be retrieved from PubChem. The ligand *5-Hydroxy Tryptamine* is cross-linked to the MeSH concept related to *Parkinson 's disease*.

Correlation of the data results in:
The ligand *5-Hydroxy Tryptamine*, a therapeutic agent for *Parkinson 's disease* binds to the receptor *D1* in the *Distal Dendrite* area of the dopaminergic neuron in the substantia nigra.

Details on how this is implemented using semantic web technologies are presented next.

4.2 Implementing the Data Integration Solution

The data integration solution can be implemented using two broad architectural approaches:

- A centralized approach where the data available through web-based interfaces is converted into RDF and stored in a centralized data repository.
- A federated approach where the data continues to reside in the existing data repositories. An RDF-based mediator or gateway converts the underlying data into the RDF format.

Common to both these approaches are the functionalities required for converting non-semantic web data such as relational or XML data into RDF; mapping ontological concepts into RDF graphs, possibly via association with appropriate SPARQL queries; and merging of RDF graphs based on matching of IDs and URIs and declaratively specified mapping rules.

4.2.1 Conversion into RDF Graphs

An approach for converting XML-based data into RDF has been presented in [18]. Typically, queries requiring navigation of multiple relationships across multiple objects requires writing complex SQL and applications programming code when using relational databases. The RDF format allows us to focus on the logical structure of the information in contrast to only representational format (XML) or storage format (relational database).

There are many issues involved in the conversion of XML data into RDF format including the use of unique identifiers, preservation of original semantics of the converted data, resolution of bidirectional relationships and filtering of redundant element tags from the original XML record. Unlike traditional XML to XML conversion, XML to RDF conversion should take into account the advantages of the RDF model in representing the logical structure of the information and the modeling of the relationships between concepts. The underlying objective of converting XML data into RDF is to capture the semantics of the data and leverage such semantics in querying the repository to not only retrieve the explicit but also the implicit knowledge through inference. Some issues that need to be considered in the conversion are:

- The use of a specific identifier allows the unique identification of the nodes (*subject* and *object*) and *predicates* in an RDF repository. But, there is no globally accepted biomedical identifier schema that may be used. The bioinformatics community is currently debating this issue and there are many candidate schemas that may be used including the Life Science Identifier (LSID) [19] and solutions based on the HTTP protocol (i.e., URIs (Universal Resource Identifiers), URLs (Universal Resource Locators) and URNs (Universal Resource Names)). NLM resources such as the Unified Medical Language System [20] could provide the basis for the identification of biomedical entities.
- It is important to decide whether to reflect the native nesting of elements in the original XML format or modify the structure to reflect one of the many possible perspectives on the data.
- In case there is an existing domain ontology, the RDF structure can be based on the domain ontology. However, if different ontologies need to

be mapped to the same set of data, one may need to specify mappings between ontological concepts and a given choice of an RDF representation.

We now discuss issues related to mapping ontological concepts to their associated RDF Graphs.

4.2.2 Mapping Ontological Concepts to RDF Graphs

As presented in Section 4.1, the use case query is satisfied by correlating fragments of data retrieved from the individual data sources. These fragments correspond to the following concepts and relationships from the ontology:

- Compartment located_on Neuron
- Receptor located_in Compartment
- Ligand binds_to Receptor
- Ligand associated_with Disease

In order to enable this data integration, we need to map these concepts and relationships to RDF graphs in the underlying data sources. We assume that there is a wrapper which transforms the results retrieved from these data sources into RDF graphs. We would need to capture the following information in these mappings:

- **Ontological Element:** This would represent the ontological concept mapped to the RDF graphs in a given data source. One may also chose to specify the properties or edges that the given data source supports. In some precise cases, one may want to specify the actual triplet which a data source may support.
- **Data Source:** This would represent the data source which would support the retrieval of data corresponding to the concepts in the ontology
- **Target SPARQL Query:** This would represent the SPARQL query which would need to be executed on the underlying data source to retrieve the required RDF graph.
- **Identifier Scheme:** This would represent the identifiers for the nodes and edges for a particular RDF graph. We adopt the approach of using URIs to identify “object” nodes in an RDF graph, i.e., those that are not “literals”. Each URI is constructed on a set of base URIs corresponding to each data source that delineates the respective name spaces.

An example table illustrating the representation of mappings is illustrated in Table X-1 below:

Table X-1. Mapping Table

Ontological Element	**Data Source**	**SPARQL Condition**	**Namespace**
Neuron	Neuron Database	?x rdf:type Neuron	Neuron DB
Neuron, {located_in}	Neuron Database	?x located_in ?y, ?x rdf:type Neuron	Neuron DB
Neuron located_in Compartment	Neuron Database	?x located_in ?y, ?x rdf:type Neuron, ?y rdf:type Compartment	Neuron DB
Receptors located_on Compartment	Neuron Database	?x located_on ?y, ?x rdf:type Receptor, ?y rdf:type Compartment	Neuron DB
Ligand binds_to Receptor	PDSP KI Database	?x binds_to ?y, ?x rdf:type Ligand, ?y rdf:type Receptor	PDSPKI DB
Ligand associated with Disease	PubChem	?x associated_with ?y, ?x rdf:type Ligand, ?y rdf:type Disease	PubChem DB
…	…	…	…

The above table represents some of the possibilities for representing some of the information required to enable data integration. This table is referenced by the system to invoke the appropriate queries on the underlying data sources to retrieve RDF graphs for display to the biomedical researcher.

An interesting omission is standardized ontologies that might be used by a given data source to represent some information, for e.g., the PubChem data source uses MeSH concepts to represent diseases. We currently make the assumption that there are certain standardized ontologies included in the domain ontology and that the RDF wrapper “maps” concepts from one ontology to another. For instance, if the domain ontology uses Snomed for representing disease concepts, and the local data source uses MeSH, the RDF wrapper will map the MeSH ID corresponding to Parkinson ’s disease to the Snomed ID corresponding to Parkinson ’s disease. Alternatively this can be done by invoking a specialized “terminology mapper” program at the global level.

4.2.3 Generation and Merging of RDF Graphs

As discussed earlier, RDF wrappers perform the function of transforming information as stored in internal data structures in LIMS and EMR systems into RDF-based graph representations. We illustrate with examples (*Figure X-4*), the RDF representation of neurological and chemical data from the various datasources.

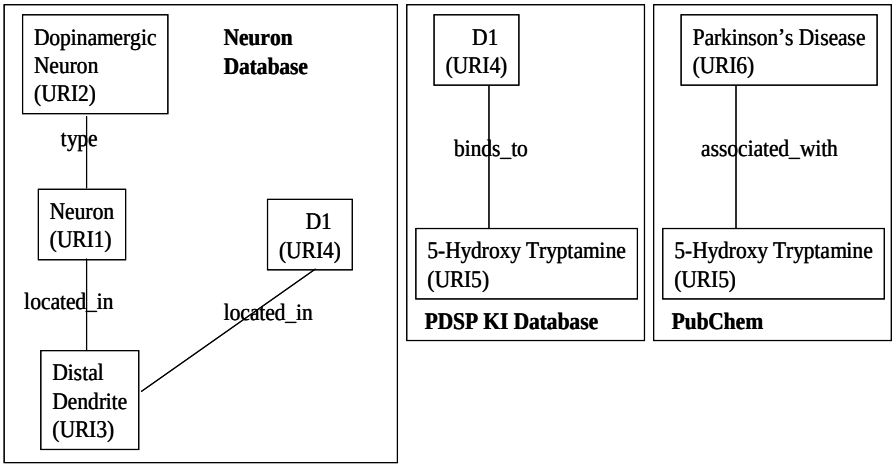

Figure X-5. RDF Representation of Neurological and Chemical Data

Each box is labeled with the name of the data source from which the associated RDF Graph can be generated. In the box corresponding to the Neuron Database, the RDF graph generated in response to the query may consist of nodes corresponding to instances of a Neuron which is linked with edge labeled *type_of* to the node representing the concept of a Dopinamergic Neuron; and with an edge labeled *located_in* to the node representing the Distal Dendrite region. The node corresponding to the D1 receptor is in turn linked with the edge labeled *located_on* to the node corresponding to the Distal Dendrite region. In the box corresponding to the PDSP KI database, the edged labeled *binds_to* represents the relationship between the ligand 5-Hydroxy Tryptamine and the D1 receptor. The box corresponding to the PubChem database uses the edge labeled *associated_with* to represent the association between the ligand 5-Hydroxy Tryptamine and Parkinson 's disease.

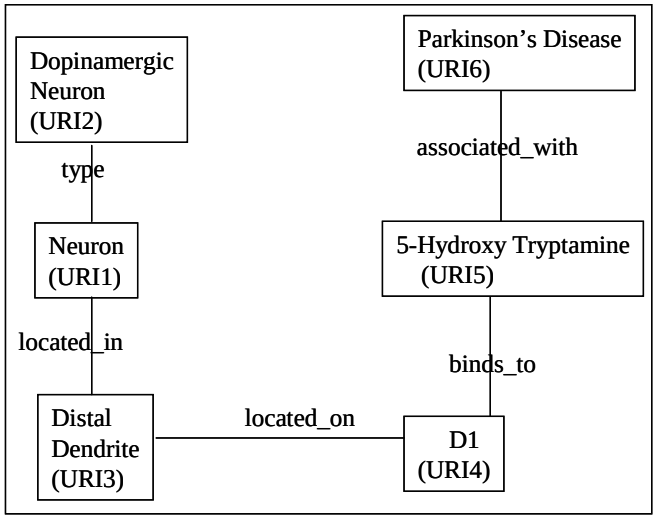

Figure X-6. The Integrated RDF Graph

The data integration process is an interactive one and involves the end user, who in our case might be a *biomedical researcher*. RDF graphs from different data sources are displayed. The steps in the process that lead to the final integrated result (*Figure X-5*) are enumerated below.

1. RDF graphs are displayed in an intuitive and understandable manner to the end user in a graphical user interface.
2. The end user previews them and specifies a set of rules for linking nodes across different RDF models. An example of linking rule could be something as simple as matching IDs of nodes in the various graphs.
3. Merged RDF graphs that are generated based on these rules are displayed to the user, who may then decide to activate or de-activate some of the rules displayed.
4. New edges may be added to the merged graph and be added back to the system. For instance, one may choose to add an edge *promising_candidate* between the nodes corresponding to the ligand 5-Hydroxy Tryptamine and Parkinson 's disease.

4.3 Approaches for data integration

As discussed in the beginning of this section, there are two primary approaches to implement the data integration solutions presented above:

- **Centralized approach:** In this approach the data sets are extracted from various data sources converted into the RDF representation and stored in a single RDF data store. The advantage of this approach is that it is likely to be very efficient and can be simply and quickly implemented.

The disadvantage of this approach is that there are a huge number of biomedical data repositories on the web and it is infeasible to load them into one big data store. Furthermore, as the data and structures in the underlying data sources change, the centralized data store is likely to become out of date and will need to be periodically refreshed and restructured.

- **Federated approach:** In this approach, the data sets sit in their native locations. An RDF wrapper is responsible for converting SPARQL queries into the native query language of the database, e.g., SQL. Results returned, e.g., in the tabular form are mapped into a RDF graph representation. The advantage of this approach is that it is more likely to scale to cover the large number of biomedical data repositories on the web. Furthermore and change in the underlying data collections are immediately reflected as, data is retrieved only in response to a query and is not pre-stored. The disadvantages of this approach are that the performance can be slow as data is fetched and possibly joined over the network and changes in the organization and structure of the data repository on the web will require the RDF wrapper to be reconfigured or rewritten. A mitigating factor is that changes in a RDF wrappers can be isolated to a local site and can be handled via reconfiguration by the local developer.

5. CONCLUSIONS AND FUTURE WORK

Semantic Web technologies provide an attractive technological and informatics foundation for enabling the Bench to Bedside Vision. Many areas of biomedical research including drug discovery, systems biology, and personalized medicine rely heavily on integrating and interpreting data sets produced by different experimental methods, in different groups, with heterogeneous data formats, and at different levels of granularity. Further more, there is a need for evolving synthetic understandings across disciplines, and across the continuum from basic research to clinical applications, applies to most complex diseases and many health care issues. A useful synthesis must combine not only data, but interpretations. Ontologies play a critical role in integrating data. They are also used to represent interpretations in the form of mappings to the underlying data on one hand and definitions and axioms on the other.

In this chapter, we present a real world use case related to Parkinson ’s disease and discuss with the help of an illustrative example how ontologies and semantic web technologies can be used to enable effective data integration. We first present pragmatic issues and design choices that arise

while designing ontologies. A discussion of real world biological data repositories represented in our approach is presented and steps for implementing data integration solutions are discussed with the help of illustrative examples.

This is part of ongoing work in the framework of the work being performed in the Healthcare and Life Sciences Interest Group chartered by the W3C. The examples presented in this chapter are a part of broad set of use cases that will be implemented as demonstrations of Semantic Web technology for the healthcare and life sciences. We will evaluate different approaches for data integration, viz., centralized vs. federated in this context.

Acknowledgements
A significant portion of this work was performed within the framework of the Health Care and Life Sciences Interest Group of the World Wide Web Consortium. We appreciate the forum and the resources given by this Interest Group. We also acknowledge the feedback provided by Alan Ruttenberg at the HCLSIG Face to Face in Amsterdam. Kei-Hoi Cheung was supported in part by NSF grant DBI-0135442.

6. REFERENCES

1. http://www.translational-medicine.com/about/
2. Davenport T. H. and Prusak L., Information Ecology: Mastering the Information and Knowledge Environment, Oxford University Press, 1997
3. Lesne' et al. 2006. Nature 16;440(7082):352-7
4. http://esw.w3.org/topic/HCLS/ParkinsonUseCase
5. Semantic Web Best Practices and Deployment Working Group, http://www.w3.org/2001/sw/BestPractices/
6. V. Kashyap, A. Borgida, Representing the UMLS Semantic Network in OWL (Or "What's in a Semantic Web Link?"), Proceedings of the Second International Semantic Web Conference (ISWC 2003), October 2003.
7. SNOMED, http://www.snomed.org
8. International Classification of Diseases (ICD), http://www.who.int/classifications/icd/en/
9. The Gene Ontology, http://www.geneontology.org/
10. Bowden DM, Martin RF (1995) NeuroNames brain hierarchy. *Neuroimage* 2:63–83.
11. Enzyme Nomenclature, http://www.chem.qmul.ac.uk/iubmb/enzyme/

12. Marenco L, Tosches N, Crasto C, Shepherd G, Miller P. L., Nadkarni P. M., Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. J Am Med Inform Assoc. 10(5):444-53, 2003.
13. The PDSP KI Database, http://pdsp.med.unc.edu/kidb.php
14. PubChem, http://pubchem.ncbi.nlm.nih.gov/
15. Entrez, http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
16. PubMed, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
17. Medical Subject Headings, http://www.nlm.nih.gov/mesh/
18. S. Sahoo, O. Bodenreider, K. Zeng and A. P. Sheth, Adapting resources to the Semantic Web: Experience with Entrez Gene, First International Workshop on the Semantic Web for the Healthcare and Life Sciences.
19. Life Sciences Identifier (LSID) Resolution Project, http://lsid.sourceforge.net/
20. Unified Medical Language System, http://umlsinfo.nlm.nih.gov/
21. Cardoso, J., Sheth, Amit (Eds.), "Semantic Web Services, Processes and Applications", 2006, Springer, Hardcover, ISBN: 0-38730239-5, 2006,
22. Matteo Cristani, Roberta Cuel: A Survey on Ontology Creation Methodologies. Int. J. Semantic Web Inf. Syst. 1(2): 49-69 (2005).
23. V. Kashyap, Design and Creation of Ontologies for Environmental Information Retrieval, In the 12th International Conference on Knowledge Acqusition, Modeling and Management, Banff, Canada, 1999.

7. QUESTIONS FOR DISCUSSION

Beginner:

1. Why is Google, Yahoo! or MSN search not good enough for searching biological data?

2. There are various Web 2.0 approaches based on collaborative annotations of various web resources such as in Flickr. Explain why or why not these approaches are able to capture semantics.

Intermediate:

1. There have been various approaches to using Taxonomies on the Web, for example, Yahoo! and the International Classification of Disease, Medical Subject Headings (MeSH), etc. Explain how these classifications can help in the data integration process.

2. Explain if the classifications identified above are enough to support data integration or more sophisticated ontologies are required.

Advanced:

1. Explain why you think there is value in creating an ontology? Shouldn't the ability to link RDF graphs via URIs be enough to achieve data integration?

2. Explain why you think a Semantic-Web/RDF-based approach is better than current relational database and web based solutions? If not, why not?

3. Compare and contrast the following: classifications and taxonomies, database schemas, ontologies; with respect to the knowledge expressed and their utility in the context of data integration.

8. ADDITIONAL SUGGESTED READING

- Baker, Christopher J.O. and Cheung, Kei-Hoi (Editors); Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Springer 2007, 450 pp; This book is a nice collection of articles illustrating the application of semantic web and ontology-based techniques in the domain of biomedical informatics
- Alan Ruttenberg, Tim Clark, William Bug, Matthias Samwald, Olivier Bodenreider, Helen Chen, Donald Doherty, Kerstin Forsberg, Yong Gao, Vipul Kashyap, June Kinoshita, Joanne Luciano, M. Scott Marshall, Chimezie Ogbuji, Jonathan Rees, Susie Stephens, Gwen Wong, Elizabeth Wu, Davide Zaccagnini, Tonya Hongsermeier, Eric Neumann, Ivan Herman and Kei-Hoi Cheung, Advancing Translational Research with the Semantic Web, BMC Bioinformatics, Vol. 8, suppl.2, 2007. This paper is a nice overview paper of various semantic web technologies such as RDF, OWL and Rules to important informatics problems in the area of Translational Medicine.