BioMap: Discovering Schema Mapping Using Ontologies

Zoé Lacroix, Sumedha Gholba, Hervé Ménager
Scientific Data Management Lab, Arizona State University
{ *zoe.lacroix, sumedha.gholba, herve.menager*}*@asu.edu*

Abstract

Biologists currently devote significant time and effort searching information for their research. The wide diversity in terminology used inhibit effective computerized and manual data retrieval. For example, say a user wants to find all the gene products that are involved in bacterial protein synthesis, and that have sequences or structures significantly different from those in humans. If one database describes these molecules as being involved in 'translation', whereas another uses the phrase 'protein synthesis', it will be difficult for the user - and even harder for a computer - to find functionally equivalent terms. A schema mapping tool, which interprets results from one database in terminology used by a second database, can solve such problems. We started our project by developing schema mapping for UniProt[1] *and Genbank*[2] *protein resources, both of which can be rendered in XML format, as a large part of scientific community uses proteomic resources. The approach will later on be extended to other scientific databases. Here we present a novel idea of mapping schemas using ontologies.*

1. Introduction

The aim of a schema-mapping tool is to express the syntactic correspondence between semantically equivalent concepts from separate schemas. A schema mapping tool takes two schemas, as input, and generates a mapping (translation) from one schema to another as output. Traditional approaches follow syntactic and semantic methods [1],[2].

The mapping can be achieved in terms of database "structure" (tree, network, relation, etc. as well as attributes names, types) and "values" (actual instance values). Syntactic approaches match different object names (types), to detect similarity. This requires syntactic metadata knowledge, data structure (tree, network, relation, etc.) as well as attribute names and types. Semantic approaches, on the other hand, map schemas on the basis of the attribute values.

These techniques can only assist users in detecting correct mappings, but usually do not result in a completely automated schema mapping process. The actual mapping is done using local ideas (pertaining to two-three databases) however the aim is to achieve global mapping (mapping any database. To increase the relevance of the results, combination of different techniques (syntactic and semantic) is necessary [1].

2. BioMap

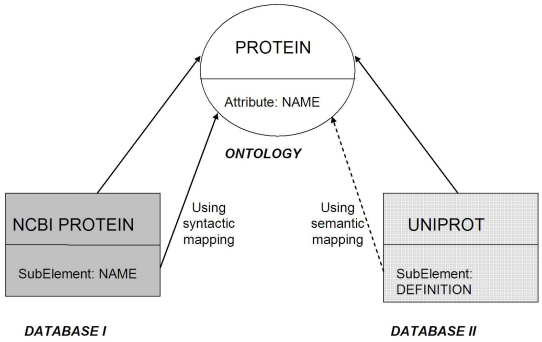
The main challenge of schema mapping is to map the information contained in the resources regardless of the intrinsic format in which they are presented. To address this challenge a reference can be used. This reference represents different concepts of the concerned domain such as Protein and forms a learning phase that lets users map the metadata (syntax or semantics) of one or a set of data sources to itself. The use of an ontology as such reference offers many advantages including:

- expressing the semantics of data instead of an alternative syntax.
- offering useful reasoning possibilities, that can be used in the automated detection process.

Ontologies help in creating a global schema mapping, thus allowing bi-directional mapping. It is better than traditional techniques which allow uni-directional mapping. Many different resources on the web can deliver information about a same scientific object, and it can be very useful to gather this information from different resources to get more complete results, as well as to detect potential inconsistencies. In this context, the use of an automated data mapping should allow, after having specified the correspondences between a set of initial data sources and the ontology, to detect these correspondences with other equivalent resources.

For analyzing how to map the two selected resources, a same protein, p53, was selected from both the resources and its attributes were studied. The front-end formats

[1] http://www.pir.uniprot.org/
[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein



(XML-UniProt and XML-GenBank) were considered for this analysis. After thoroughly understanding the structure of the XML files and the attributes associated with the entity 'protein', few common features were recognized. Using these common attributes, we formed our own ontology using the Protégé software [3]. We included in the ontology identifiers and important attributes for each object.

3. BioMap Ontology

An ontology was built, in order to accurately represent the different concepts expressed in both the protein resources, using guidelines from [3] and [4]. Not all concepts related to the proteomics were identified and were represented in this ontology, but we concentrated on selecting only a few important ones.

The BioMap ontology represents the following classes of scientific objects such as Citation, Sequence, NucleotideSequence, Codon, Protein, Organism, etc. Each class is related to another class. For example a protein *is a* sequence of AminoAcids resulting from the translation from a Gene. A Gene *is a* sequence of DNA nucleotides that *codes for* a specific product in the metabolism. A Gene *is a kind* of NucleotideSequence.

Each of these classes is characterized with a unique identifier, that has to be as widely used as possible by the scientific community. For example citations are identified by a *PubMed ID*. This identifier comes from NCBI's bibliographic database, PubMed.

4. Schema mapping toolkit

The semi-automated schema mapping process is based on the statements issued by different agents that analyze the data. In syntactic mapping, agents are based on the analysis of the syntax of the resources that have to be mapped. For instance, an agent could compare the names used for two structures to determine their degree of similarity (are the names of the two structures identical, or is one a substring of the other?). Semantic mapping agents analyze the instances. For example, by detecting similar values in two structural components (e.g., an attribute), the agent can infer they are equivalent. An ontology identifies semantic mapping, which is difficult to find in a syntactic schema and guides the mapping process. We follow this approach as it introduces a solution to map the first source to the ontology, and also the second source to the ontology, thereby allowing mapping either resource to the other (bi-directional).

For instance, as illustrated in Figure 1, we can identify that the two resources model a common entity - Protein. Next, we can identify the same value "Cellular tumor antigen p53" in a component's label "Name" in UniProt and as a part of the component value called "Definition" in NCBI Protein. If the ontology specifies a concept "Protein" which has an attribute called "Name", a syntactic mapping easily identifies "Name" as being the name attribute of the Protein concept. The mapping between the component "Definition" in UniProt and "Name" in the ontology is semantically devised, thus mapping the two resources through ontology. Similarly, the amino acid sequence of the protein is labeled "sequence" in UniProt, and "origin" in NCBI Protein. The syntactic mapping from UniProt to the ontology is therefore easy, and based on the value of the UniProt associated value; the semantic mapping identifies the NCBI Protein structure.

5. Conclusions

BioMap can be used for integrating databases by translating one resource schema into another and provides interoperability between applications. It is not only useful for generating but also maintaining adapters, parsers and wrappers.

In future, we plan to consider other formats (nested text, relational, HTML, etc.) and resources other than proteomics.

Acknowledgement: This project is partially supported by the NSF grant IIS-0223042.

References

[1] A. Doan, P. Domingos, and A. Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. 2001.

[2] L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernandez, and R. Fagin. Translating web data. pages 598–609, 2002.

[3] H. Knublauch. An AI tool for the real world - Knowledge modeling with Protégé, Jun 2003.

[4] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. (KSL-01-05), Mar 2001.