Mahoui, Malika, Kulkarni, Harshad, Li, Nianhua, Ben-Miled, Zina and Börner, Katy.
(2005) Semantic Correspondence in Federated Life Science Data Integration Systems.
In Bertram Ludäscher & Louiqa Raschid (eds.) Data Integration in the Life Sciences,
Lecture Notes in Computer Science, Vol. 3615, Springer-Verlag, pp. 137-144.

Semantic Correspondence in Federated Life Science Data Integration Systems

Malika Mahoui[1], Harshad Kulkarni[2], Nianhua Li[2],
Zina Ben-Miled[2], and Katy Börner[3]

[1] School of Informatics, IUPUI, Walker Plaza, 719 Indiana Avenue, suite WK307,
Indianapolis, IN 46202, USA
1-317-278-9205
mmahoui@iupui.edu
[2] Department of Electrical and Computer Engineering, IUPUI
{hkulkarn, niali, zmiled}@iupui.edu
[3] School of Library and Information Science, Indiana University
katy@indiana.edu

Abstract. For execution of complex biological queries, data integration systems often use several intermediate data sources because the domain coverage of individual sources is limited. Quality of intermediate sources differs greatly based on the method used for curation, frequency of updates and breadth of domain coverage, which affects the quality of the results. Therefore, integration systems should provide data provenance; i.e. information about the path used to obtain every record in the result. Furthermore, since query capabilities of web-accessible sources are limited, integration systems need to support refinement queries of finer granularity issued over the integrated data. However, unlike the individual sources, integration systems have to handle the absence of data and conflicts in the integrated data caused by inconsistencies among the sources. This paper describes the solution proposed by BACIIS, the Biological and Chemical Information Integration System, for providing data provenance and for supporting refinement queries over integrated data. Semantic correspondence between records from different sources is defined based on the links connecting these data sources including cross-references. Two characteristics of semantic correspondence, namely degree and cardinality, are identified based on the closeness of the links that exist between data records and based on the mappings between domains of data records respectively. An algorithm based on semantic correspondence is presented to handle absence of data and conflicts in the integrated data.

1 Introduction

The rapid development of experimental biology has led to the emergence of a large number of web-accessible biological data sources [1]. Together, these data sources cover a wide range of subjects and data types. But each individual data source often focuses on a specific subject area; and thus represents only a fraction of all the

B. Ludäscher and L. Raschid (Eds.): DILS 2005, LNBI 3615, pp. 137 – 144, 2005.

Mahoui, Malika, Kulkarni, Harshad, Li, Nianhua, Ben-Miled, Zina and Börner, Katy.
(2005) Semantic Correspondence in Federated Life Science Data Integration Systems.
In Bertram Ludäscher & Louiqa Raschid (eds.) Data Integration in the Life Sciences,
Lecture Notes in Computer Science, Vol. 3615, Springer-Verlag, pp. 137-144.

available data. Cross-references are provided between different data sources to connect data into a network. In addition of cross-references, integration systems [2, 3] use field values in records produced by one data source to connect to another data source to address complex queries. So a query plan can contain chains of links connecting several sources. One important issue is data quality control. Due to factors such as the method of annotation, update frequency and overall coverage of the subject area, not all of these sources are trusted equally. The quality of a data source will affect both the quality of data fields and the quality of cross-references. It is therefore important to specify the source of every piece of result data, and to aggregate records with cross-references to each other.

Query capabilities of web-accessible data sources are limited and it is often not possible to use every characteristic of biological entities in the query predicate. Therefore, most initial queries are not specific enough and their results contain several unwanted records. In such cases, once the integrated result of an initial query is available, the integration system can allow scientists to issue refining queries. However, unlike the initial query, while processing a refining query, complex inter-relationships among records from different sources must be considered. Most integration systems assume that different sources cover different characteristics of the biological entities and hence, do not deal with absence of data or contradictory data [4]. However, this does not represent the true nature of relationships among records and consequently, the results of such systems are not complete and reliable. In reality different sources have significant overlap of information and data inconsistencies are present in the overlapping portions due to different methods of curating the data. Therefore, to process refining queries in a comprehensive and correct manner, we must assume an overlapping coverage of the global schema by different sources and deal with the data absence and inconsistency.

The objective of this paper is to describe the solution proposed by BACIIS, the Biological and Chemical Information Integration System [5-7], for providing data provenance for result records and for supporting queries over integrated results. Section 2 briefly introduces BACIIS system and its main data integration features. Section 3 defines the concept of semantic correspondence and its characteristics. In section 4, processing of refinement queries over the integrated data is discussed and an algorithm is presented to handle the conflicts in the integrated data.

2 BACIIS: An Ontology Augmented Database Integration System

BACIIS (Biological and Chemical Information Integration System) is a highly coupled federation of life science web-databases. It uses a mediator-wrapper approach, augmented with a knowledge base. The wrapper extracts information from a given remote data source. The mediator transforms data from its format in the source database to the internal format used by the integration system. The BACIIS knowledge base has two components: the ontology and data source schema. The ontology provides a method for mapping differences in terminology to a common term that is recognized throughout the domain. In addition to syntactic reconciliation, the ontology is used for semantic reconciliation as well as a global schema in BACIIS. Global queries are built by using concepts from the ontology. These global

queries are decomposed within BACIIS into database specific sub-queries. The query planner in BACIIS [7] identifies the data sources that can answer the sub-queries based on the description of the data sources that is included in database specific data source schema. The results of the graph planner is a graph where nodes represent data sources and an edge is present between two nodes if a link can be established between the corresponding data sources (see section 3.1). Finally, each database is associated with a specific wrapper and these wrappers are responsible for executing the sub-queries on the web databases and retrieving the result.

3 Semantic Correspondences Among Heterogeneous Data

In section 3.1, the concept of semantic correspondence is explored in the context of integration of web-accessible life science data sources. Two characteristics of semantic correspondence, degree and cardinality, are then introduced in section 3.2 and 3.3. Degree is a measure of how closely two data records from different databases correspond with each other. Cardinality is a measure of domain mapping between two real world objects with some semantic correspondence.

3.1 Concept of Semantic Correspondence (SC)

The issue of semantic correspondence between two objects that have significant representational differences was examined in [8]. It also provides a way to distinguish between different degrees of semantic correspondence using factors like the context, abstraction, domains and the state of objects. However, in the context of integrating domain specific data from autonomous, heterogeneous and semi-structured sources, we maintain that the SC is established between two records when field values of one record can be used to identify the other record. Sometimes, the link between records is explicitly given by the data sources. For example, SwissProt records provide hyperlinks to related records in PDB. This is similar to the concept of hyperlink authority explored in [9]. However hyperlinks are not the only way to establish SC. Consider the case of BIND [10], which does not have explicit hyperlinks to SwissProt records. However, BIND records contain attributes ‘protein-name’ and ‘organism-name’, which can be combined to identify a protein sequence record from SwissProt. The roles of ‘protein-name’ and ‘organism-name’ here are similar to the role of foreign keys in relational databases.

The idea of SC can be illustrated by an example query: “Which protein family does chaperonin hsp60 precursor in Arabidopsis thaliana belong to? What is its coding gene sequence? What are the 3D structures of proteins that belong to the same family?” The predicate of this query has two constraints (i.e., Protein Name = chaperonin HSP60 precursor and Organism Name = Arabidopsis thaliana), and the output requires four characteristics (protein family, coding gene sequence, and 3D structure). To the best of our knowledge, no individual life science data source can answer the above query directly due to limited query capabilities and domain coverage [7]. Information from multiple data sources has to be combined together for a complete answer. Figure 1 shows one possible query plan and some results for illustration purpose. The predicate criteria ‘protein-name’ and ‘organism-name’ are

Mahoui, Malika, Kulkarni, Harshad, Li, Nianhua, Ben-Miled, Zina and Börner, Katy.
(2005) Semantic Correspondence in Federated Life Science Data Integration Systems.
In Bertram Ludäscher & Louiqa Raschid (eds.) Data Integration in the Life Sciences,
Lecture Notes in Computer Science, Vol. 3615, Springer-Verlag, pp. 137-144.

combined and submitted to SwissProt, which provides the sequence-info part of the output. The SwissProt data record matching the protein name, also provide hyperlinks to related data records in GenBank and PROSITE. These sources provide the gene sequence information and the pattern description part of the output, respectively. Finally, PROSITE data records provide hyperlinks to the related PDB data records, which contain the 3-D structure part of the output. Thus, the result of this query consists of data records that are obtained from four different sources.

Consider the SC between SwissProt record P29197 and the PROSITE record PS00296 in figure 1. This semantic correspondence is established because the SwissProt record has a hyperlink to the PROSITE record. In terms of domain knowledge, this SC denotes the fact that the protein represented by the sequence in SwissProt record belongs to the family represented by the PROSITE record. Similarly, the SC between SwissProt and GenBank records denotes the domain knowledge that the protein represented by the SwissProt record is a product of the gene represented by the GenBank record.

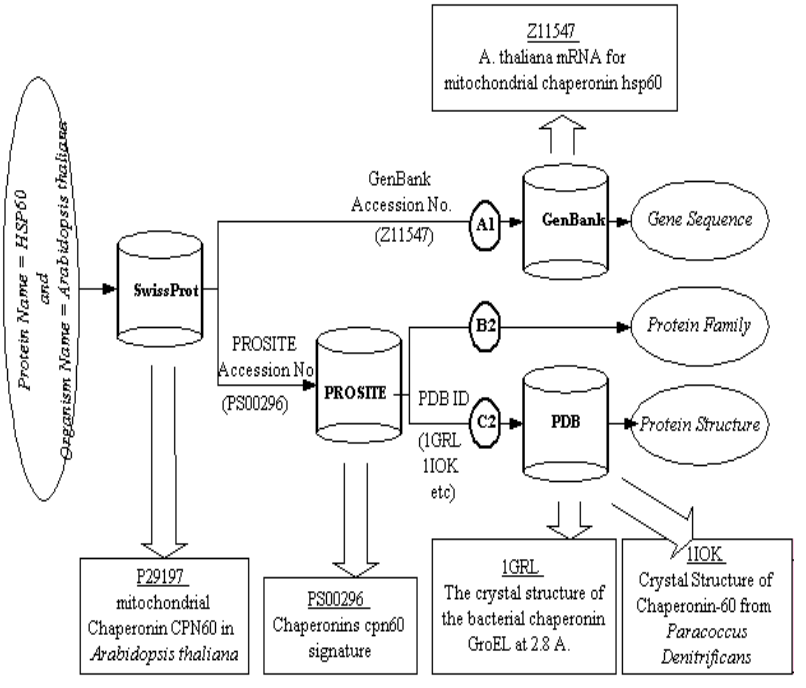

Fig. 1. Partial result of a query

3.2 Degree of Semantic Correspondence

Now consider the PROSITE and GENBANK records in figure 1. Do they have SC among them? In terms of domain knowledge, the protein family represented by PROSITE record and the gene represented by the GENBANK record, both are

definitely related to the sequence represented by the SwissProt record, i.e., these records represent different characteristics of the same protein. Therefore, GENBANK and PROSITE records do have certain SC. However, this SC is not as strong as the one that links GENBANK and SwissProt records; because there are neither direct hyperlinks nor matching field values between the GENBANK and PROSITE records. Formelly, we define two degrees of SC: strong SC and weak SC.

Strong SC (SSC): Two data records are said to have strong SC, if they are linked directly either by matching field values or by hyperlinks. These data records are immediate neighbors in the query plan. For example, the SwissProt and PROSITE records mentioned in the example above, have strong SC, as do SwissProt and GenBank records.

Weak SC (WSC): Two data records are said to have weak SC, if they are connected using a chain of SSC that travels through at least one other data source. These data records are connected but not immediate neighbors in the query plan. Records connected by WSC may represent different characteristics of the same biological entity. However, WSC is just a possibility and its validity must be confirmed using some other means as explained in the next section.

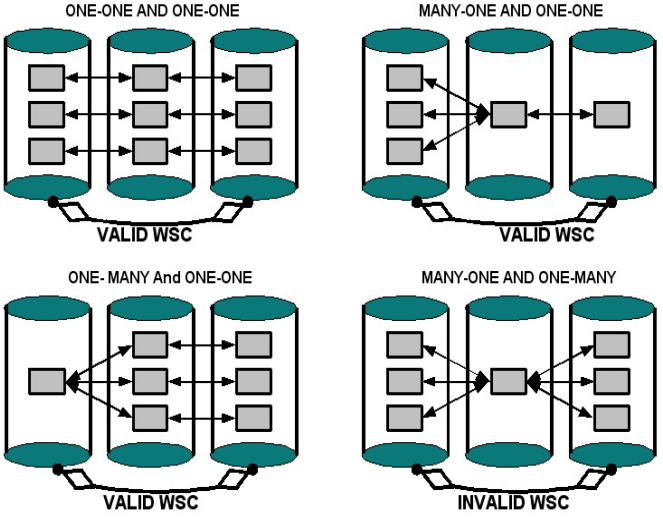

Fig. 2. Domain mapping and validity of WSC

3.3 Cardinality of Semantic Correspondence

According to the above definitions, SwissProt record P29197 and PDB record 1GRL in figure 1 are connected by WSC because both of them have SSC with PROSITE record PS00296. However it is misleading to connect them (P29197 and 1GRL) together because they represent two different proteins. On the other hand, the WSC between GenBank record Z11547 and PROSITE record PS00296 makes more sense

Mahoui, Malika, Kulkarni, Harshad, Li, Nianhua, Ben-Miled, Zina and Börner, Katy.
(2005) Semantic Correspondence in Federated Life Science Data Integration Systems.
In Bertram Ludäscher & Louiqa Raschid (eds.) Data Integration in the Life Sciences,
Lecture Notes in Computer Science, Vol. 3615, Springer-Verlag, pp. 137-144.

because the gene record and the protein family record are both characteristics of the protein represented by the connecting SwissProt record. In other words, the WSC between SwissProt and PDB records is invalid while the WSC between PROSITE and GENBANK is valid. The validity of WSC between two records thus depends on whether or not we can biologically pair them with each other.

This can be determined from the mapping of domains of the biological entities involved. For example, each protein record will have a corresponding Gene record that it can be biologically paired with; however, each protein family record can be paired with several corresponding protein records. Figure 2 shows the possible cases of domain mappings and corresponding validity or invalidity of the WSC. The mapping between the domain of intermediate source and its neighbors is the most important factor in deciding validity of WSC. If the intermediate record maps to multiple records with both its neighbors, its SC has a plural cardinality; otherwise it has singular cardinality. When an intermediate record has plural cardinality, we cannot reliably pair its neighbors and the WSC between them is labeled invalid.

4 Refinement Query Processing over Integrated Data

Query processing capabilities of web-accessible data sources are limited and not every field in the record can be used in the predicate. For example, it is not possible to use the field 'induction' as predicate in the initial query. Therefore, there will be many records in the result of the initial query with values of 'induction' different than the desired value. However, since BACIIS now has all the data locally, it can apply the additional criterion to that data regardless of the sources' capabilities. In general, BACIIS can provide refinement query capabilities of arbitrary granularity over the global schema and process those queries over the integrated data. For example, refinement query 'induction=heat shock' will only keep those records that contain 'heat shock' in field 'induction', and their related records.

Given the rich population of biological databases available online, it is not surprising that some portions of the domain be covered by multiple sources. For example, protein sequence information is available from several sources such as SwissProt, PIR, etc. Since BACIIS collects information from multiple data sources, it may get multiple values about the same data field from records of different sources. Those values may be inconsistent, but it is impossible to eliminate the wrong ones automatically. So, BACIIS will present all the data to users by default. If a user wants to further refine the result based on the value of one data field, inconsistent field values may cause a problem.

Consider the following query issued to BACIIS "What is the GENE ontology classification of protein featured for the protein phytochrome B in Arabidopsis thaliana?" Along with many others, the result for this query contains the following three records: At2g18790 from TIGR, NF00659007 from iProClass and 1005515 from TAIR. Figure 5 shows the cross-references among these records and using those along with domain mapping information; we can state that there is a valid WSC between the iProClass record and the TAIR record.

Now, consider the following refinement query issued on this result: "cellular component = membrane". From the data source schema, BACIIS finds out that only

TIGR and iProClass records can be directly evaluated for this predicate. Therefore, TAIR record's selection is completely dependent on its having valid semantic correspondence with a record that can be evaluated. However, the value of this field is different in both of such records, where iProClass record satisfies the predicate and TIGR does not. This inconsistency of data can be attributed to the different methods of annotation employed by the two sources. Nevertheless, the TAIR record now corresponds to one record that satisfies the predicate and another that doesn't. However, BACIIS has to take into account the relative degree of SC between these records and since the TAIR record has a SSC with the mismatching TIGR record, its WSC with iProClass record should be considered invalid and it should not be included in the result.

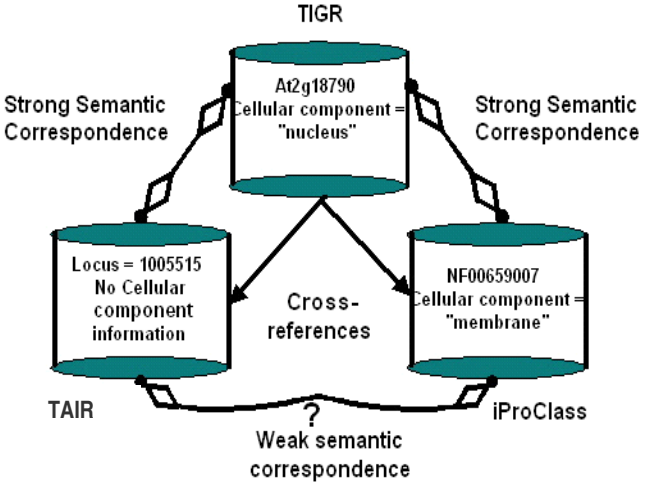

Fig. 3. Invalid WSC due to Data Inconsistency

To solve this problem, we propose an algorithm that first finds out all the records that can be directly evaluated for the predicate, and marks them as either valid or invalid. And then, the algorithm uses semantic correspondence to evaluate other records. For each record that has not been marked yet, if its predecessor is not marked invalid and if any of its neighbors are marked valid, then that record itself becomes valid. The rest of the records are invalid. Thus, the validity of a record for the refinement query is based on its being part of an unbroken chain of records in a path expression. Therefore, records with SSC to invalid records are eliminated.

5 Conclusion

In this paper two challenges were addressed; providing provenance for records in integrated data and processing queries over integrated data in a semantically meaningful way. The concept of semantic correspondence was introduced for

Mahoui, Malika, Kulkarni, Harshad, Li, Nianhua, Ben-Miled, Zina and Börner, Katy.
(2005) Semantic Correspondence in Federated Life Science Data Integration Systems.
In Bertram Ludäscher & Louiqa Raschid (eds.) Data Integration in the Life Sciences,
Lecture Notes in Computer Science, Vol. 3615, Springer-Verlag, pp. 137-144.

heterogeneous data obtained using several query paths. Two characteristics of semantic correspondence were also defined. First, the degree of semantic correspondence which represents the closeness of entities represented by different records and second, the cardinality which represents the mapping between domains of entities.

Data quality in biological data sources varies greatly based on several factors. Therefore, integrating data from overlapping data sources may generate results with missing data items or results that contain inconsistencies. The algorithm provided in this paper deals with these conflicts based on the characteristics of semantic correspondence among the records. It makes no assumption about the correctness of any data source involved. Furthermore, by removing semantically distant records from the integrated data, it achieves a better consistency for the integrated results.

References

[1] A. D. Baxevanis, The Molecular Biology Database Collection: 2003 update, Nucleic Acids Res., 31(1): 1-12, 2003.

[2] E. M. Zdobnov, R. Lopez, R. Apweiler, and T. Etzold, "The EBI SRS server-recent developments", Bioinformatics, 18(2): 368-73, 2002.

[3] C. A. Goble, R. Stevens, G. Ng, S. Bechhofer, N. W. Paton, P. G. Baker, M. Peim, and A. Brass, "Transparent access to multiple bioinformatics information sources", IBM Systems Journal, 40(2): 532-552, 2001.

[4] T. Hernandez, and S. Kambhampati, "Integration of Biological Sources: Current Systems and Challenges Ahead", To appear in SIGMOD Record, Vol 33, No3, September 2004.

[5] Z. Ben Miled, O. Bukhres, Y. Wang, N. Li, M. Baumgartner, and B. Sipes, "Biological and Chemical Information Integration System", Network Tools and Applications in Biology, Genoa, Italy, May 2001

[6] Z. Ben Miled, Y. Webster, N. Li, and Y. Liu, "An Ontology for the Semantic Integration of Life Science Web Databases," International Journal of Cooperative Information Systems, Vol. 12, N0.2, 2003

[7] Z. Ben-Miled, N. Li, G. Kellett, B. Sipes, and O. Bukhres, "Complex Life Science Multidatabase Queries", Proceedings of the IEEE, 90(11), 2002.

[8] A. Sheth, and V. Kashyap, "So Far (Schematically) yet So Close (Semantically)", Proceedings of the MT DS-5 Conference on Semantics of Interoperable Database Systems, Lorne, Australia, Elsvier Publishers, November 1992;

[9] J. Kleinberg., "Authoritative sources in a hyperlinked environment", Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997

[10] [http://www.blueprint.org/bind/bind.php]