SEMANTIC WEB TECHNOLOGY AND PHYLOINFORMATICS

by

MARYAM PANAHIAZAR

(Under the direction of Professor. Jim Leebens−Mack)

ABSTRACT

Phyloinformatics is an interdisciplinary study, combining *Phylogenetics* and *Informatics*. It involves gathering, storing, computing, and reusing the product of *phylogenetic analyses*. One of the major challenges in the domain of *phylogenetics* is the accessibility of published phylogenetic trees and the data that are used to estimate these trees. With the help of Semantic Web technology, we can make *phylogenetic* resources more understandable to the web agents and facilitate the reusability of accessible phylogentic trees and associated metadata. This thesis presents an ontology-driven, semantic problem-solving environment for *phylogenetic* analysis, including four research efforts to achieve this goal. *PhylOnt*, as an ontology for *phylogenetic* analysis. *PhylAnt-D* and *PhylAnt-X*, to annotate *phylogenetic* documents and *NeXML* files. Finally, *MUDDIS* as a MUlti-Dimentional semantic integrative approach for knowledge DIScovery to reconstruct the gene tree based on different gene similarities. The outcome of this research is to advanced *phyloinformatics* to reuse data sources in *phylogenetic* analysis with the help of Semantic Web technologies.

INDEX WORDS: Semantic Web technology, Phyloinformatics, Ontology, Semantic annotation, Data integration, knowledge discovery

SEMANTIC WEB TECHNOLOGY AND PHYLOINFORMATICS

by

MARYAM PANAHIAZAR

M.Sc., Computer Engineering, 2006

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

IN BIOINFORMATICS

ATHENS, GEORGIA

2012

SEMANTIC WEB TECHNOLOGY AND PHYLOINFORMATICS

by

MARYAM PANAHIAZAR

Approved:

Major Professor: Jim Leebens−Mack

Committee: Amit P. Sheth
John A. Miller
Hamid R. Arabnia

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

Semantic Web Technology and Phyloinformatics

Maryam Panahiazar

December, 2012

Contents

List of Figures

List of Tables

Acknowledgments

Over the past five years I have received support, education, and encouragement from a number of individuals. First and foremost, I would like to thank Professor Jim Leebens-Mack and Professor Amit Sheth for their support. They taught me how to follow my interest, how to do interdisciplinary research and be successful.

I will be eternally grateful to Dr. Hamid Arabnia for his great support. Dr. John Miller for his great help during my dissertation. I would like to thank Dr. Olivier Bodenreider and Dr. Bastien Rance of the National Library of Medicine for their assistance and support during my internship as well as the many hours discussing the ups and downs for my research. Thanks to Dr. Jeffery Dean for helping me through the graduation process. Thanks to Dr. Farough Dowlatshahi for his wonderful help at the beginning of my PhD. He taught me to be in right place and do not give up.

I would also like to thank my colleagues and dear friends throughout the years, in particular Dr. Reza Farivar, who always encouraged me during graduate school to start my M.Sc. in Sharif University of Technology. He has always been supportive, taught me the alphabet of research and always encouraged me to be a good researcher and person. He helped me go through the hard years of graduate school.

I would also like to express my gratitude to Dr. Ajith Ranabahu, Kno.e.sis Center, for detailing the finer points of semantic annotation. Dr. Payam Barnaghi spent hours and hours helping me in my research.

I would also like to thank my colleagues from Kno.e.sis, Dr. Christopher Thomas for his brilliant suggestions and help during writing my dissertation. Tonya Davis, Pramod Anantharam, Hemant Purohit, Vahid Taslimi and Hima Yalamanchi. I thank them for being great colleagues in the lab during my research there and making the lab such a nice and

productive place. Finally I thank all the teachers, professors, mentors and colleagues I have had throughout the years.

Special thanks to my best friends and families.

I dedicate my thesis to the best educator Prof. Amit Sheth and my parents for their support during my education.

Chapter 1

Introduction and Overview of the Thesis

Phylogenetic analyses can resolve historical relationships among genes or organisms based on evolutionary similarities and differences. Understanding such relationships can elucidate a wide range of biological phenomena including the role of adaptation as a driver of diversification, the importance of a gene and genome duplications in the evolution of gene function, or the evolutionary consequences of the biogeographic shifts. Since Darwin in 1859 and Haeckel in 1866 published their iconic tree figures around 150 years ago, *phylogenies* have provided a historical framework to interpret the evolution of form and function (Leebens-Mack et al., 2006). *Phylogenies* can address issues ranging from the prediction of genes and protein functions to organismal relationships (Barker and Pagel, 2005; Gaudet et al., 2011). It can be used to transfer knowledge from the genes having known functions to the genes with unknown functions. Finally, *phylogenies* provides the unifying context across the life sciences for investigating diversification of biological forms and functions from genotype to phenotype. However, the increased interest in using and reusing *phylogenies* has exposed major limitations in the accessibility of published *phylogenetic* trees and the data used to estimate these trees. Most of published *phylogenetic* trees can only be found in graphical format embedded in printed or electronic versions of the research publications. This greatly limits the ability of biologists to reuse gene and species trees in meta-analyses with other data sources. The published *phylogenetic* trees are typically inaccessible for semantic integration, also underlying data and methods of analysis are often not adequately described.

In the process of inferring *Phylogenic* trees, there are some methodological concerns in regards to finding and selecting the best methods, models, protocols or data sets. Another concern is how to access the results of the *Phylogenic* studies and reuse them in a similar or new study. One of the challenges is the variety of resources. Experimental data, raw data for computation and analysis, results of the studies, and provenance information for each study are different data sources in *Phylogenic* studies. Each of these data sources has different formats:

1. Unstructured data such as text and image.

2. Semi-structured data such as table, key delimited record.

3. Structured data such as database, *XML* data.

4. Published document data such as technical report, scientific literate, academic article, and manuscript.

As a result, this diversity of data and application of analysis poses informatics challenges such as a) integrating diverse resources, b) finding ways to reuse the published data, and c) generating federated queries given different data sources to answer research questions. All of these issues need the help of informatics in *phylogenetic*, which is called *Phyloinformatics*. In summary *Phyloinformatics* is an interdisciplinary study involving in biology and informatics, which can be useful for storing, organizing, accessing, computing, and finally reusing information in *Phylogenic* studies.

The current state of the field of *phylogeny* study, like other sciences, is that most of the resources can now be accessed on the *Word Wide Web*. Increasingly all the informatics components are public on the Web and follow the open access principles. These include data, publications, sources of the applications, web services, and documents along with the provenance information about them. However, the process of data integration including different sources is a time consuming process for humans. Therefore, finding a way to

make these resources more accessible for computers has became a major challenge in the field of *Phyloinformatics*.

Thus, my research objective is to create a foundation to make the *phylogenetic* resources more understandable for machines, by using Semantic Web technology in *Phyloinformatics*. The specific objective of my research is to develop and deploy a novel, ontology-driven, semantic problem-solving solution for *phylogenetic* analyses and downstream use of *phylogenetic* trees. This leads to development of an integrated platform in *phylogenetic-based* comparative analyses and data integration. For this purpose, it is required to handle, analyze and interpret data in a different and more formalized way. The Semantic Web approach enhances data exchange and integration by providing standardized formats such as RDF(Resource Description Framework) and OWL(Web Ontology Language), to achieve a formalized computational environment, and the solution presented in this thesis is based on them.

The variety of analysis methods and data types typically employed in *phylogenetic* analyses can pose challenges for techniques based on Semantic Web technologies due to significant representational and computational complexity. One of the fundamental foundations for this purpose is to make a common vocabulary. Toward this concern, I have made a network of concepts and defined them in an ontology, which is the formal representation of the set of concepts with the relationships between them. Then, I tagged different resources in the web with the shared concepts, which has defined in the ontology with an uniquely identified URL.

For this purpose, I designed and developed *PhylOnt* ontology for *phylogenetic* analyses, which establishes a foundation for semantics-based workflows including meta-analyses of *phylogenetic* data and trees. *PhylOnt* is an extensible ontology, which describes the methods employed to estimate trees given a data matrix, models and programs used for *phylogenetic* analysis and descriptions of *phylogenetic* trees as well as provenance information for *phylogenetic* resources.

The common vocabulary included in *PhylOnt* will facilitate the integration of heterogeneous data types derived from both structured and unstructured data sources. To illustrate the utility of *PhylOnt*, I annotated scientific literature to support semantic search, and also annotated NeXML formatted files as well. Vos et al. (2012) proposed *NeXML* as an exchange standard for representing *Phyloinformatics* data which is inspired by the commonly used *NEXUS* format (Hladish et al., 2007), but is more robust and easier to process. The annotation platform is designed in three steps; annotation, indexing and searching for annotated data.

I evaluated my ontology to guarantee that what has been built can in fact meet the application requirements and that the elements of the ontology covered the concepts in different resources. I used an annotation-based approach and a metric-based approach to validate quality and quantity of *PhylOnt*. The results of the relationship richness show that more than half of the connections between classes are rich relationships. For the quality validation, I annotated a set of papers which are selected by the field experts. Since *PhylOnt* is specifically developed for *phylogenetic* operations, methods, models and programs, the results show that more than half of the concepts in the selected set of papers were annotated correctly using *PhylOnt*.

Finally, I designed and implemented a semantic integrative approach to comparing genes for knowledge discovery based on multiple gene annotations. The motivation of creating this part of research is that, *Phylogenies* themselves are intrinsically interesting, but their real utility to scientist comes when they integrate *phylogenetic* data with other biological and biomedical data. The driving principle in using a multi-dimensional approach is to create effective domain specific knowledge discovery, based on gene annotation with the use of scientific and provenance information from different resources. The novelty of this part of my research is to make the queries through heterogeneous data sources and create a data collection for similarity calculations and compare that with gene trees as a use case. Semantic Similarity is calculated on different levels of granularity. Data from

literature, open public databases such as Online Mendelian Inheritance in Man (OMIM, 1960), and genes annotations information from National Center for Biotechnology Information (NCBI, 2005) are used as individual resources for different features of the gene. Each additional feature increases the value of the knowledge that can be explained within individual resources. The major motivation for this part of research is to make a foundation for scientists to study and search for different genes within or across diverse species form different perspective.

In chapter 2, I present the *PhylOnt* ontology as a foundation for using semantic technology in *Phyloinformatics*. Chapter 3, describes the use of the ontology for annotating documents in *Phyloinformatics*. Chapter 4, explains the annotation of *NeXML* files, which is an *XML* fromat for *phylogenetic* study. Chapter 5, is about the *MUDDIS* system, a multidimensional semantic integrative approach for knowledge discovery. In Chapter 6, I conclude the work and discuss possible future work such as *SemPhyl* as a platform of integrating, querying and visualizing of *phylogenetic* related sources. The relationships between chapters are showed in Figure 1.1. Although Chapter 2 is based on my published paper (Panahiazar et al., 2012b), it still differs from the published work and it is being submitted to BMC Medical Genomics Special Issue, 2012. A part of chapter 3 is published in IEEE ICSC 2011(Ranabahu et al., 2011a), iEvoBio 2011 (Panahiazar et al., 2011), W3C Workshop on Data and Services Integration (Ranabahu et al., 2011b), and AMIA Annual Symposium (Panahiazar et al., 2012c). The research underlying Chapter 5 was conducted during my internship in summer 2012 at NIH and it is in the process of submitting as a journal manuscript. A part of chapter 6 is introduced in the Translational Medicine Conference AMIA 2012 (Panahiazar et al., 2012a).

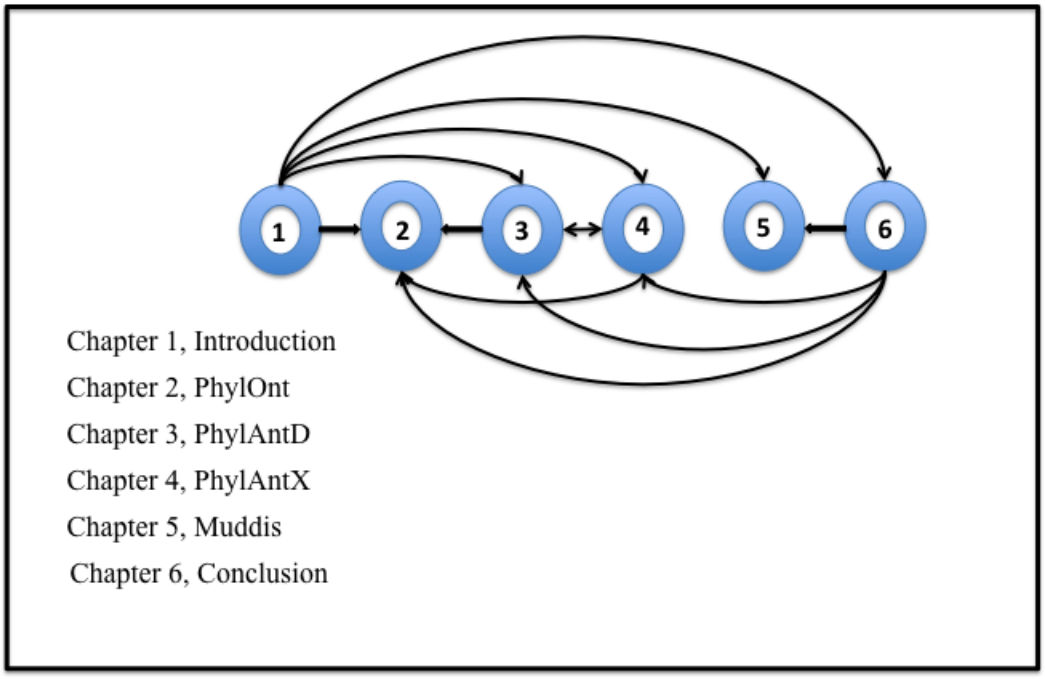

Figure 1.1: Relationships Between Chapters

Chapter 2

PhylOnt, A Domain-Specific Ontology for Phylogenetic Analysis

In this chapter, I discuss the development of *PhylOnt* - an ontology for *Phylogenetic* analyses. *PhylOnt* is an extensible ontology, that describes the methods employed to estimate trees given a data matrix, models and programs used for *phylogenetic* analysis and descriptions of *phylogenetic* trees including branch-length information and support values. It also describes the provenance information for *phylogenetics* analysis data such as information about publications and studies related to *phylogenetic* analyses.

2.1 Introduction

The specific objective of this chapter is to develop and deploy an ontology for a novel ontology-driven semantic problem solving approach in *phylogenetic* analysis and downstream use of *phylogenetic* trees. This is a foundation to allow an integrated platform in *phylogenetically* based comparative analysis and data integration. The variety of methods of analysis and data types typically employed in *phylogenetic* analysis can pose challenges for semantic reasoning due to significant representational and computational complexity. These challenges could be ameliorated with the development of an ontology that is designed to capture and organize the variety of concepts used to describe *phylogenetic* data, methods of analysis and the results of *phylogenetic* analyses.The vocabularies included in

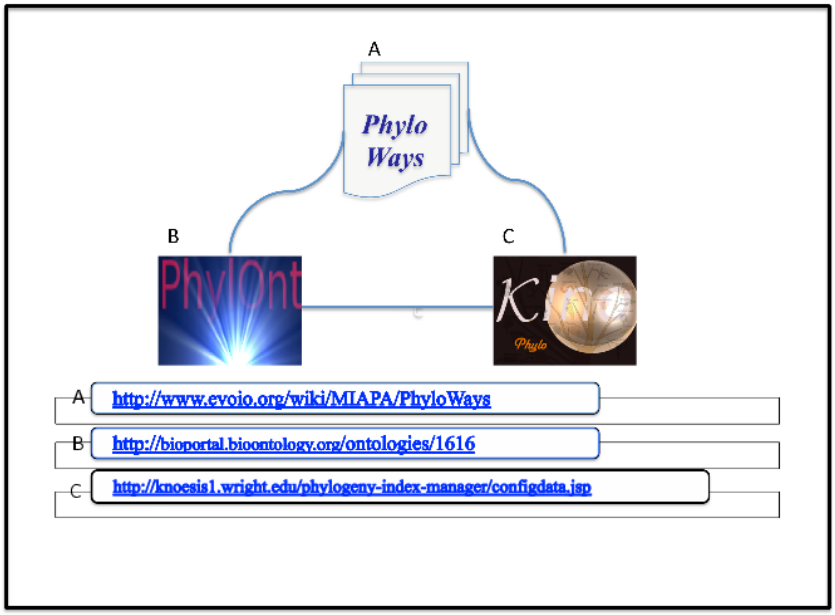

Figure 2.1: Major Components in Design and Implementation of PhylOnt

the *PhylOnt* ontology facilitate the integration of heterogeneous data types derived from both structured and unstructured sources. To illustrate the utility of *PhylOnt*, I annotated scientific literature and files to support semantic search. The semantic annotations can subsequently support workflows that require the exchange and integration of heterogeneous *phylogenetic* information. Figure2.1. shown the three major components in this chapter with the link for each of them on the web. There are different kind of Ontologies in life science domain. Stevens et al. (2000) divided ontologies based on their usage into three types:

1. Domain-oriented, which could be either domain-specific ontologies, such as PEO-Parasite Experiment ontology (Parikh et al., 2012) or domain generalisations such as Gene Ontology (Botstein et al., 2000).

2. Task-oriented, which could be task specific. *EDAM* is one of the example for these kind of ontologies (Lamprecht et al., 2010).

3. Generic, it comprises ontologies to capture common high level concepts, known as upper ontologies such as OBI-ontology for Biomedical Investigations (Smith et al., 2007).

In this research, I consider *PhylOnt* as both domain and task oriented ontology. It is domain oriented, because it contains concepts specifically defined for *phylogenetic* studies. It is task oriented, because it is defined for the operations in *phylogenetic* studies, and it is used for annotation of data sources in *phylogenetic* analyses (for example, it is used to annotate documents and *NeXML* files). *PhylOnt* describes the methods employed with estimate trees given a data matrix, models, programs and provenance data associated with *phylogenetic* analyses. *PhylOnt* also supports the Minimum Information About *phylogenetic* Analyses (*MIAPA*) specification by providing a formal vocabulary for it (Leebens-Mack et al., 2006). *PhylOnt* has been publicly shared through the *BioPortal* (Noy et al., 2009) at the National Center for Biomedical Ontologies (*NCBO*), which is a web based portal universally accessible over the Internet. Thus, the contributions are the following:

1. I described the *PhylOnt* ontology, an extensible ontology targeted towards data integration in *Phyloinformatics*.

2. I described the systematic process taken in developing *PhylOnt*.

3. I provided a comprehensive use case of using *PhylOnt* in annotation. I used a subset of our *Kino* annotation Tools (Ranabahu et al., 2011a; Panahiazar et al., 2011), which enables annotation and faceted search over the annotated publications.

The subsequent sections are organized as follows: Section 2.2 reviews the background and related works in *phylogenetic*. Section 2.3 presents the challenges and opportunities in this field. Section 2.4 explains developing a data set and foundation for ontology development. Section 2.5 describes the ontology development process. Section 2.6 describes the annotation use case. Section 2.7 presents the evaluation and finally section 2.8 includes the conclusions and discussion.

2.2 Background and Related Works

The rapidly increasing number of published genes and species trees creates significant opportunities for addressing a variety of biological questions. Further, this trend is certain to pick up pace with the explosion of data generated by the next generation of sequencing technologies. One of the major challenges in this space, data integration, has been successfully addressed by using ontologies. Ontologies are being used as the core knowledge component in a number of sophisticated, integrated platforms for data analysis and integration (Gaudet et al., 2011; Cruz and Xiao, 2005).

2.2.1 Background

Leebens-Mack et al. (2006) defined the steps of a workflow for *Phylogenic* studies as follows:

1. Formulation of hypotheses and questions.
2. Identifying steps for a gene or taxon sampling scheme for the questions.
3. Data collection, in both scientific and informatics contexts.
4. Constructing the data matrix.
5. Estimating trees with support values.
6. Publishing the results.

As shown above, *phylogenetic* workflows are more complicated than many other types of studies that have well-developed ontologies such as Gene Ontology (Van Auken et al., 2009). Because of this complexity, development of an ontology to support *phylogenetic* studies is more challenging. When creation of workflow is important, this kind of complexity can be even more problematic. Storing data items such as documents, publications, underlying data and workflow in structured, exchangeable and easily retrievable formats

would facilitate interoperability among various researchers. Such practices would allow researchers to access, explore and reuse the products of *phylogenetic* studies including innovative workflow. With these considerations in mind, domain scientists with an interest in archiving and reuse of *phylogenetic* data have outlined the requirements of a reporting standard, which is called *Minimum Information About Phylogenetic Analyses (MIAPA)* (Leebens-Mack et al., 2006). The main objective of the proposed *MIAPA* standard is to enable the interpretation of *phylogenetic* data by multiple researchers. The need for such a reporting standard is clear, but specification of the standard has been hampered by the absence of controlled vocabularies to describe *phylogenetic* methodologies and workflow with common concepts.

2.2.2 Related Works

There are two ontologies that stand out in this domain of study. Both of these include concepts related to *phylogenetic* analysis but they have some limitations which I explain here.

2.2.2.1 Comparative Data Analysis Ontology

CDAO (Prosdocimi et al., 2009; Chisham et al., 2011), is an ontology for comparative data analysis that provides a formal semantic descriptions of data and transformations commonly found in the domain of *phylogenetic* analysis. This ontology was developed as a part of *EvoInfo* group supported by the National Evolutionary Synthesis Center. The focus of ontology is to cover concepts commonly used in evolutionary analyses including *phylogenetic*.

The development strategy in *CDAO* consider that the resources to create ontology is from usage, such as sequence alignment, to challenging projects such as comparing developmental gene expression patterns across species. At the same time they gathered a list of related artifacts, file formats, database schemas, software interfaces, which have been

proposed in use in the evolutionary analysis domain. The key concepts in the *CDAO* ontology are not about the operations and methods, which are used in *phylogenetic* applications. They are mostly about character-state data and sequences (Chisham et al., 2011).

However, there seems to be a major gap in *CDAO* between what is available and what is needed by a *phylogenetic* researcher for *phylogenetic* analyses. *CDAO* does not cover certain concepts related to *phylogenetic* analysis such as methods, models, and programs, which need to be described for the community to estimate trees in organized way. For example, *CDAO* includes concepts such as node, edge, branch, and network that explain the structure of a *phylogenetic* tree/network, but not the analysis of the *phylogenetic* study. To provide more abstraction and a clean hierarchy in *phylogenetic* analysis, I developed *PhylOnt* ontology to continue and extend work in *phylogenetic* ontologies both for covering context base data and *CDAO* knowledge, knowledge needed for *phylogenetic* tools and applications in *Phyloinformatics*, and provenance information related to *phylogenetic* resources.

2.2.2.2 Embrace Data And Methods

EDAM (Lamprecht et al., 2010) is an ontology developed for general bioinformatics concepts including operations, topics, types and formats as a EMBRACE data and methods ontology. *EDAM* can be used as background knowledge for the composition of Bioinformatics workflow. It was developed in the scope of European Model for Bioinfomatics Research and Community Education as an ontology for describing life science web services. One of the advantages of this ontology in my concern is that in contrast to many well known ontologies such as Gene Ontology or the majority of Open Biomedical Ontologies (OBI), which are context-base and focus on the description of biological content, it provides a vocabulary of terms and relations, which are used to annotate web services for the operations, inputs, and outputs. As discussed in (Lamprecht et al., 2010) *EDAM* is not a single ontology. It is really broad and consist of six separate sub-ontologies: Biolog-

ical entity, Topic, Operation, Data resource, Data, and Data format. The most services that *EDAM* has used to annotate them are about sequencing and a few related to *phylogenetic* analysis.However, there are some mentioned limitation of using the *EDAM* ontology, such as, *EDAM* has used to annotate services in a workflow that "contain services like ReadFile (requiring no input but producing new data that is planted into the workflow) or Write-File (without producing any new analysis result" (Lamprecht et al., 2010). It has used for workflows that contain services that make no progress within the workflow.

In conclusion even though *EDAM* includes *phylogenetic* related concepts,but some of the concrete terms forming the core of *phylogenetic* analysis including methods, models and programs have neither been explicitly defined under the correct hierarchy nor reported in *EDAM*. For instance, the character-based method, *maximum-parsimony* inferred in a *phylogenetic* tree was included in *EDAM* as *parsimony-methods* under a general classification of "*phylogenetic* tree construction". By contrast, *PhylOnt* includes *maximum-parsimony* in a very specific classification of *phylogenetic* methods, well defined with its usage, illustrated with an example, related metadata and relationships with programs, and models.

Also some of the important programs for *phylogenetic* analysis like *PAUP**, *NINJA* are missing in *EDAM*. Having a detailed description of the classes and their applications in tree estimation is the key concept behind developing *PhylOnt* ontology. Another issue with *EDAM* ontology is about the limitation of different data and object properties in this ontology. In *EDAM* the majority of the relationships are type taxonomy such as is-a or is-instance-of. There is no way to define a relationship between methods, models and program which are used in *phylogenetic* studies. I overcame these limitations when implementing the *PhylOnt* ontology. In the following, I show some of the statistics for the number of terms, methods, models, programs, metadata and formats of the current work compare to other ontologies. As is shown in Table 2.1. the number of related term in *EDAM* is less than other ontologies.

Parameters	EDAM	CDAO	PhylOnt
total terms	2658	124	116
phylogenetic related terms	25	124	116
phylogenetic methods	5	4	13
phylogenetic models	1	NA	6
phylogenetic programs	2	NA	50
Provenance	NA	NA	21
phylogenetic data formats	3	NA	5

Table 2.1: Comparison of Ontologies (*EDAM, CDAO, PhylOnt*)

2.3 Challenges and Opportunities

The most significant challenge in *phylogenetic* studies is the variety and complexity of data being used in *phylogenetic* reconstruction. Some of the reasons that challenge the reuse of this data are incomplete and non-tractable provenance data, insufficient method descriptions to reproduce the results and the lack of semantic annotations. The different types of data, used in a typical molecular *phylogenetic* analysis workflow, include (Leebens-Mack et al., 2006):

1. Sample description, including taxonomy, collection locality, *DNA/RNA* preparation.
2. Raw sequence data, sequencing methods, sequence assemblies, assembly method.
3. Alignments and trees including branch lengths (with units) and support values.
4. Alignment programs and their parameter settings.
5. *phylogenetic* estimation programs, models of evolution, methods, search algorithms, support assessments, and relevant parameter settings.

My focus in this research is on the last component, formally characterizing these types of data and identifying the relationships between them to develop an ontology for *phylogenetic* analysis. Developing an ontology and using it to annotate the data and services in workflow can provide a foundation for other semantic technologies, such as concept based searches and comprehensive federated queries over the data sources.

2.4 Data Collection

The first step of ontology development is to understand the domain of study for which it is being developed. This is usually achieved by reviewing and harvesting concepts from exemplary publications and data sets. For this study, I reviewed exemplary papers identified by *phylogenetic* experts. I used *Phyloways* (MIAPA/PhyloWays, 2011), as a list of interpreted *Phyloinformatics* workflows environment as a base repository for adding selected papers and results of analysis. In order to perform data extraction, a standard reporting method and formalized methods to extract and classify data from the papers is required (Panahiazar et al., 2012b). I identified all the information required to repeat the analysis done for each tree presented in the *Phyloways* papers such as programs used in the paper, method, models and provenance information related to each paper. I added these extraction methods for *Phyloways* to provide a description of *Phyloinformatics* data and workflows extracted from publications. These descriptions pave the way for classification of concepts associated with *phylogenetic* data (including provenance information) *phylogenetic* workflows, and the results of *Phylogenic* analyses.

2.4.1 Advantages of Data Collection with PhyloWays

The *PhyloWays* data collection has been used as a foundation for making and evaluating diagrams depicting the relationships among concepts that will ultimately evolve into an extensible ontology for *phylogenetic* analyses. *PhyloWays* also serves as an archive to share comments and descriptions of *phylogenetic* documents (Edstam et al., 2011). Finally, *PhyloWays* includes a set of exemplary publications for annotation and validation of the *PhylOnt* ontology.

2.4.2 Candidate Cases for Analysis

Phylogenetic analyses included in *PhyloWays* were categorized into Protein-based and DNA-based groups. My initial step has focused on analyses of molecular data rather than other data types such as morphology. For each paper I harvested the following information: publication, data type, alignment method, method of tree estimation, models, programs, parameters, provenance data and additional comments. In Section 2.4.3, I provide an example of analysing a publication and writing descriptions for the trees and methods used for *phylogenetic* estimation.

2.4.3 Example Analysis

As an example, I provide the detailed steps taken in analysing the paper from (Soltis et al., 2011). The study mentioned in the paper was first converted to a more structured description using commonly used concepts. Compilation of information in *PhyloWays* provides a foundation to develop workflow diagrams, lists of concepts, and the classification that captures relationships among concepts used in *phylogenetic* analyses. Some of the categories are as follows.

- Goal of the Selected Paper: the main descriptive statement can be the goal of the paper.
- About the Publication: *Pub1 has_authors Soltis DE,Smith SA*.
- About the *phylogenetic* results:*phylogenetic Result1. has_value* *; *phylogenetic result1. has_method* *.
- About the Tree:*Tree. is_inferred_by* *; *Tree. has_substitution_model* *.
- About the Program: *Program. is_used_in* *.

The actual values are not included for brevity. The relevant place holders are marked with * symbols. *phylogenetic* result1 is an instance of *phylogenetic* results. has_method is object

properties and has_value is a data type property. I omit the rest of the details for brevity. Check the *Phyloways* and *PhylOnt* ontology for more details.

2.4.4 Domain and Source

Data Sources in *phylogenetic* studies can be classified into scientific and metadata categories. Scientific data exists as published data, such as literature with text, images, excel files, and other supplemental materials. Scientific data also refers to methods, models, programs and even parameters used in programs. Metadata includes data about whom, when and where the data was created. This information plays a very important role in the reusability of valuable resources. For example, when researchers want to reuse or repeat any kind of experiments, this kind of knowledge not only helps them to find the data, it also allows them to evaluate the data source, and methods of analysis. The availability of metadata can also aid disambiguation between experimental data entities.

2.5 A Systematic Approach for Ontology Development

Based on discussions with domain experts, literature reviews and the data in *PhyloWays*, I created the diagrams to describe methods of sequence analysis, models, and widely used *phylogenetic* programs.

2.5.1 Methods in Phylogenetic Analysis

Phylogenetic methods vary considerably in the concepts upon which they are developed and the way they are used to infer relationships. One of the main methods in *phylogenetic* is the optimal tree under the maximum parsimony criterion, which is the minimal tree after evaluating different trees. This method gives the order not the branch length. It means this method searches all possible trees to find the best tree. The optimal tree under the maximum likelihood criterion is the method of search for the tree with the highest probability or

likelihood. Bayesian inference in *phylogenetic* generates a posterior distribution for a parameter, composed of a *phylogenetic* tree for that parameter and the likelihood of the data, generated by a multiple alignment. All of these methods are character-based. The second main group is distance-based methods. Neighbor joining and *UPGMA* are two different distance-based methods.

Two main validation methods are bootstrap and jack knife resampling. Bootstrap resampling is a method to test how good a dataset fits on an evolutionary model by checking the branch arrangement topology of the tree with a bootstrap value. The basic idea behind jack knife resampling is to re-compute the statistical estimate leaving out one observation at a time from the sample set in *phylogenetic* for validation (Harrison and Langdale, 2006). Usually more than 1 site left out in each replicate. Figure 2.2 demonstrates the data diagram for the most popular methods in *phylogenetic* study.

2.5.2 Models in Phylogenetic Analysis

Model selection is very important and effects most of the stages in *phylogenetic* inference. The rational development of a *phylogenetic* method needs a model of evolution as a starting point. Maximum likelihood, Bayesian Inference and most distance-based methods rely on substitution models, but parsimony simply assumes all types of change are equally possible. Substitution models classified as *DNA* model and protein model at the first level of classification. *JC, HKY, SYM, F81, GTR*, and *K80* are all models of Nucleotide Substitution Models (Posada and Buckley, 2004). Figure 2.3 is a diagram for the most popular substitution models used in *phylogenetic* studies. For more details such as definition, description, and resources for each of theses concepts check the *PhylOnt* ontology (Panahiazar et al., 2012b).

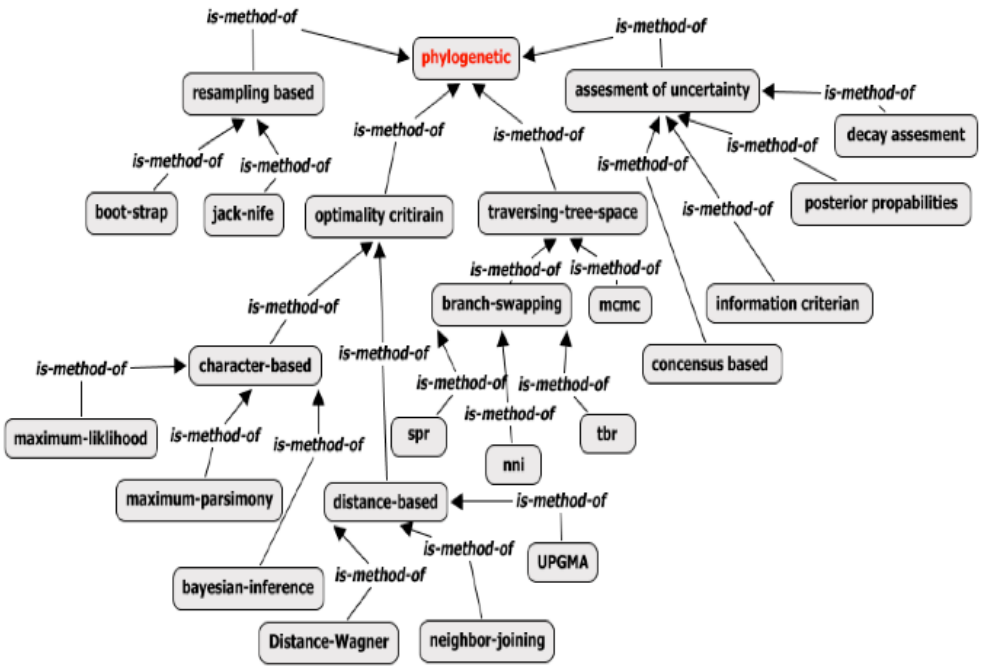

Figure 2.2: Data Diagram for Most Popular Methods in Phylogenetic

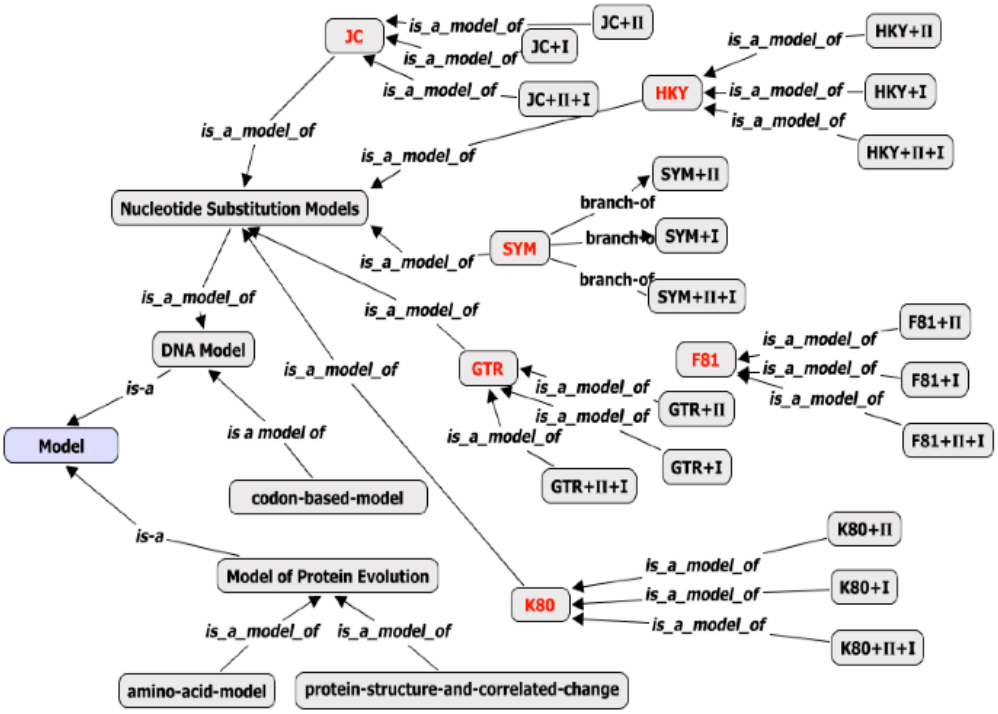

Figure 2.3: Data Diagram for Most Popular Models in Phylogenetic

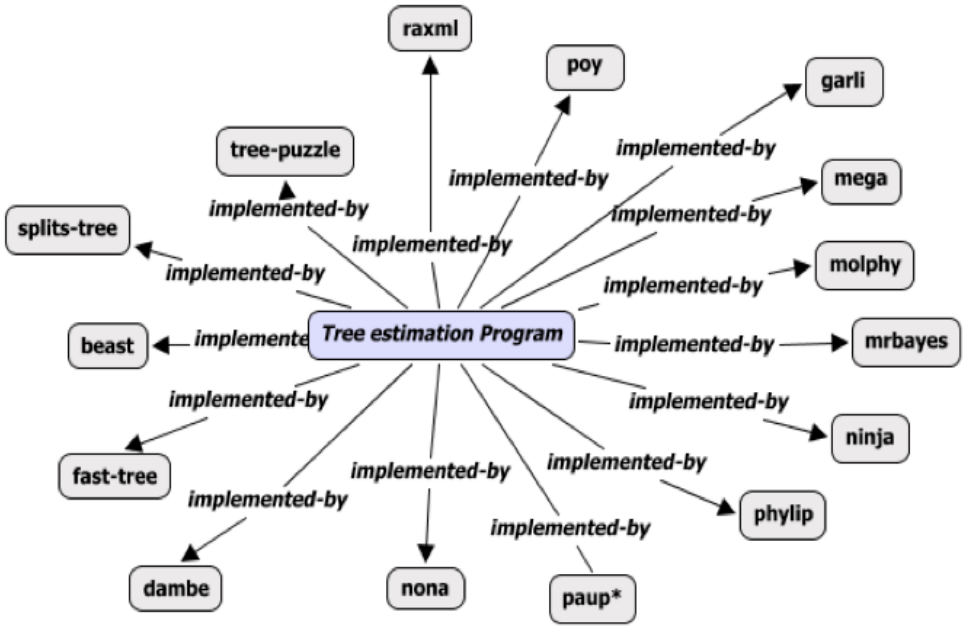

Figure 2.4: Diagram for Most Popular Programs in Phylogenetic

2.5.3 Most Popular Programs in Phylogenetic Analysis

There are approximately 400 *phylogenetic* packages and more than 50 free web servers for such as analysis (Panahiazar et al., 2012b). *PhylOnt* currently includes the most commonly used *phylogenetic* inference programs such as *fast-tree, mrbayes, dambe, nona, garli, paup*, raxml*, and *mega*. Programs can be categorized based on the method they used. For example *paup** can be used to perform most major methods of analysis such as parsimony, and maximum-likelihood. Figure 2.4 shows an example of the programs in *phylogenetic* analysis. For more detail about the programs such as description for each and relation between program, model and method check the *PhylOnt* Ontology (Panahiazar, 2011).

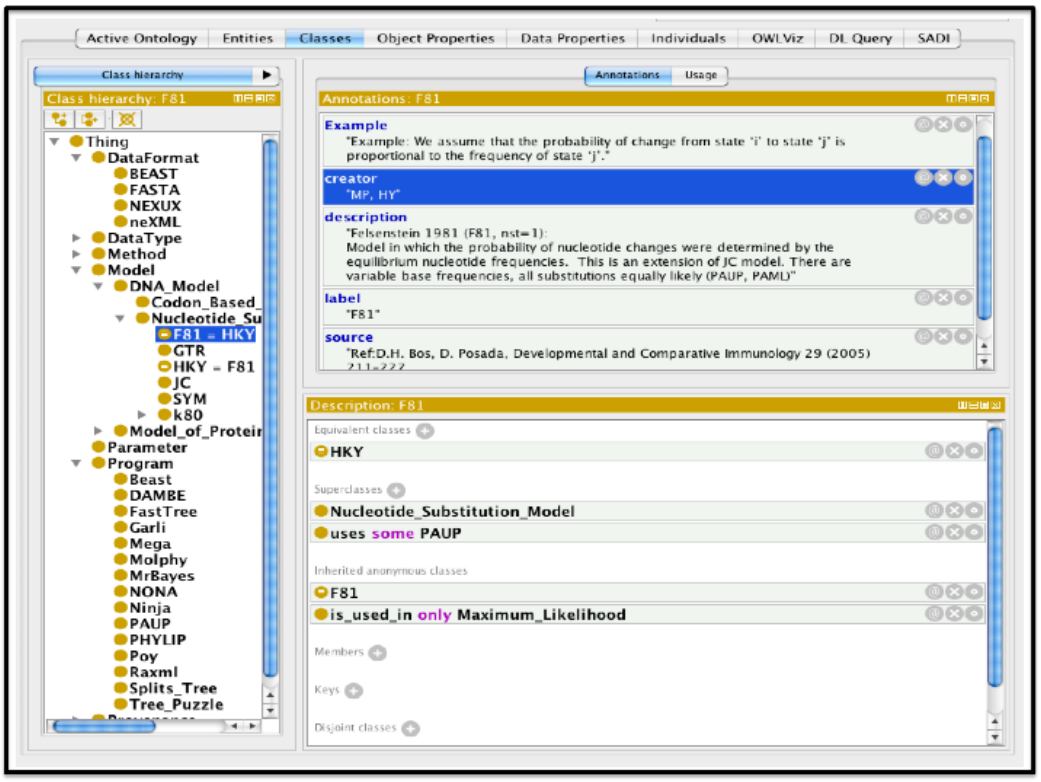

Figure 2.5: PhylOnt Implemented with Protege

2.5.4 Development of the Ontology

An extensible and broadly applicable ontology for description of *phylogenetic* analyses can only be developed though synergistic efforts of both the *phylogenetic* community and computer scientists. For this research, I worked closely with *phylogenetic* researchers and computer scientists to develop and validate the ontology. *PhylOnt* is being developed in *Protege 4.1.0*, which supports the Web Ontology Language (*OWL*). As shown in Figure 2.5. this ontology includes descriptions of classes, definitions, properties, metadata, usage of classes with an example and relations between them. Also object and data properties in this ontology are domain specific for *phylogenetic* such as is-inferred-by, has-substitution-model, etc. For making the ontology publicly available, with the help of *NCBO* researchers, the *PhylOnt* Ontology has already been deployed within *BioPortal* at *NCBO*. The *BioPortal*

is a web based portal designed to enable universal accessibility over the Internet. The deployment of *PhylOnt* in *BioPortal* maximizes its exposure. One of the advantages of *PhylOnt* is a rich relationship defined between each single concept in the ontology. Check the PhylOnt in BioPortal for more information.

2.6 Using Ontology for Annotation - Use Case

A fundamental driving principal for the development of ontologies is their utility for data object annotation and management (Ranabahu et al., 2011a). Therefore, as I developed *PhylOnt*, I used it to annotate publications from the *phylogenetic* literature. Here, annotation refers to embedding labels pointing to ontologies from documents. Using accurate annotations pointing to even a single ontology can improve the quality of lookups in a scientific document management system dramatically. From the perspective of database searches, it is very important to have the ability to link from ontology concepts to concepts in publications. Annotating publications with ontology concepts highlights the utility of an ontology in the targeted field of study, and literature searches (Ranabahu et al., 2011a). One should note, however, that annotation of scientific literature still remains a human-oriented task. My intention is to provide both producers and consumers of phylogenetic trees with a convenient tool to annotate a large volume of documents and retrieve annotated documents for future use. In the process of annotation, I could determine if the concepts encountered in a paper being annotated does not exist in a target ontology, one can search for them in other ontologies or extend *PhylOnt*.

2.6.1 Annotation the *Phylogenetic* Related Papers with *PhylOnt*

Kino for *Phylogenetic*(*Kino-Phylo*) (Panahiazar et al., 2011) is an integrated suite of tools that enables scientists to annotate *phylogenetic* related web-based documents as a branch of *Kino* (Ranabahu et al., 2011a). *Kino-Phylo* can annotate documents by accessing *Phy-*

lOnt and other *NCBO* ontologies. *Kino-Phylo* consists of an *NCBO* integrated front-end that allows the convenient annotation and submission of web documents through a browser plugin, and an annotation aware back-end, capable of providing faceted search capabilities. These annotations have a variety of uses, ranging from extended search capabilities to advanced data mining. In *Kino-Phylo*, Annotated documents are indexed using a faceted indexing and search engine that provides search capabilities for the scientists (Ranabahu et al., 2011a). Thus, *Kino-Phylo* is a comprehensive architecture for annotating and indexing *Phylogenetic* oriented documents that should be of great use for the *phylogenetic* community. More details about *Kino-Phylo* tools is described in chapter 3.

2.7 Evaluation

Ontology evaluation is an important task that is needed in many situations. For example, during the process of building of an ontology, ontology evaluation is important to guarantee that what has been built meets the application requirement. There are different approaches for ontology evaluation, such as evolution-based, metric-based and application-based (Vrandečić and York, 2007). In this study, I used an annotation-based approach and a metric-based approach to validate quality and quantity of the *PhylOnt*.

2.7.1 Metric-Based Approach

These metrics scan through the ontology to gather different types of statistical criteria about the structural knowledge represented in the ontology. In this paper, I followed *OntoQA* framework that is one of the metric based approaches and used schema metrics (Vrandečić and York, 2007). These metrics evaluate ontology design and its potential for rich knowledge representation. In the following, I list metrics with a brief description and then show the results of the evaluation in Table 2.2. In this table, relationship richness reflects the diversity of the types of relations in the ontology. Attribute richness indicates

both quality of ontology design and the amount of information pertaining to instance data. Inheritance richness describes the distribution of information across different levels of the ontologys inheritance tree. The results of the relationship richness show that more than

Metric name	Metric formula[1]	Metric value
Relationship Richness	$RR = \frac{\|P\|}{\|H\|+\|P\|}$	0.54
Attribute Richness	$AR = \frac{\|T\|}{\|C\|}$	0.18
Inheritance Richness	$IR = \frac{\|H\|}{\|C\|}$	0.94

[1] $|H|$: inheritance relationships, $|P|$: non-inheritance relationships, $|C|$: classes, $|T|$: attributes

Table 2.2: Metric-Based Approach to Evaluate the Quantity of PhylOnt

half of the connections between classes are rich relationships compared to all of the possible connections. Inheritance Richness describes my ontology as deep vertical, which indicates that it covers a specific domain in a detailed manner.

Ontology	Precision	Recall	F-measure
PhylOnt	0.64	0.43	0.51
EDAM	0.17	0.013	0.024
CDAO	0.07	0.15	0.095

Table 2.3: Precision, Recall and F-measure Results for Annotation-Based Approach

2.7.2 Annotation-Based Approach

In this approach, I annotated the papers selected by experts using *PhylOnt*. I investigated which concepts are missing in our ontology, in practice by trying to annotate using *PhylOnt*. The rationale is that we could determine the quality of *PhylOnt* by counting the relevant concepts encountered in a paper that are not present in *PhylOnt*, but are present in other relevant ontologies. This approach is used to compute precision, recall, and F-measure (Cross et al., 2011). Suppose that $C_{\{P \cap O\}}$ is the set of concepts from the papers which have been annotated using *PhylOnt*. Then $Precision$ and $Recall$ can be calculated by the

following equations.

$$Precision = \frac{\left|C_{\{P \cap O\}}\right|}{|C_P|} \tag{2.1}$$

$$Recall = \frac{\left|C_{\{P \cap O\}}\right|}{|C_O|} \tag{2.2}$$

C_P and C_O refer to the concepts of the paper and concepts in ontology respectively. The F-measure is the harmonic mean of precision and recall and it is calculated as

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.3}$$

For this experiment, I annotated selected papers using *PhylOnt*, *EDAM* and *CDAO*. As it is shown in Table 2.3 the precision of *PhylOnt* demonstrates that more than half of the concepts in the validation set papers are annotated correctly using *PhylOnt*. Around 43 percent of the all concepts in the paper that should have been annotated are annotated correctly with *PhylOnt*. And this 43 percent make sense, it means close to 50 percent of the concept in validation set papers are covered with the concepts of *PhylOnt*. Also the F-measure of *PhylOnt* is 5 times more than that of *CDAO* and 20 times more than that of *EDAM*.

2.8 Conclusion and Discussion

With the growing importance of using semantic technology in life science, having a well-defined ontology is necessary to make a foundation and facilitate the integrity and accessibility of data and services. To the best of our knowledge and the feedback from *phylogenetic* community (Panahiazar et al., 2011), *PhylOnt* is the first ontology specifically created for *phylogenetic* analysis operations and related metadata. As of Sep 2012, the 6th version of *PhylOnt* has been submitted to *NCBO*. My results show that *PhylOnt* is a rich ontology, compared to other alternatives. Annotating *phylogenetic* documents with ontology is

the foundation for the use of other semantic technologies and it is a preliminary step to semantic search, information retrieval, and heterogeneous data integration that can support *phylogenetic* workflows. *PhylOnt* has been introduced as an important component in my integrated *SemPhyl* platform (Panahiazar et al., 2012a) and has been used for annotation documents and *NeXML* files in *phylogenetic* studies.

Chapter 3

PhylAnt-D, Semantic Annotation of Phylogenetic Documents

In this chapter, I describe the process of describing the publications with *PhylOnt* concepts. This annotation helps to use the existing valuable information in publications for integration and reusability of published data.

3.1 Introduction

A large volume of *phylogenetic* related information is buried in the fast-growing literature and academic articles. One of the requirements of the Semantic Web to understand the web content is to describe web content semantically with ontologies (Jonquet et al., 2009). Through annotation or tagging of data with the ontology concepts, unstructured data becomes standardized and understandable for machine to use. In my research, these annotations contribute to create a *phylogenetic* Semantic Web that facilitates phylogeny scientific discoveries by integrating annotated data from literature. This annotation is possible with the concept of all ontology from *NCBO*. My concern is the annotation with *PhylOnt* ontology in this research. I explain how to annotate terms in publications with the concepts from *PhylOnt*. I have already introduced literature annotation in chapter 2, as a use case of using *PhylOnt* and validation *PhylOnt* Ontology in the comparison of *CDAO* and *EDAM*. In this chapter, I introduce the related work and explain the details of annotation process.

3.2 Background and Related works

Usually semantic annotation needs resources such ontologies to map the concepts from resources to the target data set for annotation. In the process of annotation, extra information is added to resources, which connect a part of the data to corresponding concept in the ontology. Today, much web information still are unstructured. However, they need to be structured to reuse with the software agent in the web. Annotation is one of the solution, but still it is time consuming and not completely automated. In the following, I introduce some of the related work in this field. Then, I discuss my approach as a semi-automatic annotation. One of the limitations of annotation is the variety of different available ontologies. Ontologies change often, and sometimes they have overlap with each other. If users want to annotate the resources manually with ontologies, they have to be always aware of the last version of the ontologies and the overlap between ontolgies. With the semi-automatic approach, I give the opportunity to annotator to look at the all *NCBO* ontologies, which have the term of interest. Then, they can scan through ontologies and make a decision to select one of them, based on their preferences for annotation.

3.2.1 NCBO Annotator, Semantic Annotation of Biomedical Data

The National Center for Biomedical Ontology Annotator is an ontology-based web service for annotation textual data with biomedical ontology concepts (Jonquet et al., 2009). This annotator has used by biomedical community to tag data base with more than 200 ontologies from *UMLS* Metathesaurus and *NCBO* Bioportal. With this annotation, unstructured data sets turn to be understandable for future use (Uren, 2006) with agents in the web. Jonquet et al. (2009) proposed a web service that allows scientists to utilize most of the public biomedical ontologies to annotate their datasets automatically. The Annotator web service is publicly available and can be used by the community for annotation. The annotation process has two steps: in the first step annotation created from the text based on

concepts from the dictionary. This dictionary uses terms and their synonyms from *NCBO* and *UMLS* ontologies. In second step, using the knowledge represented in one or more ontologies, different components expand the annotation set from first step. The annotation results could be saved as *XML*, tab delimited or even *OWL* file. However, one of the issues with this annotator is about full automatically selected concept for annotation and does not give the opportunity for expert to select the concepts and flexibility of selecting and checking the concept in different ontologies to make the right decision about annotation (Jonquet et al., 2009).

3.2.2 TextPresso,Text literature Searches

TextPresso (Müller et al., 2004) is an ontology-based information retrieval and extraction system for biological literature. It is a text processing system, that splits papers into sentences, and sentences into words. Each word or phrase is then labelled using *XML* according to the lexicon of their ontology. *TextPresso* is one of the projects for full text literature searches for specific organism, text classification and mining literature for database curation and make a link between biological entities in *RDF* and on line journal articles to on line databases. In *TextPresso* for each of the selected words and phrases, they manually determined to which of the three categories, cellular components, essay terms, and verbs, it should be assigned (Van Auken et al., 2009), which it is a time-consuming task. Another limitation with *TextPresso* is that it is specifically defined and it used *TextPresso* ontology for annotation. So, it is not appropriate for the annotation of phylogeny data.

3.2.3 Artemis, Annotating Sequence for Microbial Genomes

Carver et al. (2008) provides tools for annotating a sequence on the database. These annotations are prepared via rapid information and knowledge exchange between teams of literature annotators and data curators. Annotations include literature and other database cross references such as; GO terms inferred from the literature and user comments, and

phenotype curations. Indeed having other ontologies as a rich resource for annotation may further improve efficiency of the re-usability and acceptability with other resources. Therefore Artemis has the same issue that other annotators have so far and does not give the opportunity for user to annotate the text with the selected ontology.

3.3 Data Collection for Annotation

PhyloWays is a part of an *Evoio, MIAPA*, collaboration (Panahiazar et al., 2012b) through the process of data collection and standard definition. It has been used as a foundation for making and evaluating diagrams depicting the relationships among concepts that will ultimately evolve into an extensible ontology for *phylogenetic* analyses. *PhyloWays* will also serve as an archive where users can share comments and link to workflow descriptions of *phylogenetic* documents. Finally, *PhyloWays* includes a set of exemplary publications for annotation and validation of the *PhylOnt* ontology. So, for the purpose of annotation, I have selected data from *PhyloWays* repository or other publications selected with exports. More details about *PhyoWays* is explained in chapter 2.

3.4 Kino-Phylo, A Platform for Literature Annotation

Kino for *Phylogenetic*, also known as *Kino-Phylo* (Panahiazar et al., 2011) is a part of *kino* platform (Ranabahu et al., 2011a). It is an integrated suite of tools that enables scientists to annotate *phylogenetic* related web-based documents. *Kino-Phylo* can annotate documents by accessing *PhylOnt* and other *NCBO* ontologies. *Kino-Phylo* consists of an *NCBO* integrated front-end that allows the convenient annotation and submission of web documents through a browser plug in, and an annotation aware back-end, capable of providing faceted search capabilities. These annotations have a variety of uses, ranging from extended search capabilities to advanced data mining. Annotated documents are indexed using a faceted indexing and search engine that provides fine grained search capabilities to the scientists.

Thus, *Kino-Phylo* is a comprehensive architecture for annotating and indexing *phylogenetic* oriented documents that should be of great use for the *phylogenetic* community. This system is designed around a basic workflow consisting of three steps; annotate, index, and search. The remainder of this section, describing the architecture of *Kino*, is reproduced from our *Kino* paper (Ranabahu et al., 2011a).

1. Annotation: In the annotation step, users provide annotations via various tools. The illustrated case is the use of browser plugin, but, it can be through a Web site, when the annotations are added, the augmented document will be submitted to the indexing engine.

2. Indexing: Indexing is performed using Apache *SOLR*. It can be installed as an independent application and exposes multiple interfaces for client programs. SOLR provides the isolation for an index as well as built in faceting support, which can be controlled via a configuration file.

3. Search: The search used driven Web user interface. It presents a typical search engine , and gives the ability to filter the results via the facets. The current UI is based on the *Kino JSON API* , which can be used to integrate any other tool.

3.4.1 Browser Plugin for Phylogenetic Annotation

Figure 3.1. shows the user interface of the annotator plugin. When the user highlights and right clicks in a word for annotation, the browsers context menu includes the annotation as a *phylogenetical* concept menu item. Selecting this menu item brings the annotations window where the highlighted term is searched using the *NCBO RESTful API* and a detailed view of the accessible ontological terms is shown to the user for selecting. The user can search or browse for a concept in any ontology in *NCBO*. Once the annotations are added, users can submit the annotations to a predefined Kino instance, by selecting the publish annotations item from menu. The annotator, modifies the *HTML* source as exemplified in Listing 3.1.

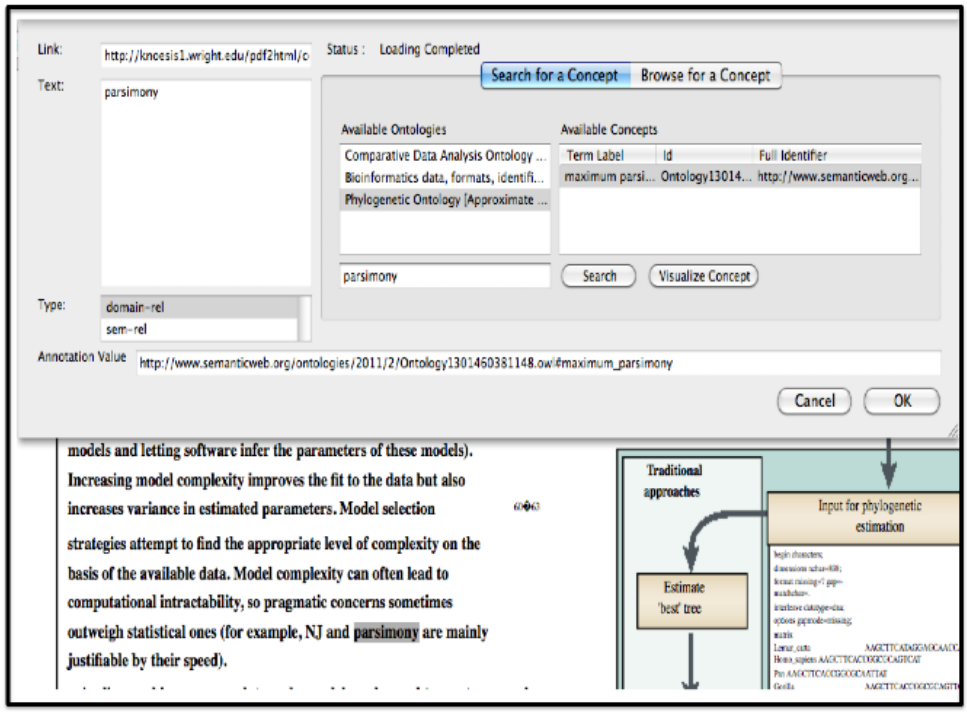

Figure 3.1: Annotation Document with kino-phylo tools

When the text is already contained by a logical grouping element such as a div, the plugin modifies the existing element instead of adding a new one. Title and class attributes are mandatory fields, that are required by the *SAREST* specification (Ranabahu et al., 2011a).

```
<span
    sarest : displayname=``parsimony"
    sarest : conceptid=``maximum_parsimony"
    sarest: ontologyid=``1616"
    title=`` http://www.semanticweb.org/ontologies/
    ...2011/7/Ontology1314368515010.owl#Maximum_Parsimony
    class=``sem-class">parsimony</span>
```

Listing 3.1: HTML Source Annotation added by Browser Plugin

In submitting the documents, the plugin has to send the full serialization of the internal document, in *XML* form, to the indexer. The index manages content of each annotation, the annotated text and the content of the document, therefore the users have the flexibility to search on the annotated concept in annotated documents.

3.4.2 Kino-Phylo Index and Search Manager

The *Kino* index manager is based on the Java *JSP/Servlets* technology and includes two major components, Document Submission *API* and Search *API*. After submitting the document with this *API*. Then the document will be indexed and the response will send to the submitter. As shown in Figure 3.2, Search API includes the facet selection section that helps user to filter the results. For example, if the user searched for the the parsimony, she can find all the documents, which have annotated with that concept so far. The *UI* includes a facet selection section that helps user to filter the results.

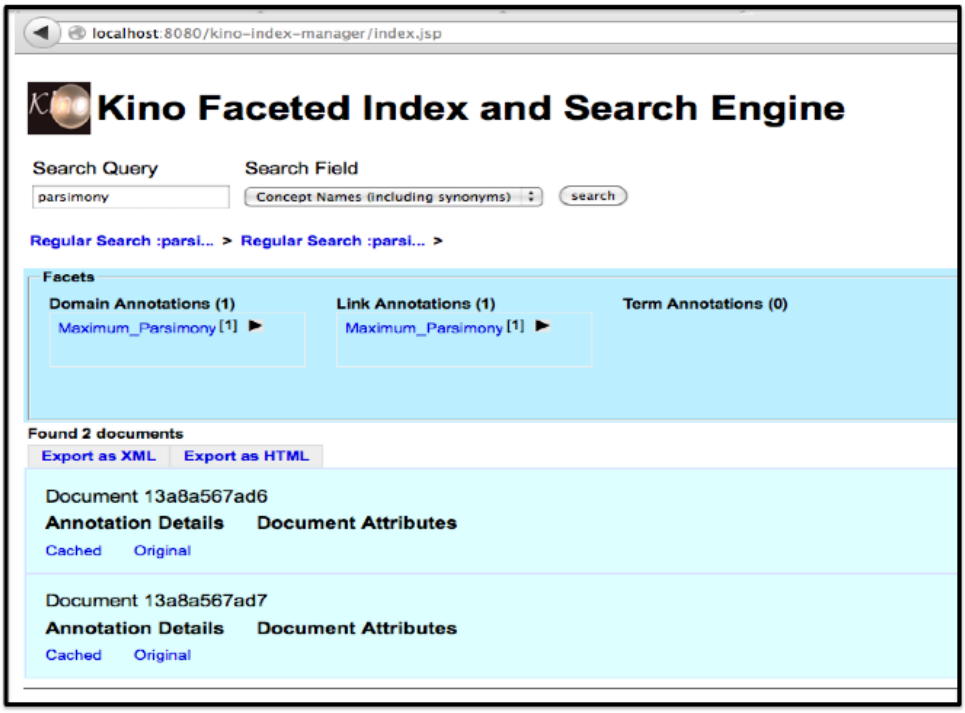

Figure 3.2: User Interface to Search the Annotated Documents

Chapter 4

PhylAnt-X, Semantic Annotation of NeXML files

In this chapter, I describe the process of annotation of the *NeXML* file as a specific *XML* format file for phylogeny study. This annotation is with the concepts from *PhylOnt*. Annotation the *NeXML* file with these concepts will be useful to make the *NeXML* files more understandable with the agent of the web for searching, integrating and re-usability.

4.1 Introduction

Usually semantic annotation needs resources such as ontologies to map the concepts from resources to the target data set for annotation. In previous chapter, I discussed how to annotate the document to add the extra information to our resources which says this part of data is about corresponding concept in ontology. In the following, I discuss the semi-automatic annotation of *NeXML* files. Users do not always know the structure of an ontology 's content or how to use the ontology to do the annotation themselves. In my approach, I overcome this limitation by showing the all possible ontologies for a selected concept for users to evaluate when assigning an annotation. After selecting an element from *NeXML* file for annotation, user have access to ontologies and select one of them to assign an annotation.

4.2 Background, NEXUS and NeXML

Vos et al. (2012) proposed *NeXML* as an exchange standard for representing *phylogenetic* data which is inspired by the commonly used *NEXUS* format (Hladish et al., 2007), but more robust and easier to process. *XML* file plays an essential role in promoting the accessibility and reusing data in the web. Using this technology can simplify and improve robustness in the processing of rich phylogenetic data and help to reuse phylogeny data as well. The purpose of *NeXML* is to leverage *XML* technologies in the development of a data standard that translates *NEXUS* concepts into a syntax that is more easily validated and processed. "*NEXUS* is a file format designed to contain systematic data for use by computer programs. The format is modular, with a file consisting of separate blocks, each containing one particular kind of information, and consisting of standardized commands" (Maddison et al., 1997).

Hladish et al. (2007) proposed that the *NEXUS* file format is modular, placing data, runtime commands, and trees in separate blocks. A node in *phylogeny* tree is shown as *TREES*, character state data shown as *CHARACTERS* and operational taxonomic units*OTUS*. *NEXUS* does not have an explicit means to link data objects with ontology concepts to include citations, or to convey some of the key annotations (study objectives, specimen vouchers, and methods descriptions) (Vos et al., 2012).

There are some issues regarding using the *NEXUS* file, which *NeXML* file overcomes them. For example, there is no formal specification for using the NEXUS file and no way for validation of them. Another issue of using NEXUS file is that their semantics is not defined. Using a *NeXML* file gives the possibility of annotating the file with ontologies. As a conclusion because of all advantages of *NeXML* files compare with *NEXUS* files, the *phylogenetic* community and developers in the *NESCent* working group decided to develop and replace this data exchange based of *XML* format. The design of *NeXML* facilitates the developing the tools, and querying the documents in the web.

TreeBASE is the database and repository that support *NeXML* (Vos et al., 2012). There

are some programming libraries that support *NeXML* file are as follows: *BioPhylo, BioPerl, BioRuby, DendroPy*, and *NeXML* Java Library. By using these libraries in applications or simple scripts, data expressed in *NeXML* could be accessible for reusability. Currently *NeXML* schema annotated with the Character Data Analysis Ontology. In this section, I explain the process of annotation of *NeXML* as the first step. Then I explain how I adapted this annotation process to annotate data with *PhylOnt* and other ontologies, which they are accessible through *NCBO* BioPortal.

4.3 Introduce TreeBase, A Database that Support NeXML

Community resources such as *TreeBASE* (Piel et al., 1997) produce large amounts of metadata that can be expressed using an method. *TreeBASE* is a database for *phylogenetic* knowledge that support *NeXML* files and it is a repository for *phylogenetic* information including the user-submitted the *phylogenetic* tree and data used for generating that tree. Any kind of data in *phylogenetic* from underlying data to the information in the scientific literature, academic articles and thesis can be submitted to *TreeBASE*. *TreeBASE* is produced and governed by the The *Phyloinformatics* Research Foundation, Inc. It generates *NeXML* output as an option accessible via its web interface. I describe *NeXML* which has used by the *TreeBASE* for annotation as an example.

4.4 Annotation Process, Metadata Annotations in NeXML

For the re-usability of *phylogenetic* resources, it is required to annotate *phylogenetic* data with metadata such as provenance information related to publications or methods which have used in specific *phylogenetic* study. Vos et al. (2012) proposed that one of the big advantages of *NeXML* is the ability to encode metadata annotations that are linked to data elements.

4.4.1 Representing the Metadata with Meta Element in NeXML

Annotations are expressed using recursively nested meta elements. The annotations are based on triples and the triples include subject, predicate, object. Before starting the annotation, I extract the triples from *PhyLOnt* with the the help of *Jena* which is shown in Listing 4.1. Knowing the triples is useful during the annotation process. *NeXML* has the facility to annotate the *phylogenetic* data object such as tree, character state matrix and taxa with ontology predicates. One of the important features of *NeXML* is the ability to encode metadata annotations that are linked to data elements. Instead of trying to provide vocabulary for all meta data types in the NeXML, use of defined vocabularies in the ontologies to annotate the NeXML. In the process of annotation there are 3 kind of object values (Vos et al., 2012):

1. Literal object value, meta elements of this type are of the subclass *nex:LiteralMeta*. In this case the object value is enclosed inside the meta element. The predicate is defined with property attribute and data type is specified by the data type attribute such as *xsd:string* or *rdf:Literal*. In this case *href* attribute is used to specify the location of the the object. The predicate of this triple is specify using the *rel* attribute.

2. Remote resource as an object value, meta elements of this type are of the subclass *nex:ResourceMeta*. As the same as previous one, the predicate of this triples is specify using the *rel* attribute. (Vos et al., 2012) confirmed that "While *NeXML* supports the representation of metadata, generalized tools and libraries for this task do not exist at present". In the rest of this chapter, when I annotate *NeXML* file for metadata with *PhylOnt* Ontology, I explain the *Kino-Phylo* tools to annotate *NexMl* File as well.

3. A nested annotation as an object value, the predicate is specified using the *rel* attribute. In this case enclosing meta element has to be transformed to an anonymous RDF node. So, this can be identified with assigning the subject with the about at-

tribute.

Listing 4.2 is an example annotation from (Vos et al., 2012). This example shown the otu element as a single container with the single *otu* element. This element has submitted to the database with the lable “Zenodorus cf. orbiculatus”. Then this label has matched to *uBio* web service. The next step is to find the match from *uBio* to the concept from *NeXML* file. The matched concept called “Zenodorus orbiculatus” with the namebank identifier “3546132” which is “close match” to “Zenodorus cf. orbiculatus”. Also it has matched to *NCBI* taxonomy for “Zenodorus cf. orbiculatus d008” with taxon identifier “39321”. The original *OTUS* which is the starting point of this annotation is defined under the context of *TreeBASE* study called “S1787”. All predicates are define with rel and all the subjects for selected object to annotate are define by $href$ in the nested metadata section.

```
public class Main {
  /**
   * comment: get the triples from PhyloNT with Jena
   * Author: mp
   */
  public static void main(String[] args) throws IOException {
    //create ontology Model
    Model m = ModelFactory.createOntologyModel(OntModelSpec.
        OWL_DL_MEM_RULE_INF);
    //red the PhylOnt
    m.read("file:///Users/mary/Desktop/mary/UGA/project/PhylOntRDF.rdf",
        "RDF/XML");
    StmtIterator itr = m.listStatements();
    Statement stmt;
    while (itr.hasNext()) {
      stmt = itr.nextStatement();
      Resource subject = stmt.getSubject();
      Property predicate = stmt.getPredicate();// get the predicate
      RDFNode object = stmt.getObject();// get the object
      if(subject.isURIResource()){
        System.out.print("Subject:");
        System.out.println(((Resource) subject).getLocalName().toString
            ());
      } else if(subject.isAnon()){
        stmt.removeReification();
        stmt = itr.nextStatement();
      }
      System.out.print("Predicate:");
      System.out.println(predicate.getLocalName().toString()+"\t");
      if(object.isURIResource()){
        System.out.print("Object:");
        System.out.println(((Resource) object).getLocalName().toString()
            );
      } else if(object.isLiteral()){
        System.out.print("Object:");
        System.out.println(((Literal)object).getString());
      }
        System.out.println(" .");
    }
  }
}
```

Listing 4.1: Extracting triples from PhylOnt With Jena API

```
<nex : nexml
  xmlns :nex=http://www.nexml.org/2009
  xmlns :rdfs=http://www.w3.org/2000/01/rdf-schema#
  version= 0.9
  xml :base=http://purl.org/phylo/treebase/phylows/
  <otus id= otus1 label= combined >
    <otu
      about= #otu3
      id= otu3
      label= Zenodorus cf. orbiculatus >
    <meta
      href= http://purl:uniport.org/taxonomy/393215
      id= meta8
      ref= skos:closeMatch
      xsi:type= nex:ResourceMeta />
    <meta
      href= ...
    </out>
  </oyus>
</nex:nexml>
```

Listing 4.2: Example of NeXML file from (Vos et al., 2012)

The root element of the schema is called *nexml*. It includes two attributes: version attribute and generator attribute. Also the root element usually define a number of *xml* namespace prefixes such as *XML* namespace prefix, and *NeXML* name space prefix. To associate the instance document with the *NeXML* schema, also needs an attribute for specify the schema location, and the namespace it applies to. The root element contains some semantic annotations, some *OTUs* elements, some characters elements, and some trees elements. In a *NeXML* file, the *OTUs* block can be involed using a set of labels which sequences and nodes in the file must refer to it. The *OTUs* element and its contained *otu* elements, it has an id attribute and an optional label which is readable for human and contain semantic annotations.

Trees or networks in a *NeXML* file are nested within a trees tag. A *NeXML* file can contain zero or more trees elements containing one or more *phylogenetic* tree or network inside it. Trees are linked to an *otus* with the compulsory *otus* attribute. Trees must be id tagged and may have an optional label (Vos et al., 2012). As I mentioned before *phylogenetic* trees are defined in *NeXML* with the tree tag, with a compulsory id and an optional label attribute. Nodes and edges are nested within tree. Nodes are defined with the node

NO	NeXML file
1	characters.xml
2	edgelabels.xml
3	metataxa.xml
4	nexml.xml
5	phenoscape.xml
6	sets.xml
7	taxa.xml
8	timetree.xml
9	tolskeletaldump.xml
10	tolweb.xml
11	treebaserecord.xml
12	treesuris.xml
13	trees.xml

Table 4.1: List of the Candidate files for Annotation

tag and they have id as well. Nodes can optionally be linked to an *OTU* with the *OTUs* attribute. Edges are defined with the edge tag. An edge must have a direction, defined by the compulsory source and target attributes.

4.4.2 Annotation the NeXML file with PhylOnt

For annotation the *NeXML* file, I follow the existing standard, which has proposed with (Vos et al., 2012). I annotate the *NeXML* file with adding metadata. First, I explain the annotation process, then will explain the *Kino-Phylo* annotation tools to annotate the *NeXML* file semi-automatically. Table 4.4.2. is the list of *NeXML* file, which I got them from *NeXML* page (all modified in July 2012). I annotated *tree.xml* with adding metadata from *PhylOnt* ontology.

As shown in Listing 4.3. I have annotated *tree.xml* file from the Table 4.4.2. As an example

I annotate the tree as a subject with the predicate called $is_inferred_by$ from *PhylOnt* and the $maximum - likelihood$ as object of this triple, which is defined with *href*. The metadata id also assigned to the element.

```
<nex:nexml version="0.9" xml:base="http://example.org/" xsi:
    schemaLocation="http://www.nexml.org/2009 ../xsd/nexml.xsd">
  <otus id="tax1" label="RootTaxaBlock">
  <trees otus="tax1" id="Trees" label="TreesBlockFromXML">
    <otu id="t1">
    <tree id="tree1" xsi:type="nex:FloatTree" label="tree1">
    <meta
       <!-- href shows the URL for the selected concept in ontology -->
        href=http://www.co-ode.org/ontologies/ont.owl/#tree
        <!-- id for meta can be assigned aytomatically -->
        id="meta23"
        <!--rel is the predicate -->
      rel="is-close-match"
      <!--xsi:type shows the type of annotation which is resource base
          now-->
      xsi:type="nex:Resourcemeta"/>
      href=http://www.co-ode.org/ontologies/ont.owl/maximum-likelihood
      id="meta2"
      rel="is-inferred-by"
      xsi:type="nex:Resourcemeta"/>
    </tree>
    </otu>
    <otu id="t2"/>
    <otu id="t3"/>
    <otu id="t4"/>
    <otu id="t5"/>
  </otus>
</nex:nexml>
```

Listing 4.3: Annotation of tree.xml file as an NeXML file with metadata from PhylOnt ontology

4.4.2.1 Annotation Process with kino-Phylo for NeXML

As the same as *Kino-Phylo* for document annotation, the system is designed in three steps, which includes annotation, indexing and searching. In the annotation process with semi-automatic *Kino tools*, as shown in Figure 4.1 user start the annotation with browser plugin tools with the righ click menu switch to *NeXML* annotation view. The *NeXML* annotation Context Menu in *Kino-Phylo* Browser Plugin Tools, with the right click menu in the selected concept for annotation the context menu (Popup Menu) can be used to switch to *NeXML* annotation View. Then the new window shows the list of *NCBO* ontology, which *PhylOnt* is included as well. User can select the concept from any of those ontologies to annotate selected element in *NeXML* file. The annotation process is somehow is different from document annotation. which I explained in previous chapter. After selecting the term

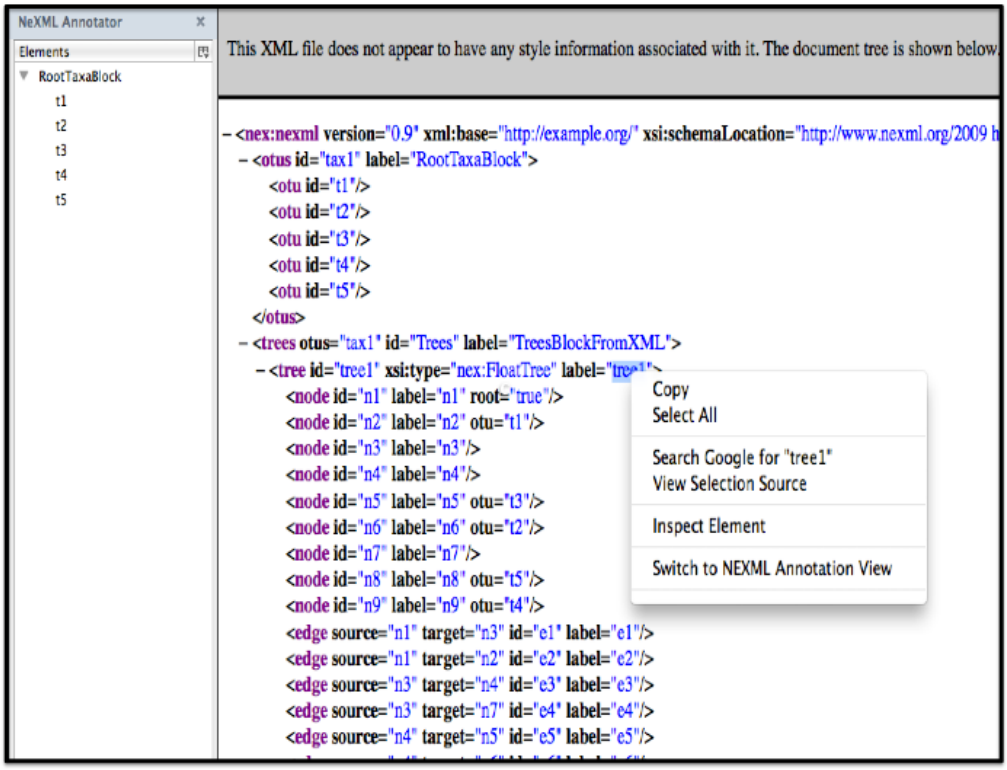

Figure 4.1: The NeXML annottaion Context Menue in Kino-Phylo Browser

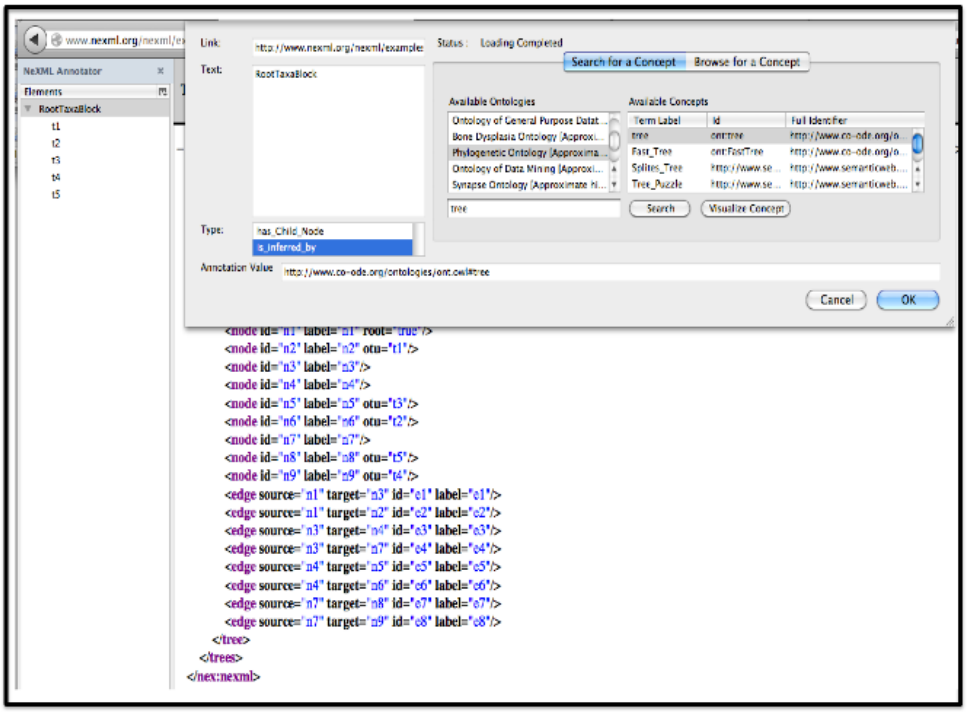

Figure 4.2: Find an element in PhylOnt with Kino-Phylo Tools

for annotation as shown in Figure 4.2 a toolbar shows in the left panel and all extracted element shows in this panel. Therefore, user can select any element and add the metadata to selected element. This metadata will be added with *href, rel, id,* and *xsi:type* as *nex:Resourcemeta*. As shown in Figure 4.3 first the element will find in ontology and annotate that as exact-match or close-match. Then user can annotate the element with desire triples, for example tree $has-substitution-model$ as $necleotide-substitution-model$, which shown in Figure 4.4. Then, the next step will be submit this annotation through the submit menu to be published for future use, which it shown In Figure 4.5 and Figure 4.6 In the second step, as an indexing process which is with *Apache SOLR* and expose multiple client for user. For more detail about this part see our *kino* paper (Ranabahu et al., 2011a) and previous chapter in this document. Finally, as a searching process with web user inter-

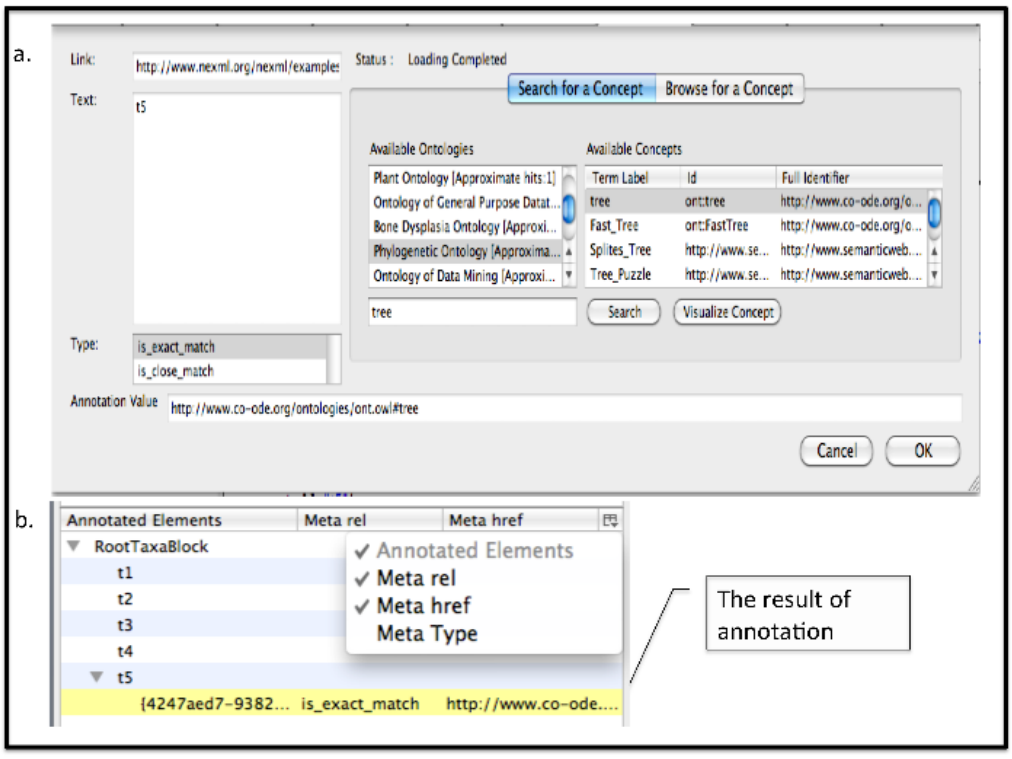

Figure 4.3: Annotation NeXML file with Kino-Phylo, find the element in PhylOnt

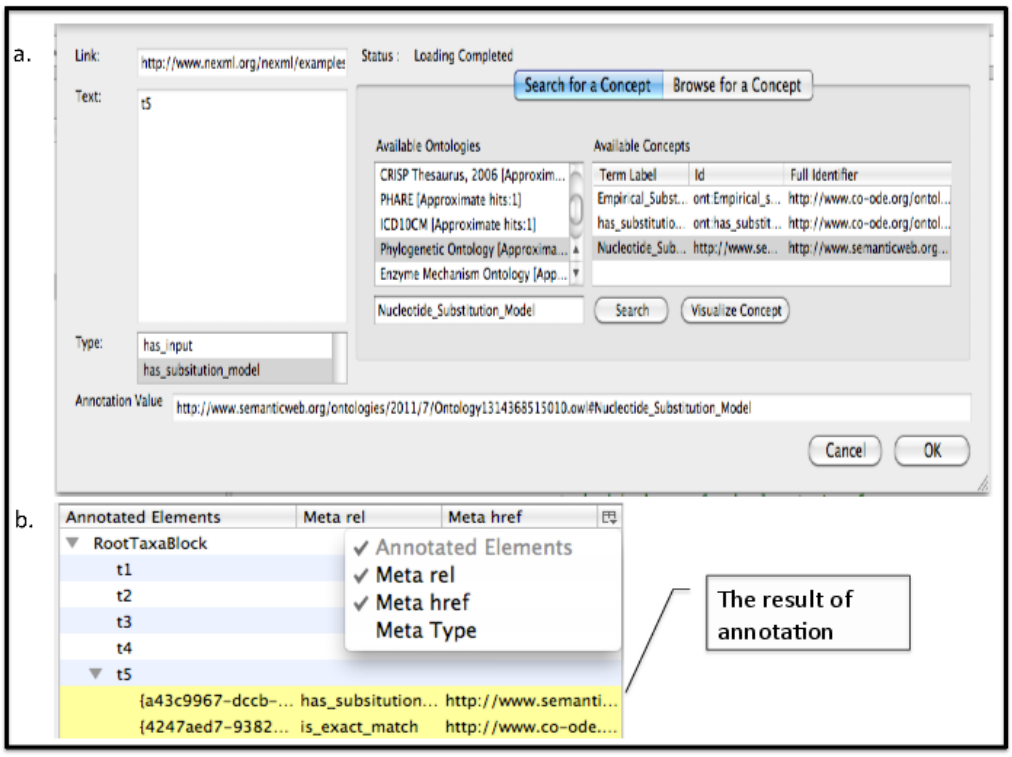

Figure 4.4: Annotation NeXML file wiht Kino-Phylo, annotate the element

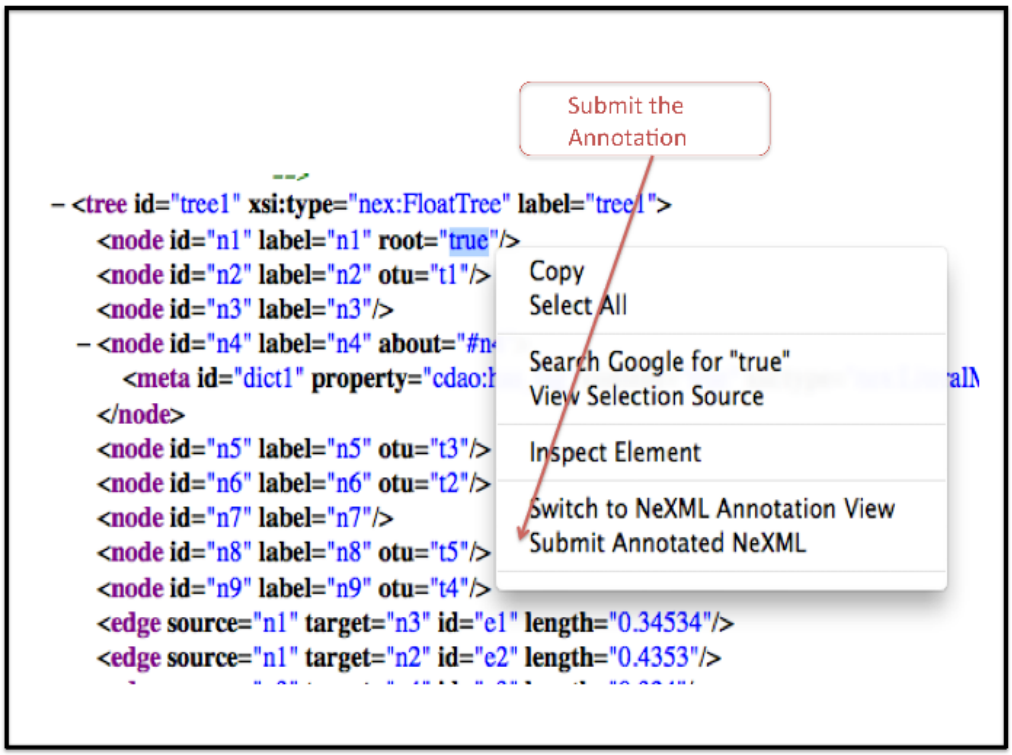

Figure 4.5: Annotation Submission

face, user can filter the results via the facets and search for annotated data. This part is the same as I explain in previous chapter for document annotation.

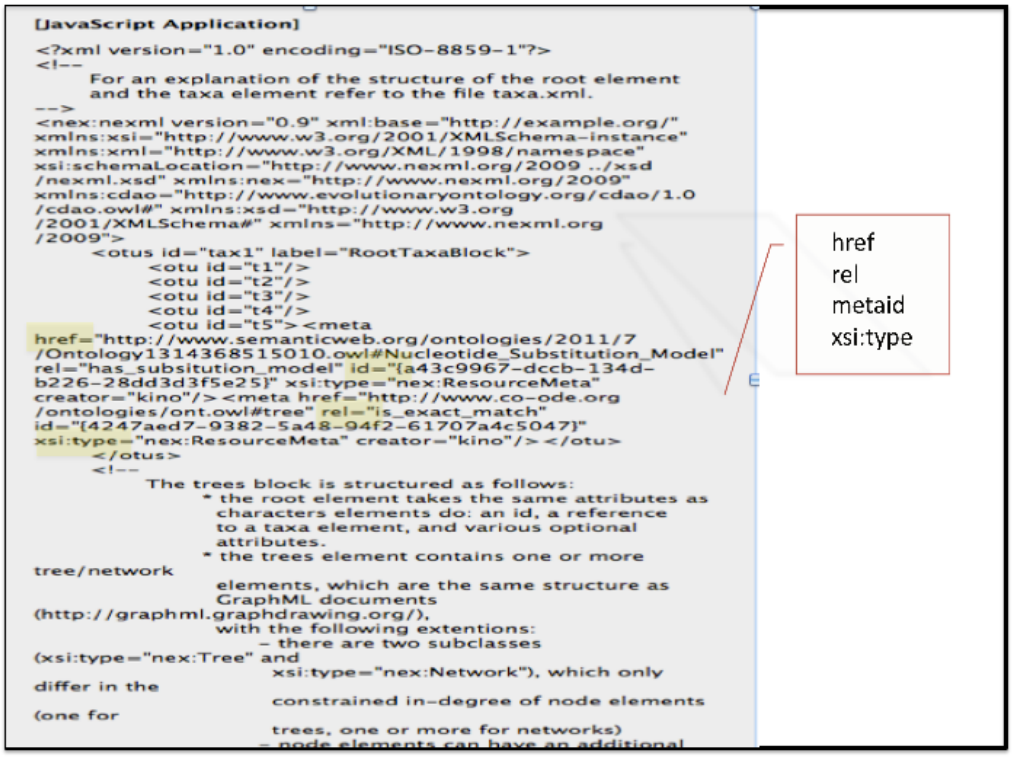

Figure 4.6: Publishing NeXML annotation

Chapter 5

MUDDIS, A Semantic Approach for knowledge Discovery

In this chapter, I discuss *MUDDIS*: a MUlti-Dimensional semantic integration approach to comparing genes for knowledge DIScovery. I regroup genes based on their functional annotations, structural annotations, genes responsible for disorders and gene-drug interactions. The driving principle in using a multi-dimensional approach is to create effective domain specific knowledge discovery, based on different gene annotations with the use of scientific and provenance information from different resources. With this multi-dimensional approach, as an example, I increase the possibility of finding more evidence to select the correct location in the corresponding gene tree in phylogeny study.

5.1 Introduction

In recent years, there has been a growing interest in comparing set of genes for knowledge discovery. Finding the similarity between genes can be useful in different areas of life science and biomedical fields such as model organism research and drug discovery in humans. The novelty of this work is that it queries through heterogeneous data sources and makes a collection of data for similarity calculations. Semantic similarity is calculated on different levels of granularity. Data from literature, open public databases, such as *OMIM* (OMIM, 1960) and gene-centered information at *NCBI* (NCBI, 2005) are used

as individual resources for different features of the gene. Each additional feature increases the value of the knowledge that can be explained within individual resources. To illustrate the utility of *MUDDIS*, I designed an evaluation framework and discussed the correlation between the *MUDDIS* similarity score and the structural similarity score, such as *HomoloGene* (HomoloGene, 2007). Also, comparing the *MUDDIS* similarity with the similarity from curated data, such as MGI-Mouse Genome Informatics (MGI, 2003) .The significance of findings at this research is to find the candidates for knowledge discovery and supporting domain experts to identify, produce and verify hypotheses based on the gene similarities inter species and intra species.

Today, with numerous genome projects, a large amount of the gene annotations data is available. The focus of the life science domain has shifted toward not only acquiring but also meaningfully using these data sets. The existing projects and data sets provide a comprehensive set of annotation tools for investigators to understand and analyse biological and biomedical data. Examples of the applications are: inferring gene function based on sequence similarity from *HomoloGene*, mining gene-disease relationships from *OMIM*, and calculating phenotypic similarity between genes based on semantic similarity (Zhang et al., 2012). However, a number of challenges exist in this field such as data integration, defining the similarity functions, finding the similarity individually based on annotation, knowledge discovery, using this knowledge for decision making, proof of the hypothesis or creating a new hypothesis.

A growing number of scientists are using genes and their corresponding information to address their research problems. These research problems range from finding the function of the gene to finding the model organism. Since gene annotation data exists separately in different data sets from literature to curated data in structured and unstructured data sets, I believe that this project provides an efficient solution to finding gene similarity across different species based on existing knowledge in different data sets (e.g., comparing gene information in species). However, the increased interest in using this resource has exposed

major limitations in the accessibility of the data sets. Most applications (Othman et al., 2008; Azuaje et al., 2005) investigated using one or two kinds of annotations. The various data sources for each annotation greatly limits the ability of scientists to use and look at different annotations from different perspectives comprehensively to find the similarity between genes.

The specific objective of this research is to develop and deploy a multidimensional integrative approach to overcome this limitation by accessing to heterogeneous data sources, and make a foundation for an integrative platform. My *MUDDIS* platform describes extracting different gene annotation data from resources that range from structured data set, such as data from *OMIM* to unstructured data sets such as scientific publications and articles. It is worth noting that the different features of selected genes have different levels of provenance, therefore we consider two types of annotations: direct annotation such as function of the gene from Entrez Gene (EntrezGene, 2006) and indirect annotations through the text such as terms from Medical Subject Headings (MeSH, 2007). Each additional feature such as disorder related to a gene or function of a gene, increases the value of knowledge that can be explained within individual resources.

The driving principle in using a multi-dimensional approach is its utility in inferring potential missing information for the gene of interest from multiple perspectives or finding similar genes in different organisms. This can be used for effective domain specific knowledge discovery, based on gene annotation. The contributions of this work are listed below:

1. Data integration for finding the similarity between genes for knowledge discovery.

2. Describing different features for gene annotation.

3. Extracting data from different data sets from structural to unstructured data sets.

4. Providing a systematic approach for finding the similarity between genes from term-term similarity to set-set, feature-feature and finally gene-gene similarity based on

semantic similarity.

5. Providing a comprehensive evaluation frame work for this platform.

The subsequent sections are organized as follows: Section 5.2 reviews the related work. Section 5.3 presents the challenges and opportunities in this field. Section 5.4 explains how to select different features to study different gene annotations and how to select related data sets for each feature such as function annotation, disorder annotation, drug annotation and structural annotation of a gene. Section 5.5 describes the similarity functions to find the similarity between genes. Section 5.6 describes use-cases in biological and biomedical domains. Section 5.7 presents the evaluation platforms and results. Section 5.8 provides discussion and conclude the chapter.

5.2 Related Works

Rapidly increasing number of published gene and gene annotation data created significant opportunities for integrating data from various source and using the data for knowledge discovery. The task of assigning meaningful annotations to the gene, such as function of the gene, has been a considerable challenge for a long time. Finding the similar genes based on their annotation has been successfully addressed by different studies such as (Zhang et al., 2012; Azuaje et al., 2005).

However, the missing piece is to use this information as an integrated platform to integrate the gene data from various sources and extract knowledge from them to find the similarity between genes from a different perspective. In the following parts of this section, I introduce some of the related work for developing the similarity functions. Then, I discuss the similarity based on different annotations which, they are limited both from aspects of annotation and coverage of species.

5.2.1 Similarity Measures

In some domains such as finding the model organism for drug discovery, similarity function is necessary to find similar organisms from different perspective. Similarity functions could be varied based on the nature of the data and the usage. In this section I introduce different similarity functions, advantages, disadvantages of each, and I discuss how they are related to my work. Wu and Palmer (1994) proposed a taxonomy based approach for similarity measures. The idea is to exploit the geometrical model provided by concept hierarchies. They estimate the distance between two terms to find the shortest path between them, which is the minimum number of links that connect these two concepts. However, it is worth noting that base on the shortest path, the similarity between a pair of concepts in an upper level of the taxonomy is smaller than the similarity between a pair in a lower level. To this end, they proposed a path-based measure that also takes into account the depth of the concepts in the hierarchy. Leacock and Chodorow (1998) proposed a method with consideration of the maximum depth of the taxonomy and also the shortest path between two concepts.

Mubaid, H and Nguyen (2006) proposed a method that combines path length and common specificity as a cluster-based method. Clusters are defined for each branch in the hierarchy with respect to the root node. The common specificity measures by subtracting the depth of the least common concepts of the selected concepts from the depth of the cluster. It means that the lower level pairs of concept nodes are more similar than higher level pairs because they have more common nodes. The advantage of the measured based on taxonomy structure is that it only uses an ontology or the controlled vocabulary with the hierarchy as background knowledge. However, the problem is that they depend on the degree of homogeneity and coverage of the semantic links represented in the ontology (Sánchez et al., 2012; Batet et al., 2011).

Resnik (1999) used information content to evaluate semantic similarity in a taxonomy. It measures the amount of information provided by a given term based on the probability

of appearance in selected content and tags the selected terms in the content. Due to the different restrictions such as security issues of data or the amount of data, an Information content-based approach may be compromised by the availability and complexity of suitable data.

Another approach is the "context vector relatedness measure", it computes the similarity by the hypothesis that if the terms are similar, then the context included those terms is also similar. Patwardhan et al. (2003) created vectors from term extracted from Wordnet, which they represent as the ideal context of the selected term. They stated that the quality of the assessment at this approach is strongly dependent on the size and nature of the data that the context is extracted from. Just like the Information-content-based method, availability and suitability of the content is a major restriction for this similarity measure. Also, because of the complexity of biological and biomedical data, we could have similar terms but they could appear in two different contexts which they are not necessary similar to each other.

Batet et al. (2011) proposed that from the applicability point of view, path-based measures are the most adequate ones, and no pre-processing is needed. In order to take into the account the number of common information between two terms, their measure is the ratio between the amount of non-shared knowledge and the sum of shared and non-shared knowledge. Finally, from the study of different similarity measures, I can conclude that my method for similarity is based on Batet's similarity for term-term similarity and calculate set-set similarity for each feature and finally gene-gene similarity as an aggregated method over different features, on the other hand different annotations of the gene, which I will explain that in Section 5.5.

5.2.2 Investigating Semantic Similarity for Multiple Features

Most of the previous work is limited from both aspect of the comparison (Patwardhan et al., 2003; Zhang et al., 2012) and the coverage of species (Bodenreider and Burgun, 2010).

For example *HomoloGene* (HomoloGene, 2007) is an automated system for detecting homologs among eukaryotic gene sets based on sequence similarity. Another example is *MGI* as an international database resource for the laboratory mouse. *MGI* provides integrated genetic, genomic, and biological data for researching human and mice health and it is mostly curated with experts.

Azuaje et al. (2005) proposed a framework for comparing phenotype annotations of orthologous genes based on the Medical Subject Headings (*MeSH*) indexing of biomedical articles in which these genes are discussed. Pairs of orthologous genes of mouse and human from the Mouse Genome Informatics system are downloaded and linked to biomedical articles through *Entrez Gene*. *MeSH* index terms for each disease are extracted from *Medline*. I make the current research based on this study and I extended this study both for the resources and the similarity functions.

5.3 Challenges and Opportunities

The major motivation for this research is to develop a foundation for scientists to study and search for different genes between or within species. This idea came from the way PubMed searching for particular publication. As shown in Figure 5.1. in the process of searching for particular paper, all of the citations for similar papers show up in the right panel of search. As shown in Figure 5.2. My idea is, when users search for particular gene such as *BRCA1*, all of the similar genes in different species show up in the right panel. This similarity is based on different annotations of the gene such as structure similarity, function of the genes, etc.

This platform is useful for scientists that search for particular gene of interest as well. Given the ability to find the similar genes from different perspectives and show them to scientists could be useful as well. The most significant challenge in finding gene similarity from different perspectives is the variety of data in different resources. Another challenge

Figure 5.1: Motivation for MUDDIS Platform

is how to define the similarity functions for each feature or on the other hand for each gene annotation at different levels. Taking into the account semantic similarity is one of the keys to make sure the similarity is meaningful and valuable. In summary, the most important challenges in this field are the following:

1. The variety of data sources such as curated data sources, external, experimental and data from the scientific literature.

2. The variety of similarity functions because of the level of similarity, and nature of the annotation gene, which we calculate similarity for them. For example calculate the similarity between functions related to gene could be different from the way we could calculate the biological process that each gene of interest involves in that biological process.

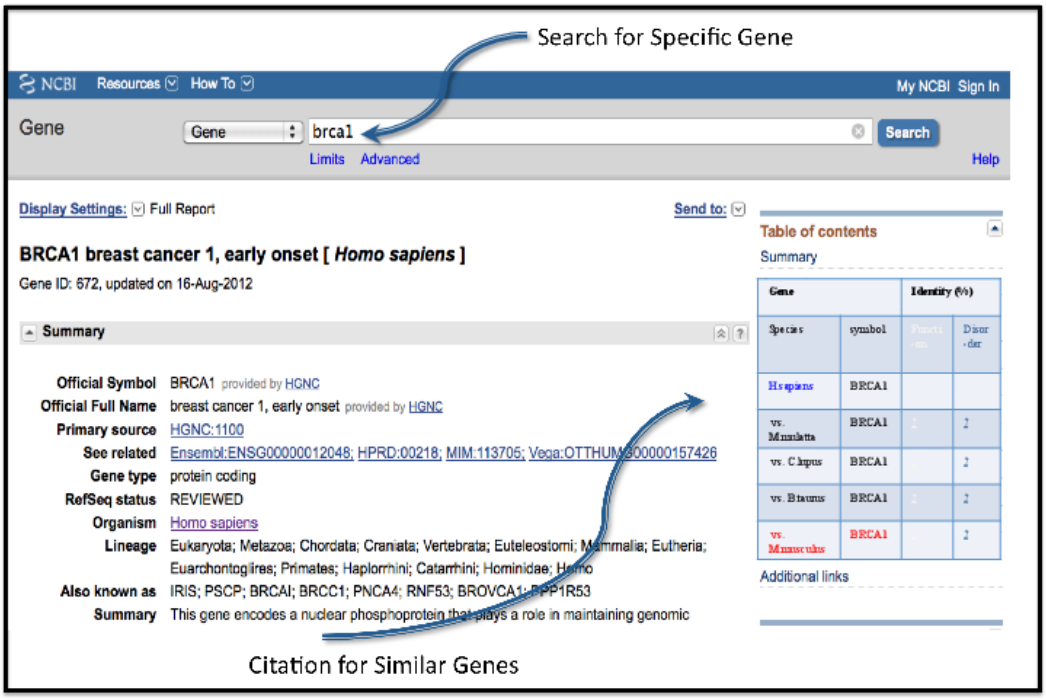

Figure 5.2: Finding the Similar Genes from Different Perspectives

3. The lack of finding semantic similarity during the calculation of similarity functions.

4. The lack of finding similarity in a comprehensive platform.

5. Lastly, the lack of information about the gene similarity, which have created by experts. Because of the nature of manually creating and gathering information, the curated data are limited. It means that not only they do not cover most of the organisms, also they are limited for finding the similarity based on different annotations of the gene in an integrated framwork. For example they just calculated the similarity based on functions of the gene, based on diseases related to gene, based on phenotype, etc. Current studies which I have mentioned in related work do not give the opportunity to scientists to study how genes are similar based on different annotation. This framework gives the scientist a good foundation to study related genes in comprehensive platform.

My focus in this research is to select different gene annotations data from structured to unstructured data sets. I also, define the similarity functions in the way that carefully captures the semantic similarity at each level of calculation from term-term similarity to gene-gene similarity. With this approach I make sure genes are being compared from different dimensions based on various gene annotations data. Therefore, if there is missing information, there will be enough knowledge to help scientists to test existing hypothesis or come up with a new hypothesis as a starting point for novel studies.

5.4 Data Collection

The first step is to select the features in the domain of study. This is usually achieved by reviewing and harvesting existing features and related data sources. In order to select the desired sources for this study, I consulted with experts and producers of data sources as well as studying data resources.

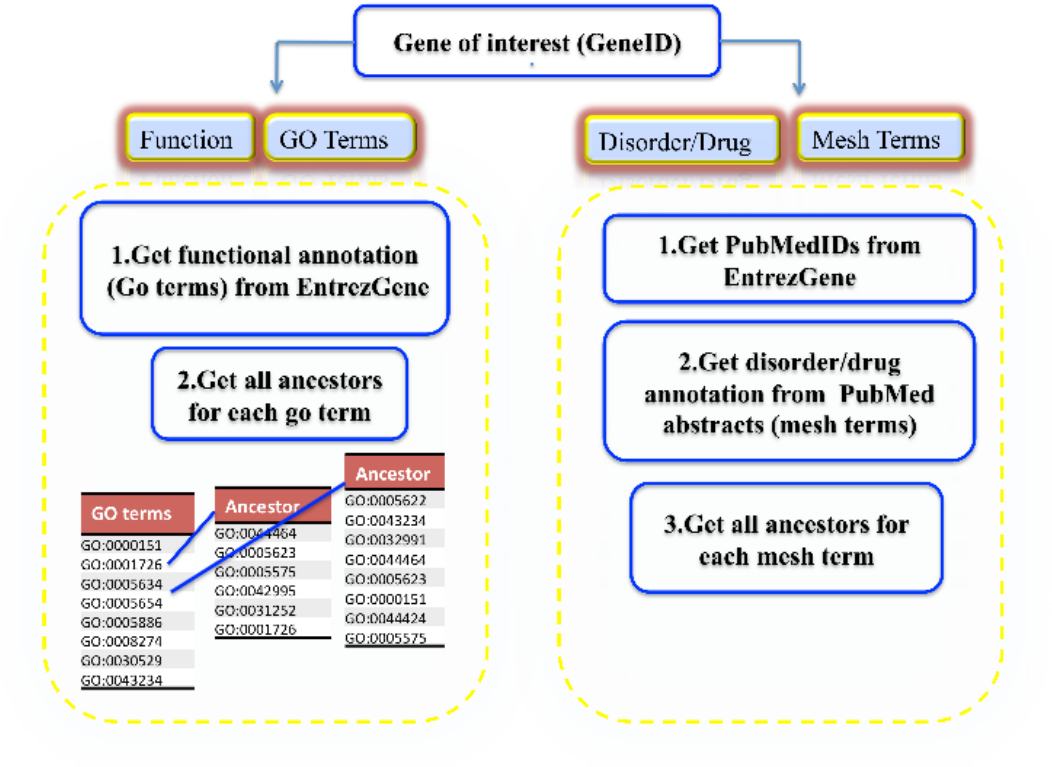

Figure 5.3: Data Collection for Selected Features

5.4.1 Candidate Features for Gene Annotation

In this study, I compare genes using annotations that include:

- Functional annotations
- Structural annotations
- Genes responsible for disorders
- Gene-drug interactions

5.4.2 Data Source of each Feature

Data sources in this study can be classified into three types of data: a) data extracted from scientific literature and academic articles, b) data from structured data sets and well known sources, and c) data curated with experts. Since part of valuable experimental data are still published as a literature and are not in structured easily accessible data sets, using data from literature is necessary in this study. Figure 5.3 demonstrates the steps of data collection. The following lists the data sources that are used in my data collection phase.

- *Entrez Gene* is an integration system developed by *NCBI* to extract data for different features of the gene of interest. *Entrez Gene* allows to access Gene Ontology terms, articles related to the gene, *PMIDs* and *MeSH* terms assigned to these articles to get disorders and drugs related to the gene of interest.

- *HomoloGene* is used for evaluation part. It detects homologs among the annotated genes of several completely sequenced eukaryotic genomes. The scores of the *HomoloGene* are from sequence alignment for Both *DNA* and Protein sequences.

- *MGI* uses Mouse Genome Informatics to access integrated curated data on mouse genes and genome features, from sequences and genomic maps to gene expression and disease models. *MGI* is a repository for raw data and detailed protocols from the Mouse Phenome Project (Aylwin et al., 2006), it collects baseline phenotypic data on genetically diverse and commonly used inbred mouse strains.

5.5 A Systematic Approach for Gene-Gene Similarity

I evaluate the relatedness among genes using heterogeneous features. Each gene of interest will be associated with the set of annotation features. With this multi-dimensional approach, there is more evidence to proof the existing knowledge or detect the missing information for gene annotation to compare with existing gene or species tree. For this purpose,

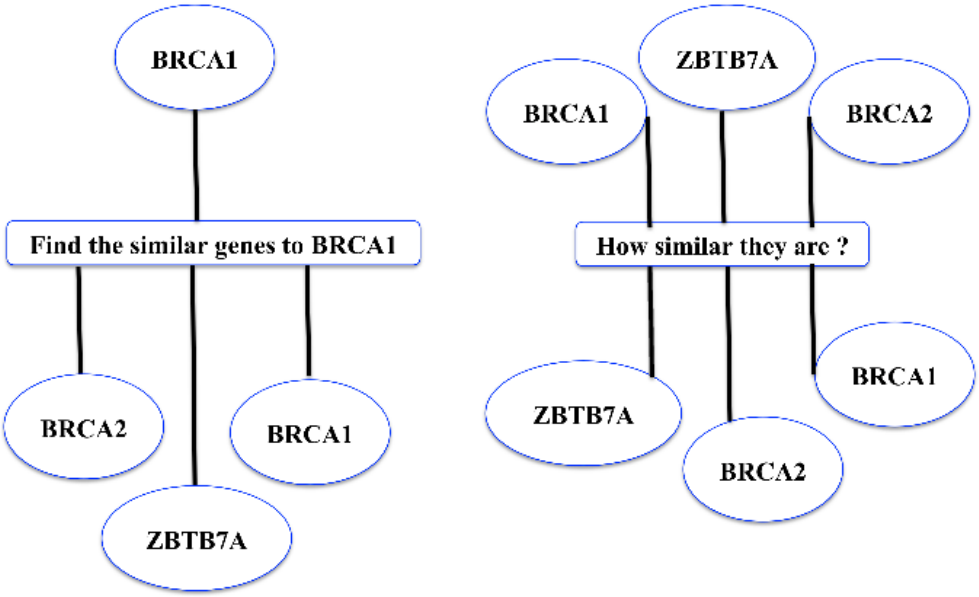

Figure 5.4: Two Scenarios of the Gene-Gene Similarities

I follow 3 steps to calculate the gene-gene similarity: term-term similarity, set-set similarity for each feature, and finally aggregation over feature-feature similarity for gene-gene similarity. I partially use some of the function introduced in previous work and extended them for this multidimensional purpose. (Batet et al., 2011; Bodenreider and Burgun, 2010).

5.5.1 Gene-Gene Similarity

Gene-Gene similarity is the similarity between genes based on multiple features. Sometimes we have a gene of interest and we want to find a similar genes or we have set of genes and we want to know how they are similar to each other. Figure 5.4 demonstrates the two scenarios of Gene-Gene similarity. As shown in Figure 5.5, the bottom-up approach to calculate Gene-Gene similarity involves calculation of the Term-Term similarity, then

finding the Set-Set similarity for each feature and finally calculating the overall Gene-Gene similarity. In Gene-Gene similarity, I assign a weight to each feature and aggregate all features. Suppose that w_q is the weight assigned to each feature from 1 to n. Where g_i and g_j correspond to selected genes, the similarity (sim) between these two genes can be calculated by the following equation:

$$FSim\left(g_i, g_j\right) = \sum_{q=1}^{n} w_q sim_q\left(g_i, g_j\right) \tag{5.1}$$

One of the advantages of this aggregation is to make a flexible framework and give different priority to features based on the expert's selection. Each feature in this equation includes a set of concepts, therefore I need to calculate Set-Set similarity for each feature.

5.5.2 Set-Set Similarity

For Set-Set similarity, a pair of genes given, g_i and g_j, which are annotated by the set of term t_i and t_j, respectively, the similarity is calculated by the average inter-set similarity between terms from both sets, by finding the maximum similarity of all the terms in first set with m members to all terms in second set with n members and vice versa and find the average through them. Set-Set similarity (Azuaje et al., 2005) can be calculated by the following equation:

$$Sim(g_i, g_j) = \frac{1}{m+n} * \left(\sum_{a \epsilon t_i, b \epsilon t_j} max\left(sim(t_a, t_b)\right) + \sum_{b \epsilon t_i, a \epsilon t_j} max\left(sim(t_a, t_b)\right) \right) \tag{5.2}$$

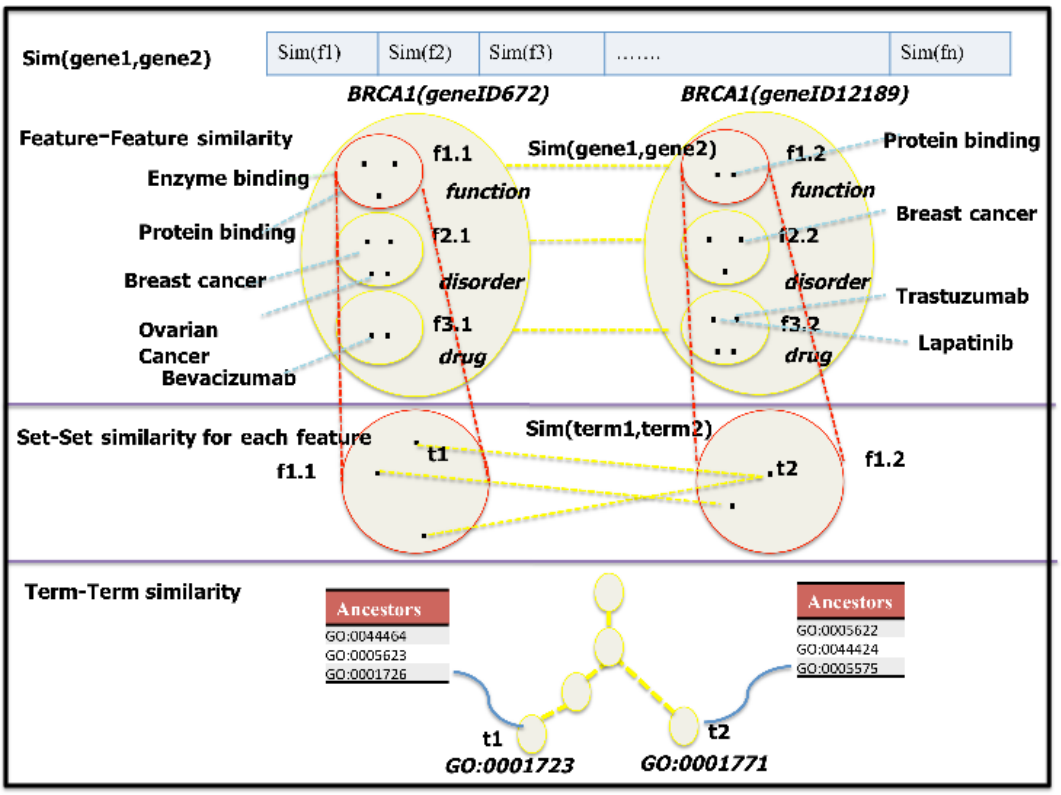

Figure 5.5: Steps to Find the Similarity Functions between Genes

5.5.3 Term-Term Similarity

Analyzing the path-based methods, which I mentioned in related work, the minimum path length between a term t_i and a term t_j is calculated, which is the sum of taxonomical links between each of the terms and their least common subsumer (LCS). The path is composed of the *LCS* and the nodes corresponding to non-shared super-terms or parents. Therefore, if one or both terms inherit from several is-a hierarchies, all possible paths between the two terms are calculated, but only the shortest one is kept. It means that, the resulting path length does not completely measure the total amount of non-common super terms in the ontology. For this reason, path-based measures waste a huge amount of knowledge. Therefore, in this study I use the method proposed with (Batet et al., 2011), that takes into account all the available taxonomical evidence to capture enough semantic evidence, for term-term similarity. To capture enough semantic evidence in the case of multiple inheritances, for term-term similarity, they take all super terms belonging to all the possible taxonomical paths connecting the evaluated terms. They consider the terms and their complete set of non-shared super-terms as an indication of their distance. By considering terms themselves in conjunction with the set of non-common super-terms, they calculate the similarity for a pair of terms that are siblings of an immediate superclass.

However, by considering only non-shared knowledge, it is impossible to distinguish terms with very few or even no super terms in common from others with more communal information. In order to take into account the amount of common information between a pair of terms, (Batet et al., 2011) define the measure as the ratio between the amount of non-shared knowledge and the sum of shared and non-shared knowledge. Considering that shared and non-shared knowledge explicitly retrieved from a repository for a term pair is not linear to their similarity/distance, they introduce the inverted logarithm function to smooth the assessments and to transform the function into a similarity.

Therefore, I redefined the similarity between each term and its ancestors from first set for the first gene to each term and its corresponding ancestors from second set in second

gene. Then I calculate the ratio between the amount of non-shared knowledge and the sum of shared and non-shared knowledge. Finally use an inverted logarithm to transform the function into a similarity. I define the all term hierarchy or taxonomy (H^t) of terms (T) of an ontology as a $is-a$ relation (H^t) and $T(t_i)$ = $\{t_i \epsilon T | t_i \; issuperterm \; of \; t_i\} \cup \{t_i\}$ as the union of the ancestors of the concept t_i and t_i itself. Then the similarity function between two terms is defined as:

$$Sim\left(t_i, t_j\right) = \log \frac{\left|T\left(t_i\right) \cup T\left(t_j\right)\right| - \left|T\left(t_i\right) \cap T\left(t_j\right)\right|}{\left|T\left(t_i\right) \cup T\left(t_j\right)\right|} \tag{5.3}$$

5.6 Use Case Scenarios

5.6.1 Make a Foundation for Knowledge Discovery

A fundamental driving principal for the similarity functions is to use them for finding the similarity between genes, and use that for knowledge discovery. In this study, as a use case, I compare a sample gene (i.e., *BRCA1*, which is a human gene that produces a protein called breast cancer type 1 susceptibility protein, responsible for repairing *DNA* in humans (Aylwin et al., 2006) versus other organisms such as *M. musculus, C. lupus, R. norvegicus, B. taurus, G.gallus, P. troglodytes,* and *M. mulatta*. "BRCA1 is expressed in breast cells and other tissue, and it helps to repair damaged DNA, or to destroy cells. If BRCA1 is damaged, damaged DNA is not repaired and this makes the possible risks for cancers" (Aylwin et al., 2006).

For this comparison, I found the similarity for the function of the gene, disorder, drug related to selected gene, and aggregate all features by assigning equal weight to each feature. As I mentioned earlier these weight are flexible and can be changed based on the experts 's need. As shown in Table 5.1, the *M. mulatta* has the highest similarity score. It means based on *MUDDIS* aggregated platform the similar gene to *BRCA1* in human is *BRCA* in *M. mulatta*. This result is not the same as the result from *HomoloGene* which

Vs. Species	GeneID	P[1]	D[1]	F[2]	Di[2]	Dr[2]	A[2]
M.musculus	12189	58.1	74.4	62.5	45.4	33.5	47.1
C.lupus	403437	74.4	84.2	67.3	58.3	55.5	60.3
R.norvegicus	24227	58.3	75.2	41.2	22.4	47.1	36.9
B.taurus	353120	72.6	83.8	65.8	33.4	39.5	46.2
G.gallus	373983	34.3	50.0	50.0	43.4	52.2	48.5
P.troglodytes	449497	98.2	99.3	63.4	60.1	58.5	60.6
M.mulatta	712634	93.1	96.1	69.5	60.8	61.5	63.9

[1]Pairwise alignment scores from *HomoloGene* for protein and DNA sequence similarity
[2]Function, Disorder, Drug and Aggregation are results from *MUDDIS* similarity functions

Table 5.1: Calculate Similarities of *BRCA1* from Homo Sapiens vs. Other Species

is the benchmark at this experiment. This could be a starting point for a researcher to figure out why the similarity scores are difference. By adapting the gene-gene similarity function and assigning the weights to different features, the possibility of finding the reason for the differences could be because of the different function or disorder or drug related annotation. Then, this could be an starting point to test in the wet lab and find a reason for this differences and discover new knowledge based on existing data.

5.6.2 A Solution for Gene Fusions Problem

Two genes in one organism can be fused into a single gene in another organism or visa versa. When the orthology mapping involves gene fusions, sequence similarity-based methods can not be effectively used (Li et al., 2011). Orthologous genes refer to genes that have evolved from a common ancestor. Li et al. (2011) propose that one issue is that the definition of orthologous genes is not operational, unless *phylogenetic* trees could be accurately derivable. Analysis methods could be available for distinguishing orthologous from paralogous genes.

The current orthology-mapping programs and tools can be classify into two groups, *phylogenetic*-based and sequence similarity-based. *phylogenetic*-based methods are generally more reliable than sequence similarity-based methods. One issue with all sequence

similarity-based methods is that they assume sequence similarity information is sufficient for finding the orthologous relationships (Li et al., 2011). However, the assumption made by similarity based methods is not always true due to the gene fusions. This issue can make a sequence similarity-based orthology mapping methods fail. The hypothesis is that using other gene annotations mechanisms such as functional annotation, in the case that sequence similarity alone is not sufficient for finding the orthologous relationships.

5.6.3 Comparing with Phylogeny Tree

There are two types of gene transfers: duplicative transfers and orthologous replacement transfers (Li et al., 2011). Orthologous replacement transfers sometimes cause false outcomes of ortholougs. For example, in comparing between a gene tree and a species tree, we find the corresponding location for the species tree in the gene tree. However, this information in some cases can be found in a different location or even in two locations. By considering some other annotations of the gene and find the similarity between genes based on these annotation instead of using just sequence similarity, we increase the possibility of finding more evidence to select the correct location in the corresponding gene tree.

5.7 Evaluation and Discussion

The formal way of evaluating the accuracy of similarity measures is based on using a set of gene pairs whose similarity has been curated with the experts. There are no widely accepted benchmark data sets of similar genes across different genomes. In *Homologgene*, homolog a gene related to a second gene by descent from a common ancestral *DNA* or protein sequence as a structural annotation based on sequence similarity (HomoloGene, 2007).

Existing curated data are limited for a specific organism, for example data from *MGI* includes integrated data on mouse genes from sequences and genomic maps to gene expres-

sion and disease models. In my validation framework, I consider both aspects and species limitations. I follow the proof by contradiction in mathematical logic. The hypothesis is that the results of *MUDDIS* similarity functions go with control similarity and I assume that this is true. The control similarity could be the structural similarity from *HomoloGene* and curated data from *MGI*.

5.7.1 Evaluation Framework Based on Structural Annotation

I calculate the correlation between *MUDDIS* similarity and structural similarity based on a pairwise sequence alignment. The hypothesis is that the multi-feature similarity functions correlate with the HomoloGene sequence similarity. If there is a correlation then support of hypothesis and no new finding. If there is not a correlation, then will reject the hypothesis. It means, if I get the Higher *MUDDIS* similarity then I Isolate the features responsible for the deviation and investigate the potential causes in wet lab experiments. If I get the lower *MUDDIS* similarity, then investigate the methods to calculate sequence similarity for *HomoloGene*.

5.7.2 Evaluation Framework Based on Curated Data

If I remove gene pairs from my data set, which are known to be orthologous based on curated data, I expect data with lower functional similarity. This framwork is based on curated data and as I mentioned before there are limitations on using curated data. For most organisms this data is not available, the only available one is data from *MGI* which supplies experimental literature-based annotation to mouse gene products. *MGI* includes gene annotations such as Phenotype, disease model, function, expression, pathways, and orthology. *MGI* could be a good benchmark for the evaluation, but limited to human-mouse comparison.

5.7.3 Discussion and Significance of Finding

Based on my results in the similarity section the correlation between the score from *MUDDIS* and *HomoloGene* is 0.67. The results of rejection hypothesis cab be used for knowledge discovery. I expect a good correlation between these two scores. If there was no correlation then it will be an starting point for analysis and finding the reasons for that. It is a sign for the possibility of missing information in the structural similarity . By scanning different annotations and following that scanning each individual feature, I can find that the function of these two genes, which are represented by the Gene Ontology term, are not the same for the gene of interest even though they have the have the same sequence.

5.8 Conclusion

With an ever-growing number of data sources available in life science, both in structured and non-structured data sets, looking at this valuable data in a comprehensive way would be helpful for scientist to analyse the data and use them for knowledge discovery, knowledge validation, decision support or hypothesis creation. I proposed a platform to create a foundation for knowledge discovery called *MUDDIS*. The results of similarity features provided by *MUDDIS* for identifying the correlation between *MUDDIS* score with the control from different perspectives could be useful in model organism research. In my future work I will apply this framework in different use cases, adding or removing dimensions for similarity comparison. I also will develop an interactive user interface for this system.

Chapter 6

Conclusions and Future Work

The research presented in this dissertation aimed at applying Semantic Technologies to *Phyloinformatics*. I addressed this aim from both a phylogenetics and a computer science perspective. The first one, from the *phylogenetic* community perspectives, states that the reusability of *phylogenetic* information is possible. The second one, from a computer science perspective, states that the integration of different data items would facilitate interoperability among various researchers. Such practices would allow researchers to access, explore and reuse the products of *phylogenetic* studies including innovative workflows.

6.1 Summary

I reported on four research efforts to achieve this goal. The *PhylOnt* Project encompasses building a repository for *phylogenetic* study and the implementation of the*PhylOnt* ontology. *PhylOnt* is the first ontology specifically extended for *Phylogenetic* analysis operations and related metadata. As of Sep 2012, the 6th version of *PhylOnt* has been submitted to *NCBO* after extensive in-house evaluation and community-feedback (Panahiazar et al., 2011, 2012a). The result of this project was published in (Panahiazar et al., 2012b).

Then, I proposed the *PhylAnt-D* project, which is about emantically annotating *phylogenetic* documents using ontologically grounded concepts as a foundation for semantic technologies. It is a preliminary step to semantic search, information retrieval, and heterogeneous data integration that can support *Phylogenetic* workflows.

A major part of PhylAnt-D is the Kino-Phylo annotation tool, which extends our previous annotation software Kino (Ranabahu et al., 2011a) to handle domain specific annotation. Related to PhylAnt-D is the PhylAnt-X project, in which I extended Kino-Phylo to annotate XML files. In particular, the NeXML file standard (Vos et al., 2012). PhylAnt-X is the package of annotation, indexing and searching tools.

As a forth part of this research, I designed and implemented the *MUDDIS* System, a MUlti-Dimentional semantic integrative approach to comparing genes for knowledge DIScovery. The major motivation for this research is to make a foundation for scientists to study and search for different genes between or within species. During the study of *phylogenetic*, I observed that there is a need to add provenance information to *phylogenetic* studies. In the process of orthologs replacement transfers (Li et al., 2011), orthologous replacement transfers sometimes cause false outcome of orthologs. For example in comparing between a gene tree and a species tree, the corresponding location for the species tree should be at the same location in the gene tree. However, this information in some cases can be found in a different location or even in two locations.

By considering some other annotations of the gene and finding the similarity between genes based on these annotations instead of using just sequence similarity, we increase the possibility of finding more evidence to select the correct location in the corresponding gene tree. This project defined a new similarly function, which is the aggregation of different levels of similarity between genes and also integration of different kind of data sources. Data from data bases such as *OMIM*, data from literature and PubMed articles, data from *EntrezGene* are integrated in this study. Then I made a federated query over different kinds of data sources, collect the data and calculate the similarity functions.

6.2 SemPhyl Platform

As a future work I will work on SemPhyl platform. I proposed the *SempPhyle* platform in *AMIA* (Panahiazar et al., 2012a), which is the integrated platform for *Phylogenetic* study from integration to visualization of results. It provides a foundation to allow the integration of different sources of information to answer questions at multiple levels of granularity using *phylogenetic* and search through different data sources.

This platform combines existing tools, some of which are described in this paper, and I adjust them for use in this platform with the consideration of reusability of *phylogenetic* resources.

6.2.1 Layers and Architecture of SemPhyl Platform

As shown in Figure 6.1. layers in this platform categorize into four layers including data layer, semantic layer, query engine layer, and interface layer. Data Layer includes all possible data in *phylogenetic* analyses such as *NeXML* files, scientific literature in *TreeBASE* and *PubMed*, experimental data, *phylogenetic* programs, web services, and results from *phylogenetic* analysis. The semantic layer is about semantic annotation data with *PhylOnt* for search, and reusability of data. Query engine layer is about making the federated queries over the heterogeneous data sources using *SPARQL*, and finally in interface layer, I show the web form for submitting data and different kinds of visualization of the system. Figure 6.2 shown the architecture of the*SemPhyl* platform.

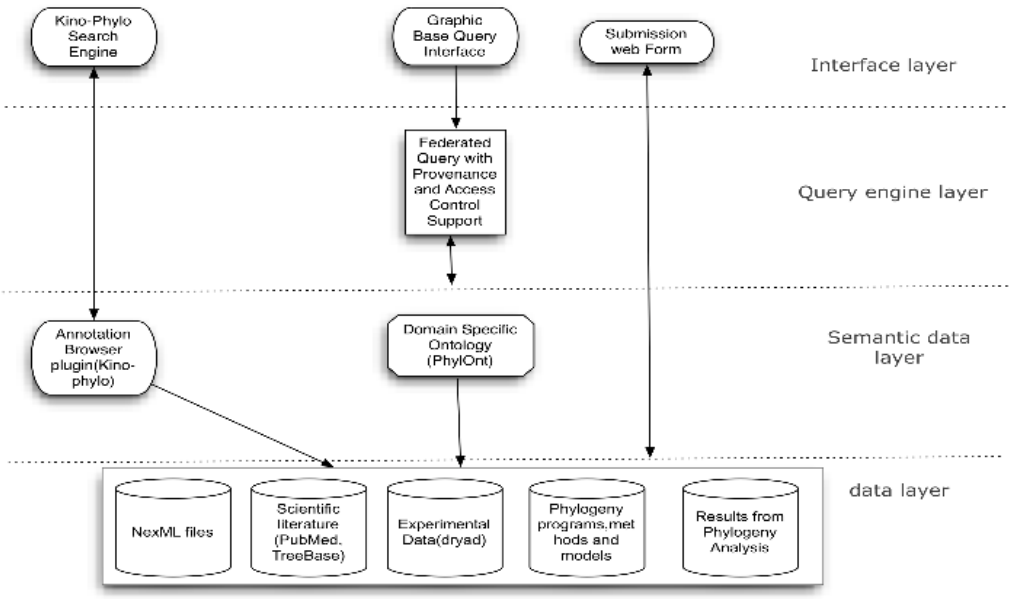

Figure 6.1: Layers of SemPhyl Platform

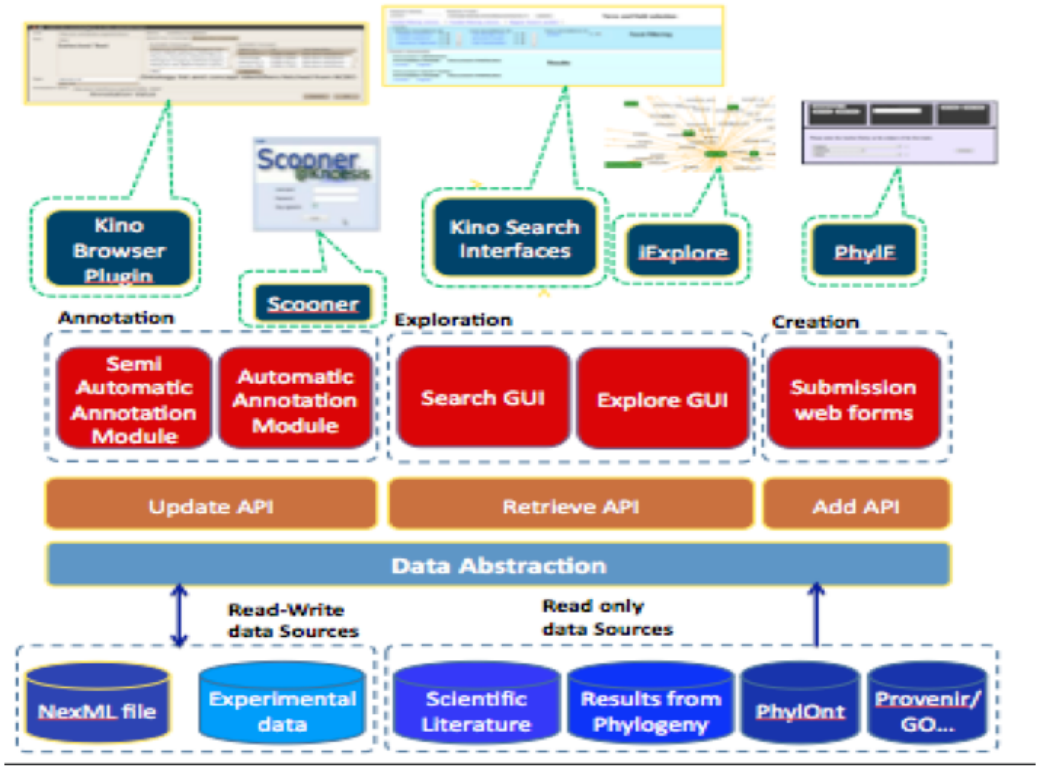

Figure 6.2: Architecture of SemPhyl Platform

Bibliography

Aylwin, N., Borisas, B., Qiong, G., Ewan, M., Marketa, Z., 2006. Resources for integrative systems biology: from data through databases to networks and dynamic system models. In: Brief Bioinform. pp. 318–330.

Azuaje, F., Wang, H., Bodenreider, O., 2005. Ontology-driven similarity approaches to supporting gene functional assessment. Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies, 9–10.

Barker, D., Pagel, M., Jun. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS computational biology 1 (1), e3.

Batet, M., Sánchez, D., Valls, A., Feb. 2011. An ontology-based measure to compute semantic similarity in biomedicine. Journal of biomedical informatics 44 (1), 118–25.

Bodenreider, O., Burgun, A., Jan. 2010. A framework for comparing phenotype annotations of orthologous genes. Studies in health technology and informatics 160 (Pt 2), 1309–13.

Botstein, D., Ball, C. A., Blake, J. A., 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., Rubin, G. M., Sherlock, G., 2000. Gene Ontology : tool for the unification of biology. Nat Genet 25 (1), 25–29.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J., Rajandream, M.-A., Dec. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics (Oxford, England) 24 (23), 2672–6.

Chisham, B., Wright, B., Le, T., Son, T. C., Pontelli, E., Jan. 2011. CDAO-store: ontology-driven data integration for phylogenetic analysis. BMC bioinformatics 12 (1), 98.

Cross, V., Parikh, P. P., Panahiazar, M., 2011. Aligning the Parasite Experiment Ontology and the Ontology for Biomedical Investigations Using AgreementMaker. In: ICBO: Internationla Conference on Biomedical Ontology. pp. 2–8.

Cruz, I. F., Xiao, H., 2005. The Role of Ontologies in Data Integration. Journal of Engineering Intelligent System, 1–18.

Edstam, M. M., Viitanen, L., Salminen, T. a., Edqvist, J., Nov. 2011. Evolutionary history of the non-specific lipid transfer proteins. Molecular plant 4 (6), 947–64.

EntrezGene, 2006. NCBI's repository for gene-specific information.
URL : ura.wi.mit.edu/entrez_gene/

Gaudet, P., Livstone, M. S., Lewis, S. E., Thomas, P. D., Sep. 2011. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Briefings in bioinformatics 12 (5), 449–62.

Harrison, C. J., Langdale, J. a., Feb. 2006. A step by step guide to phylogeny reconstruction. The Plant journal : for cell and molecular biology 45 (4), 561–72.

Hladish, T., Gopalan, V., Liang, C., Qiu, W., Yang, P., Stoltzfus, A., Jan. 2007. Bio::NEXUS: a Perl API for the NEXUS format for comparative biological data. BMC bioinformatics 8, 191.

HomoloGene, 2007. HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes.
URL : http://www.ncbi.nlm.nih.gov/homologene

Jonquet, C., Shah, N. H., Cherie, H., Musen, M. A., Callendar, C., Storey, M.-A., 2009. NCBO Annotator : Semantic Annotation of Biomedical Data. In: ISWC. pp. 2–3.

Lamprecht, A.-L., Naujokat, S., Steffen, B., Margaria, T., Dec. 2010. Constraint-Guided Workflow Composition Based on the EDAM Ontology. Nature Precedings.

Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: WordNet: an electronic lexical database. MIT Press. pp. 265–283.

Leebens-Mack, J., Vision, T., Brenner, E., Bowers, J. E., Doyle, J. J., Eisen, J. A., Gu, X. U. N., Harshman, J., 2006. NIH Public Access. OMICS 10 (2), 231–237.

Li, G., Ma, Q., Mao, X., Yin, Y., Zhu, X., Xu, Y., Dec. 2011. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes. Nucleic acids research 39 (22), e150.

Maddison, D. R., Swofford, D. L., Maddison, W. P., 1997. NEXUS: an extensible file format for systematic information. Systematic Biology 46 (4), 590–621.

MeSH, 2007. Medical Subject Headings.
URL : http://www.ncbi.nlm.nih.gov/mesh

MGI, 2003. Mouse Genome Informatics.
URL : http://www.informatics.jax.org/

MIAPA/PhyloWays, 2011. A list of interpreted phyloinformatics workflows.
URL : http://www.evoio.org/wiki/MIAPA/PhyloWays

Mubaid, H, A., Nguyen, H., 2006. A cluster-based approach for semantic similarity in the biomedical domain. In: Conference proceedings of the IEEE engineering in medicine and biology society. New York, USA. pp. 2713–7.

Müller, H.-M., Kenny, E. E., Sternberg, P. W., Nov. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS biology 2 (11), e309.

NCBI, 2005. National Center for Biotechnology Information.
URL : http://www.ncbi.nlm.nih.gov/

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., Musen, M. A., Jul. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research 37 (Web Server issue), W170–3.

OMIM, 1960. Online Mendelian Inheritance in Man.
URL : http://www.ncbi.nlm.nih.gov/omim

Othman, R. M., Deris, S., Illias, R. M., Feb. 2008. A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. Journal of biomedical informatics 41 (1), 65–81.

Panahiazar, M., 2011. PhylOnt: An Ontology for Phylogeny analyses.
URL : http://bioportal.bioontology.org/ontologies/1616

Panahiazar, M., Leebens-Mack, J., Ranabahu, A., Sheth, A., 2012a. Using semantic technology for Phylogeny. In: AMIA .Annual Symposium proceedings, TBI. No. iEvoBio. p. 175.

Panahiazar, M., Ranabahu, A., Taslimi, V., Yalamanchili, H., Stoltzfus, A., 2012b. PhylOnt : A Domain-Specific Ontology for Phylogeny Analysis. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012).

Panahiazar, M., Sheth, A. P., Ranabahu, A., Leebens-Mack, J., 2012c. Semantic Technology and Translational Genomic Research. In: AMIA .Annual Symposium proceedings, TBI. p. 176.

Panahiazar, M., Vos, R. A., Enrico, P., Todd, V., Leebens-Mack, J., 2011. Building a Foun-

dation to Enable Semantic Technologies for phylogenetically based Comparitive Analysis. In: 2011 Informatics for Phylogenetics, Evolution, and Biodiversity (iEvoBio 2011).

Parikh, P. P., Minning, T., Nguyen, V., Lalithsena, S., Asiaee, A. H., Sahoo, S. S., Doshi, P., Tarleton, R., Sheth, A. P., Jan. 2012. A semantic problem solving environment for integrative parasite research: identification of intervention targets for Trypanosoma cruzi. PLoS neglected tropical diseases 6 (1), e1458.

Patwardhan, S., Banerjee, S., Pedersen, S., 2003. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the fourth international conference on intelligent text processing and computational linguistics. Mexico City, Mexico. pp. 241–57.

Piel, W. H., Donoghue, M., Sanderson, M., Person, C., 1997. TreeBASE : A database of phylogenetic information . Authors Studies Analyses Character States Matrices Trees Character Strings Taxa. Tech. rep., Institute of Evolutionary and Ecological Sciences, Leiden University 2311 GP Netherlands,.

Posada, D., Buckley, T. R., Oct. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Systematic biology 53 (5), 793–808.

Prosdocimi, F., Chisham, B., Pontelli, E., Thompson, J., Stoltzfus, A., 2009. Initial Implementation of a comparative Data Analysis Ontology. Evolutionary Bioinformatics, 47–66.

Ranabahu, A., Parikh, P., Panahiazar, M., Sheth, A., Logan-Klumpler, F., Sep. 2011a. Kino: A Generic Document Management System for Biologists Using SA-REST and Faceted SearcRanabahu, A., Parikh, P., Panahiazar, M., Sheth, A., & Logan-Klumpler, F. (2011). Kino: A Generic Document Management System for Biologists Using SA-REST and Faceted. 2011 IEEE Fifth International Conference on Semantic Computing, 205–208.

Ranabahu, A., Sheth, A., Panahiazar, M., Wijeratne, S., 2011b. Semantic Annotation and Search for resources in the next Generation Web with SA-REST SA-REST for Service Annotation. In: W3C Workshop on Data and Services Integration. Bedford, MA.

Resnik, P., 1999. Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Articial In telligence Research , 95–130.

Sánchez, D., Batet, M., Isern, D., Valls, A., Jul. 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications 39 (9), 7718–7728.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., Lewis, S., Nov. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 25 (11), 1251–5.

Soltis, D. E., Smith, S. a., Cellinese, N., Wurdack, K. J., Tank, D. C., Brockington, S. F., Refulio-Rodriguez, N. F., Walker, J. B., Moore, M. J., Carlsward, B. S., Bell, C. D., Latvis, M., Crawley, S., Black, C., Diouf, D., Xi, Z., Rushworth, C. a., Gitzendanner, M. a., Sytsma, K. J., Qiu, Y.-L., Hilu, K. W., Davis, C. C., Sanderson, M. J., Beaman, R. S., Olmstead, R. G., Judd, W. S., Donoghue, M. J., Soltis, P. S., Apr. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. American journal of botany 98 (4), 704–30.

Stevens, R., Goble, C. A., Bechhofer, S., 2000. Ontology-based Knowledge Representation for Bioinformatics Keywords : Ontology ; Knowledge ; Concept ; relationship ; knowledge. Tech. rep., Department of Computer Science and School of Biological Sciences, University of Manchester.

Uren, V., 2006. Semantic annotation for knowledge management: Requirements and a survey of the state of the art.

Van Auken, K., Jaffery, J., Chan, J., Müller, H.-M., Sternberg, P. W., Jan. 2009. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC bioinformatics 10, 228.

Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Madison, W. P., Midford, P. E., Priyam, A., Sukumaran, J., Xia, X., Stoltzfus, A., Jul. 2012. NeXML: Rich, Extensible, and Verifiable Representation of Comparative Data and Metadata. Systematic biology 61 (4), 675–89.

Vrandečić, D., York, S., 2007. How to Design Better Ontology Metrics. ESWC '07 Proceedings of the 4th European conference on The Semantic Web: Research and Applications, 311 – 325.

Wu, Z., Palmer, M., 1994. Verb semantics and lexical selection. In: In 32nd annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico: Association for Computational Linguistics. pp. 133 –138.

Zhang, S., Chang, Z., Li, Z., Huizi DuanMu, Z. L., Li, K., Liu, Y., Qiu, F., Xu, Y., 2012. Calculating phenotypic similarity between genes using hierarchical structure data based on semantic similarity. 2012 Elsevier.