Engineering Trustworthy Ontologies: Case Study of Protein Ontology

Farookh K. Hussain[1], Amandeep S. Sidhu[2], Member, IEEE, Tharam S. Dillon[2], Fellow, IEEE, Elizabeth Chang[2], Member, IEEE

[1]*School of Information Systems, Curtin University of Technology Perth, Australia*
{Farookh.Hussain, Elizabeth.Chang}@cbs.curtin.edu.au
[2]*Faculty of Information Technology, University of Technology Sydney, Australia*
(asidhu, tharam)@ it.uts.edu.au

Abstract

Biomedical Ontologies are huge. It is not possible for any one person to manage and engineer a complete ontology. They would need the help of Research Assistants and other people to develop and maintain the ontology. In the process of developing and maintaining the ontology the Research Assistants may enter incorrect data, resulting in low quality of the ontology. In this paper we will propose a conceptual framework to solve these ontology management and ontology development issues. There can be N assistants entering data into the ontology. All the data entered initially is stored in an intermediate ontology. The administrator of the ontology has a set of rules, which makes a checklist that checks and validates the data in intermediate ontology for correctness according to the ontology schema. We use the Case Study of Protein Ontology for this proposed approach to develop interfaces for assistants and administrators. The proposed approach can easily be extended to other biomedical ontologies just by tweaking the administrator rule set according to the ontology.

1. Introduction

In recent years, the notion of the Ontology has been gaining prominence. Ontology provides explicit formalization and conceptual specification of knowledge representation. The knowledge conceptualization is modeled in terms of notional entities and their inter-relationships. Ontology or simply a conceptual knowledge map is only meaningful when it is associated with semantic data instances. As Ontology Development becomes more geographically dispersed, inter-site communications and communication between various human and intelligent agents become a key issue that often leads to inconsistent and incorrect instances of ontology. In this paper we propose a *'Trustworthy Ontology Approach'* as a response to problems in multi-site distributed ontology development. The approach is termed as *'Trustworthy Ontology Approach'* because the final developed ontology would be accurate. In other words the user can trust the ontology to be accurate and precise that the ontology that would be developed with out this approach.

We demonstrate the proposed approach using the case study of Protein Ontology (PO) which is highlighted in. PO provides the vocabulary of terms for Proteomics data and the inter-relationships between those terms. PO provides a framework for seamless data integration between major protein data sources available for public use.

2. Protein Ontology Framework

For developing Protein Ontology (PO), we will mainly deal with two main sources of protein annotations: (1) those taken from various protein data sources submitted by authors of protein data themselves from their published experimental results and (2) those that we name annotation that are obtained by an annotator or group of annotators by analysis of raw data (typically a protein sequence or atomic structure description) with various tools extracting biological information from other protein data collections. PO provides integration of heterogenous protein and biological data sources. PO converts the enormous amounts of data collected by geneticists and molecular biologists into information that scientists, physicians and other health care professionals and researchers can use to easily understand the mapping of relationships inside protein molecules, interaction between two protein molecules and interactions between protein and other macromolecules at cellular level.

The process of development of a protein annotation based on our protein ontology requires an important effort to organize, standardize and rationalize protein data and concepts.

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 IEEE

1. First of all, protein information must be defined and organized in a systematic manner in databases. In this context, our protein ontology addresses the following problems of existing protein databases: redundancy, data quality (errors, incorrect annotations, and inconsistencies), lack of standardization in nomenclature etc.
2. Secondly, the process of annotation relies heavily on integration of heterogeneous protein data. Integration is thus a key concept if one wants to make full use of protein data from collections. In order to be able to integrate various protein data it is important that concepts underlying the data be agreed upon by community. PO provides a framework of structured vocabularies and standardized description of protein concepts that helps to achieve this agreement and achieve uniformity in protein data representation.

PO consists of concepts (or classes), which are data descriptors for proteomics data and the relations among these concepts. PO has (1) a hierarchical classification of concepts represented as classes, from general to specific; (2) a list of attributes related to each concept, for each class; and (3) a set of relations between classes to link concepts in ontology in more complicated ways then implied by the hierarchy, to promote reuse of concepts in the ontology. At the moment PO currently contains 92 concepts or classes and 261 attributes or properties. The structure of PO provides the concepts necessary to describe individual protein complexes, but does not contain individual protein themselves. The PO database acts as instance store for the PO. The attribute values in the PO are not defined as text strings or as set of keywords. Most of the Values are entered as instances of Concepts defined in Generic Classes. PO is the first ever work to integrate protein data based on data semantics describing various phases of protein structure. PO helps to understand structure, cellular function and the constraints that affect protein in a cellular environment.

3. PO Semantic Framework

3.1 Semantic Relationships

Semantics in protein data is normally not interpreted by annotating systems, since they are not aware of the specific structural, chemical and cellular interactions of protein complexes. Protein Ontology Framework provides specific set of rules to cover these application specific semantics. The rules use only the relationships whose semantics are predefined to establish correspondence among terms in PO. The set of relationships with predefined semantics is: {SubClassOf, PartOf, AttributeOf, InstanceOf, and ValueOf}. The PO conceptual modeling encourages the use of strictly typed relations with precisely defined semantics. Some of these relationships (like SubClassOf, InstanceOf) are somewhat similar to those in RDF Schema but the set of relationships that have defined semantics in our conceptual PO model is small so as to maintain simplicity of the system. The following is a description of the set of pre-defined semantic relationships in our common PO conceptual model.

SubClassOf: The relationship is used to indicate that one concept is a subclass of another concept, for instance: SourceCell SubClassOf FunctionalDomains. That is any instance of SouceCell class is also instance of FunctionalDomains class. All attributes of FunctionalDomains class (_FuncDomain_Family, _FuncDomain_SuperFamily) are also the attributes of SourceCell class. The relationship SubClassOf is transitive.

AttrributeOf: This relationship indicates that a concept is an attribute of another concept, for instance: _FuncDomain_Family AttributeOf Family. This relationship also referred as PropertyOf, has same semantics as in object-relational databases.

PartOf: This relationship indicates that a concept is a part of another concept, for instance: Chain PartOf ATOMSequence indicates that Chain describing various residue sequences in a protein is a part of definition of ATOMSequence for that protein.

InstanceOf: This relationship indicates that an object is an instance of the class, for instance: ATOMSequenceInstance_10 InstanceOf ATOMSequence indicates that ATOMSequenceInstance_10 is an instance of class ATOMSequence.

ValueOf: This relationship is used to indicate the value of an attribute of an object, for instance: “Homo Sapiens” ValueOf OrganismScientific. The second concept, in turn has an edge, OrganismScientific AttributeOf Molecule, from the object it describes.

3.2 Sequences

Apart from semantic relationships defined in Section 3.1, PO also model relationships like Sequences. By itself semantic relationships described in Section 3.1, does not impose order among the children of the node. In applications using Protein Sequences, the ability of expressing the order is paramount. Generally Protein Sequences are a collection of chains of sequence of residues, and that is the format Protein Sequences have been represented unit now using various data

representations and data mining techniques for bioinformatics. When we are defining sequences for semantic heterogeneity of protein data sources using PO we are not only considering traditional representation of protein sequences but also link Protein Sequences to Protein Structure, by linking chains of residue sequences to atoms defining three-dimensional structure. In this section we will describe how we used a special semantic relationship like *Sequence(s)* in Protein Ontology to describe complex concepts defining Structure, Structural Folds and Domains and Chemical Bonds describing Protein Complexes. PO defines these complex concepts as *Sequences* of simpler generic concepts defined in PO. These simple concepts are *Sequences* of object and data type properties defining them. A typical example of *Sequence* is as follows. PO defines a complex concept of *ATOMSequence* describing three dimensional structure of protein complex as a combination of simple concepts of *Chains*, *Residues*, and *Atoms* as: *ATOMSequence Sequence (Chains Sequence (Residues Sequence (Atoms)))*. Simple concepts defining ATOMSequence are defined as: *Chains Sequence (ChainID, ChainName, ChainProperty)*; *Residues Sequence (ResidueID, ResidueName, ResidueProperty)*; and *Atoms Sequence (AtomID, Atom, ATOMResSeqNum, X, Y, Z, Occupancy, TempratureFactor, Element)*.

In this paper we propose a method by which ontologies can be engineered. The next section provides the conceptual framework for our proposed method.

4. Conceptual Framework

Here we describe a conceptual framework that we are working on, to engineer Trustworthy Protein Ontology. It is termed as 'Trustworthy Protein Ontology' as the final engineered ontology is trustworthy in the sense that it is accurate and precise. The final engineered ontology does not contain any redundant, inconsistent, and incorrect data or relationships.

Consider the scenario where we have 'N' Research Assistants. Each of these Research Assistants enters the data into an Intermediate Protein Ontology (IPO). IPO is mirror of the Original PO and contains same concepts in an exactly similar structured hierarchy as PO. However the research assistants may not be necessarily the experts in field of proteomics for which the ontology is being engineered. Hence we propose that instead of allowing research assistants to make changes directly to the Original PO, changes should be entered into the IPO. PO administrator then goes through IPO to check if the concepts, relationships and instances entered by research assistants. PO administrator is a person who is an expert in the field of proteomics for which trustworthy PO is engineered. PO administrator has knowledge about data formats of diverse protein data and knowledge sources. After research assistants enter the data in IPO, PO administrator goes through IPO in order skim out concepts, relationships and instances which are redundant, inconsistent, and incorrect. This is done by running syntax and semantic checks on IPO, to check its validity in regards to concepts, relationships and instances already present in Original PO. There are two ways in which PO administrator may choose to skim through IPO.

Method 1: PO administrator goes through the whole IPO to which changes have been submitted by the Research Assistants to determine those concepts, relationships and instances which are redundant, inconsistent, and incorrect. PO administrator then removes or fixes these concepts, relationships and instances to create the final engineered IPO. Once all discrepancies have been removed from the final engineered IPO, and it has been checked for validity with the Original PO, all the changes made to IPO are integrated into the Original PO. This method compares structure and relationships of IPO and Original PO. This method is tedious and requires a lot of time and effort by the PO administrator. PO administrators can alternatively choose Method 2 as a means to engineer trustworthy ontology which is quick, effective and does all the checks.

Method 2: PO administrator uses an administration console to skim through IPO using a defined set of rules that denotes what a correct concept would be, what a correct relationship between those concepts would be and what a correct instance of the concept would be. These set of rules utilize structure and semantics of PO to facilitate validation of any changes made to IPO by research assistants. PO structured vocabulary briefly outlined in Section 2 has 92 pre-defined concepts that belong to set of valid concepts, **SET V**. Of these 92 concepts, 12 concepts are necessary to define the basic information to enter protein complex data into the PO framework. These mandatory concepts belong to **SET M**. SET M is a subset of SET V. Semantic Relationships among the concepts of PO framework are discussed in Section 3. These Semantic Relationships belong to set of valid relationships, **SET R**.

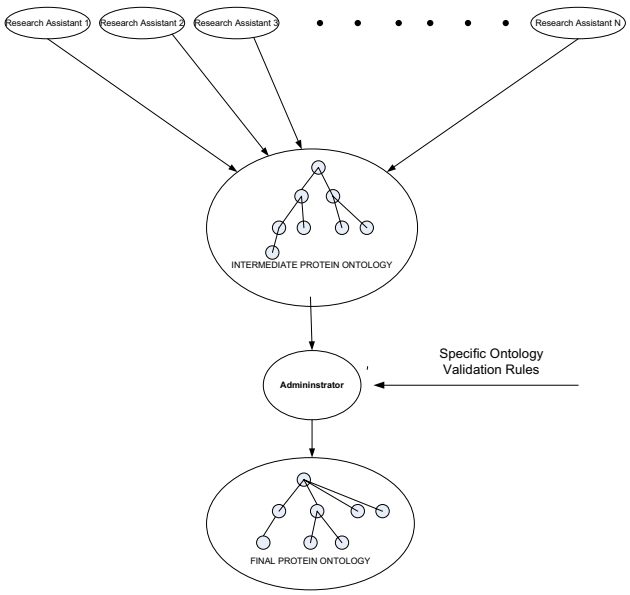

Figure 1: Conceptual Framework for Engineering Trustworthy Protein Ontology

To run structure and semantic checks using this method is followed:

1. For a concept entered in IPO by research assistants to be valid **(c)** it should be within the scope of SET V and must belong to SET M.

2. For a relationship entered in IPO by research assistants to be valid **(r)** it must belong to SET R.

3. Every tuple (c, r) in IPO belongs to a frameset F. These concepts and relationships are necessary and must be integrated with Original PO.

4. Every tuple ($c^{/}$, r) in IPO belongs to frameset $F^{/}$. Here $c^{/}$ is a concept that does not belong to SET M. These concepts are checked further to see if they belong to SET V. If they do belong to SET V, then the tuple ($c^{/}$, r) is valid and must be integrated with Original PO.

5. All the tuples that do not belong to F and $F^{/}$ are discarded.

Thus, Method 2 is much quicker and efficient way to engineer a trustworthy PO, but it adds to the complexity of the algorithm. The approach proposed here for generating Trustworthy Protein Ontology is currently being implemented to provide a non-redundant, accurate and precise PO framework for future.

5. Summary and Future Work

In this paper we discussed the process of developing and maintaining the ontology. As mentioned before the Research Assistants may enter incorrect data, resulting in low quality of the ontology. In this paper we proposed a conceptual framework to solve these ontology management and ontology development issues.

We have proposed two methods by which the administrator of the ontology (expert in the field in which the ontology is being engineered) can engineer trustworthy ontology.

Future work involves validating our approach practically. In the larger version of the paper we intend to lay out the rules that can be used by the Administrator of Protein Ontology in order to ensure the integrity of the protein ontology being engineered.

6. References

[1] **Sidhu, A. S.**, T. S. Dillon, et al. (2005). **Structured Vocabularies for Proteins.** The 12th International Conference on Biomedical Engineering **(ICBME 2005)**, Singapore, International Federation for Medical & Biological Engineering (IFMBE)..

[2] **Sidhu**, **A. S.**, T. S. Dillon, et al. (2005). **The Protein Ontology Project: Structured Vocabularies for Proteins.** 6th International Conference on Data Mining, Text Mining and their Business Applications **(Data Mining 2005)**. A. Zanasi, C. A. Brebbia and N. F. F. Ebecken. Skiathos, Greece., WIT Press.

[3] **Sidhu, A. S.**, T. S. Dillon, et al. (2006). *"Protein Ontology: Data Integration using Protein Ontology" in* **Database Modeling in Biology: Practices and Challenges.** Z. Ma and J. Y. Chen. New York, NY, Springer Science, Inc.: **In Press.**

[4] **Sidhu, A. S.**, T. S. Dillon, et al. (2006). *"Knowledge Discovery in Biomedical Data facilitated by Domain Ontologies" in* **Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data.** X. Zhu and I. Davidson. Idea Group Inc.: **In Press.**