Integration of Biological Data and Quality-Driven Source Negotiation

Laure Berti-Equille

IRISA, Campus Universitaire de Beaulieu,
35042 Rennes cedex, France
berti@irisa.fr
http://www.irisa.fr

Abstract. Evaluation of data non-quality in database or datawarehouse systems is a preliminary stage before any data usage and analysis, moreover in the context of data integration where several sources provide more or less redundant or contradictory information items and whose quality is often unknown, imprecise and very heterogeneous. Our application domain is bioinformatics where more than five hundred of semi-structured databanks propose biological information without any quality information (i.e. metadata and statistics describing the production and the management of the biological data). In order to facilitate the multi-source data integration in the context of distributed biological databanks, we propose a technique based on the concepts of quality contract and data source negotiation for a standard wrapper-mediator architecture. A quality source contract allows to specify quality dimensions necessary to the mediator for data extraction among several distributed resources. The source selection is dynamically computed with the contract negotiation which we propose to include into the mediation and the global query processings before data acquisition. The integration of the multi-source biological data is differed for the restitution and combination of the results of the global user's query by techniques of data recommendation taking into account source quality requirements.

1 Introduction

Maintaining a certain level of quality of data and data sources is challenging in distributed multiple source environments. In practice, assessing data quality in database systems is mainly conducted by professional assessors with more and more cost-competitive auditing practices. Well-known approaches from industrial quality management and software quality assessment have been adapted for data quality and came up with an extension of metadata management [20, 13, 26, 27, 25]. Classically, the database literature refers to data quality management as ensuring : 1) syntactic correctness (e.g. constraints enforcement, that prevent "garbage data" from being entered into the database) and 2) semantic correctness (i.e. data in the database truthfully reflect the real world situation). This traditional approach of data quality management has lead to techniques such as

H.S. Kunii, S. Jajodia, and A. Sølvberg (Eds.): ER 2001, LNCS 2224, pp. 256–269, 2001.

integrity constraints, concurrency control and schema integration for distributed and heterogeneous systems. Techniques such as data tracking, data cleaning and data quality controlling are costly in practice and difficult to adapt efficiently for specific application domains. In the multi-source context, as recent studies show, applications built on top of data warehouses often experience several problems due to the reliability and the quality of integrated data [12]. The main reason is that the local databases participating in providing their data contain incorrect, inaccurate, outdated or poor quality data. The quality of integrated data then becomes even worse unless suitable methods and techniques are employed during the multi-source environment design. Despite the amount of work focuses on semantic heterogeneity among data and metadata [11], the quality of integrated data has been addressed by very few research projects focused on the issues of multi-source data quality control and on the management of enriched metadata [17, 4]. As a matter of fact, data quality mainly has been and still is an important research topic independent of database integration. In the biological databanks context, more than five hundreds of databanks store potentially redundant and erroneous information. Here, the question of data cleaning and data mediation is crucial but has not been addressed so far. The richness and the complexity of biological concepts make also very difficult the implementation of a classical mediation system based on a common semi-structured description model of the distributed resources. Moreover, biological databanks are intensively cross-referenced and dependent, they cover several common domains and adopt very close and inter-dependent description models.

2 Related Works

2.1 Mediation and Cleaning of Multi-source Data

We distinguish two main research approaches concerning mediation systems [30].

The first one is a query based approach relying on a common semi-structured data model which represents data coming from heterogeneous sources, and queries expressed in a common query language (TSIMMIS with the model OEM and the language LOREL [2, 19], Strudel [6], YAT [5]). To handle the structure and the semantic discrepancies of the resources, appropriate integration views are defined over the data sources with special logic-based object-oriented languages for wrapper and mediator specification.

The second approach is a domain-model based approach. It relies on a common domain model described in the mediator level. This domain model (or metadata schema) captures the basic vocabulary used for the description of information expressed in local databases (Information Manifold [18, 14], SIMS with the language LOOM [3], Context Interchange [8], Garlic [1]). Mediators are used to resolve semantic conflicts among independent information sources. Knowledge necessary for this resolution is stored in the form of shared ontologies. In the domain-model base approach, the condition of integration for a new information

source is to provide an exhaustive description of its structure in terms of the domain model. The independence of resource-to-resource and mediator-to-resource description models is essential [11].

Data cleaning objectives are to detect matching records from several input extensional data structures (relational tables, object classes, DTDs), to find out and eliminate duplicate records in the integration process. As mentioned in [7], the main drawback of data cleaning methods is that, besides being a knowledge and time intensive task (since it implies several passes on the data), finding the suitable key for putting together similar records is very difficult. As another drawback, the support for cleaning rules offered in [16] allows matching rules to be applied only to pairs of neighbour records in the same file. Although, the problem of the possible conflict in data values was recognized, few specific solutions were offered [22]. Another problem arising from the proliferation of independent sources, some of them with overlapping information, is the inconsistency of information content, and hence there is a need for methods to resolve such inconsistencies in global answers. Inconsistencies result in multiple candidate answers ; the dual problem also exists, that is a global query might have no answer at all. Three usual approaches for reconciling the heterogeneities in data values are (1) to prefer the values from a more reliable database, (2) attach tags with data source identifications to data items and rely on the reputation of the data source [28], (3) store reliability measures of data sources from which a data item originated along with the data item itself [22]. These approaches also suffer from several drawbacks. First, it is not clear how to determine which of the sources is more reliable and how to measure reliability of a source. Second, even if the reliability of the data sources is somehow provided, it is implicitly assumed that the reliability remains the same for all data items from a particular data source. However, the reliability of the data items may be significantly different in the different parts of the source. And third, storing the reliability information or the data source tags along with the data items requires significant modifications in the conventional query processing mechanisms and increases data storage requirements.

2.2 Data Quality and Meta-data Management

There are a number of research investigating issues related to models and methodologies for data quality improvement [29, 27, 25], specifications and metrics for data quality dimensions [13]. Currently, data quality audits are the only practical means for determining quality of data in databases by using the appropriate statistical techniques. Since databases model a portion of the real world which constantly evolves, the data quality estimates become outdated as time passes. Therefore, the estimation process should be repeated periodically depending on the dynamics of the real world. The statistical aspects of data quality have been the primary focus with statistical methods of imputation (i.e., inferring missing data from statistical patterns of available data), predicting accuracy of the estimates based on the given data, data edits (automating detection and handling of

outliers in data). The use of machine learning techniques for data validation and correction is considered in [23]. For example, [24] describes a prototype system for checking correctness of the existing data (called data validation and cleanup). Utilization of statistical techniques for improving correctness of databases and introduction of a new kind of integrity constraints were proposed in [10]. The constraints are derived from a database instance using the conventional statistical techniques (e.g., sampling and regression), and every update of the database is validated against these constraints. If an update does not comply with them, then a user is alerted and prompted to check correctness of the update. Despite the growing importance of data quality for end-users and that many techniques have been proposed to improve and maintain quality of local databases, very few projects try to use quality metadata for multivalued attributes in distributed and quality-heterogeneous environments (DWQ [4], HiQiQ [17]). The use of metadata for data quality evaluation and improvement was advocated in [21] where information producers are encouraged to perform Verification, Validation, and Certification (VV&C) of their data. The metadata help in the process of estimating and maintaining the quality of data. Across different applications domains (such as geographic information systems [9] or digital libraries), a great amount of effort has been invested in the development of metadata standard vocabularies for the exchange of information. For the biological data domain, we are not aware of any kind of project (even among the current standardization works) that tries to specify metadata and to control biological data quality.

3 Example of Biological Information Retrieval from Distributed Databanks

Searching across distributed, disparate biological databases is increasingly difficult and time-consuming for biomedical researchers. Bioinformatics is coming to the forefront to address the problem of drawing effectively and efficiently information from a growing collection of 511 multiple and distributed databanks [1]. For example, suppose a biological researcher working with a gene and wanting to know what gene it is and its DNA sequence, whether the rRNA sequence is known, how the gene is transcribed into mRNA, how the mRNA is translated into that protein, what the protein function is, its cellular location... With the currently available biological banks and tools the researcher has to search the relevant databases one by one and then to locate the information items of interest within the return results.

Our mediator is designed to provide information about genes. The mediator conceptual model is presented in Figure 1. The mediator of our example will query three existing sources S_1 (EMBL[2]), S_2 (GenBank[3]) and S_3 (SWISS-

[1] see the Public Catalog of Databases : DBcat, http://www.infobiogen.fr/services/dbcat

[2] EMBL (European Molecular Biology Laboratory), http://www.ebi.ac.uk/embl/

[3] GenBank, http://www.ncbi.nlm.nih.gov/

PROT[4]) for retrieving information on Gene HFE_Human (related to the hemochromatose, a gene hepatic pathology). The raw values are automatically extracted by scripts using the DTD of each source. Four records with their respective Accession Numbers (Acc. Nb.) are extracted for the same gene in EMBL and Genbank and one record is extracted from SWISS-PROT concerning the protein related to gene HFE_Human (see Table 1). The sequence types and sizes are different according to the sources. They correspond to different submitted records with more or less detailed or complete annotations and four different submission dates. Our objective is to systematically extract and aggregate the

Source	Acc. Nb	Seq. Type	Seq. Size	Links for Protein	Date	Annotation
EMBL (S_1)	Z92910	Complete	12146	Genbank CAB07442	Mar. 97	Complete annotation
GenBank (S_2)	AF204869	Partial	3043		Nov.99	No annotation but a relevant information item
GenBank	AF184234	Partial	772	Genbank AAF01222	Sept.99	Detailed but incomplete annotation
SWISS-PROT(S_3)	Q30201	Complete	348		Oct.2000	Complete annotation

Table 1. Example Results for biological information retrieval on gene HFE_Human

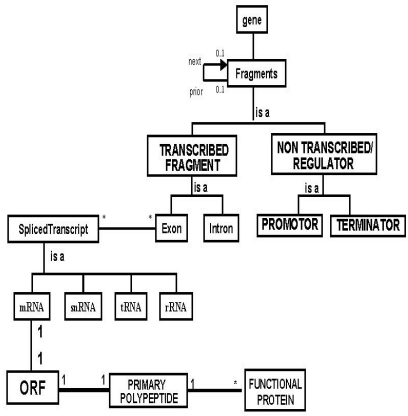
most exhaustive information on the expressed genes for instanciating and completing the gene sequence model of our mediator with considering the aspects of data and source quality for the integration process.

In order to define an appropriate data aggregation of all the available information items and rules for mapping the sources' records, obviously data conflicts have to be resolved due to the different values recorded for the same concept of gene HFE_Human. Traditional data integration approaches suggest a conflict resolution method that either chooses one value over the other (or that computes the average in case of the numerical values). A global query for the gene sequence would retrieve one data value from the source according to the specified data integration rule. For our example, the most complete sequence would be retrieved from S_1 skipping a relevant information item proposed by S_2 in the most recent record. Note that the same source (S_2) may propose several records for the same concept. Now assume the following scenarii :

Scenario 1 : S_1 updates their data every night, S_2 monthly and S_3 twice a month. In this case, the global query time may determine from which data source the most up-to-date data are retrieved.

Scenario 2 : S_3 is the server of an institute which does its own sequencing for human species. Sequence data are highly accurate for this species. S_2 data may have sometimes parsing errors and may come from other sites.

[4] SWISS-PROT, http://www.expasy.ch/sprot/sprot-top.html




```
<! DOCTYPE gene [
  <!ELEMENT gene     (name , description, fragment*)>
       <!ATTLIST AC IDREFS #IMPLIED SOURCE IDREFS #IMPLIED>
  <!ELEMENT fragment (position? , (nontranscribed_fragment | transcribed_fragment))>
  <!ELEMENT nontranscribed_fragment (#PCDATA ? ,( promotor | terminator))>
  <!ELEMENT promotor     (position , sequence)>
  <!ELEMENT terminator   (position , sequence)>
  <!ELEMENT transcribed_fragment (intron | exon)>
  <!ELEMENT intron       (position)> <!ATTLIST number  #REQUIRED >
  <!ELEMENT exon         (position)> <!ATTLIST number  #REQUIRED >
  <!ELEMENT spliced_transcript ((exon)*, ( mRNA | snRNA | tRNA | rRNA))>
  <!ELEMENT mRNA         (#PCDATA)>
       <!ATTLIST mRNA ORF IDREF #IMPLIED>
  <!ELEMENT ORF          (#PCDATA)>
       <!ATTLIST ORF primary_polypeptide IDREF #IMPLIED>
  <!ELEMENT primary_polypeptide (#PCDATA)>
       <!ATTLIST primary_polypeptide  protein IDREFS #IMPLIED>
  <!ELEMENT protein      (name , description, function)>
        <!ATTLIST AC IDREFS #IMPLIED SOURCE IDREFS #IMPLIED>
  <!ELEMENT snRNA        (#PCDATA)>
  <!ELEMENT tRNA         (#PCDATA)>
  <!ELEMENT rRNA         (#PCDATA)>  ]>
```

Fig. 1. The Gene Sequence Model and the corresponding DTD

Scenario 3 : S_1 and S_2 cover more biological domains than S_3. S_1 is one of the main genetic databanks and S_3 is one of the main databanks on proteins. Information items of S_2 are usually less complete and accurate than those of S_3 for the protein domain.

The above scenarii briefly show that the way of how and when data is populated into the local sources plays an important role for integrating local data for global queries. They also describe source dependencies and data quality dimensions such as freshness, accuracy, completeness or coverage. Up-to-date data does neither imply most accurate data nor most complete data. Actually, a global user might be interested in most complete data, and another one might be rather interested in most accurate data. In both cases, it should be possible for these users to specify tolerance thresholds for data and source quality (or, at least, to have technical means to estimate the quality of query results from the different sources). In this perspective, we propose the specification of quality contracts for the sources in order to use dynamically quality requirements in the query processing. This technique enables to differ data integration to end-user according to flexible quality criteria. Our approach is based on a standard mediator-wrapper architecture extending mediation with new functionalities such as the quality contract negotiation. A implicit assumption underlying this research is that we incorporate a set of controls on the top of data sources to enhance the global system's reliability and its integrated data quality ; the final aim is to maintain a high probability of preventing, detecting and eliminating data non-quality for data integration.

4 A Description of the Multi-source Architecture

We argue that database techniques such as having an expressive internal data model and query language, together with a meta-information repository and meta-information analysis techniques constitutes a necessary foundation for a mediator system. In order to alleviate the problem of information overload and confusion when results of a query are presented, the classical solution is to rank the results according to consistent relevance assessments. Our approach is to include quality specifications for sources and send them within the query. Such a functionality for retrieval and integration of information must be supported by an easily extensible, scalable and customizable architecture for addressing a wide range of specific applications such as the biological data domain. In this perspective, we propose a multi-source architecture (see Figure 2). From the application layer, the user can submit a global query to the mediator which conjointly sends the query and a quality contract type to the sources' wrappers. The wrappers send the corresponding local query to their respective source. Information sources may be cross-referenced, structured or not and with (or without) a meta-information repository. They respond to their wrapper with the query result and a quality contract instance. At the mediation layer, the mediator computes : 1) a conformance score for each source corresponding to the constraint satisfaction with respect to the contract specification, and 2) a conflict score for the queried integrated result which represent the importance of the data conflicts when distinct information items referring the same real-world concept are aggregated into a multi-source object structure. The mediator negotiates with the best sources and combines the query results, before sending it to the user.

5 Source Integration and Quality-Driven Negotiation

We considered the integration as a two-step process with 1) the schema integration and 2) the materialized data integration. As an input, we extract the different biological source schemata and their respective data sets in order to obtain as an output, a reconciled data view for a target schema. Our target schema is the DTD given in Figure 1.

- The first step of schema integration implies : comparing the sources' and mediator's schemata, detecting the structural analogies and correspondences, resolving the structural conflicts and transforming the source schemata into a canonical form (the target schema).
- The second step of data values integration implies : clustering and identifying the records that refer the same real object, expressing the criteria that define the conditions for the detection of matching records and eliminating data duplicates and inconsistencies.

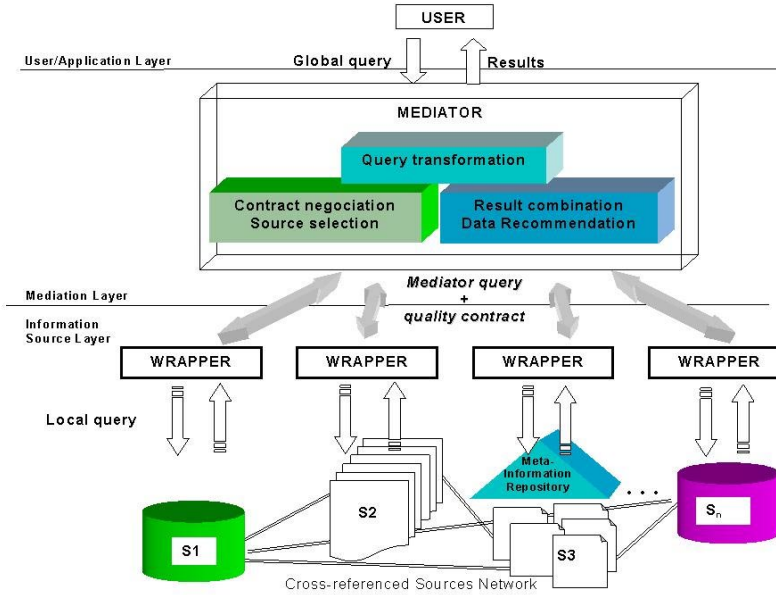


Fig. 2. The Multi-source Architecture including Quality Negotiation

5.1 Entity Identification

Anybody may submit biological information to public databanks with more or less formalized submission protocols which usually don't include names standardization or data quality controls. Erroneous data may be easily entered and cross-referenced. The available data sources have overlapping scopes with different levels of data quality. But the main problem of the biological databanks seems to be the entity identification. This problem arises when information from different information sources related to the same entity has to be merged. The databases may have inconsistent values in equivalent attributes of tuples referring to the same real-world object or may have mismatched attributes in them (i.e., attributes at the different level of abstraction), or have a combination of both. The identification problem appears as a biological description problem which requires a high-level expertise and the specialists' knowledge. Some tools (such as LocusLink [5]) propose clusters of records which identify the same biological concept accross different biological databanks : they are semantically related but biologists still must validate the correctness of the clusters and resolve interpretation differences among the records. As a starting point of our work, we considered that the biologist's expertise for entity identification is necessary. One of our objectives is to develop tools to assist the scientist in this task.

[5] LocusLink,http://www.ncbi.nlm.nih.gov/LocusLink/

5.2 Selective Structuration

In our example, each biological source propose a HTML document describing partially the gene HFE_Human identified by a particular record accession number. These documents are parsed and converted into the XML format and restructured according to our target schema (see Figure 3).

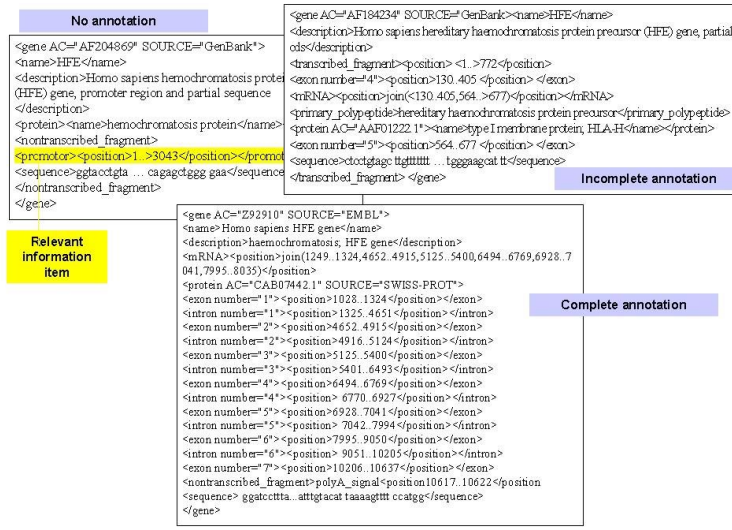


Fig. 3. Selective Structuration for Data Extraction

We use a set of mapping rules for identifying homologous structural elements (tags) based on biological shared ontologies which specifically help to determine the descriptive inclusions and the correspondances between the sources schemata and our target schema. Other rules are based 1) on the acquisition of biological expertise 2) on the results of statistical tools for sequence alignment (such as Blast[6]) 3) on format equivalence and conversion. In our example, the equivalence between exon positions elements can be automatically found out because the sequences respectively proposed by S_1 and S_2 have an alignment of 99% and the exons position proposed by S_2 is relative with an offset of 664 : exon.position(S_1.Z92910)= 6364 + exon.position(S_2.AF184234).

A multi-source data view is built by aggregating exhaustively all the information items available on gene HFE_Human proposed by the different sources if the following conditions are satisfied : i) the data refer to the same biological real-world concept ii) the data conflicts raised between more or less contradictory

[6] BLAST, http://www.ncbi.nlm.nih.gov/BLAST/

values are negligible (i.e. under a given threshold) and still allow data aggregation into a multi-source element which can be validated by the biologist. The sources' identification is mentioned for each value of the multi-source element. We obtain the following multi-source record for the HFE_Human in Figure 4.

<GENE AC="AF184234" source="GENBANK"
AC="AF204869" source="GENBANK"
AC="Z92910" source="EMBL">
<NAME>
<name AC="AF184234">HFE</name>
<name AC="AF204869">HFE</name>
<name AC="Z92910">Homo sapiens HFE gene</name>
</NAME>
<DESCRIPTION>
<description AC="AF184234">Homo sapiens hereditary haemochromatosis protein precursor (HFE) gene, partial cds</description>
<description AC="AF204869">Homo sapiens hemochromatosis protein (HFE) gene, promoter region and partial sequence</description>
<description AC="Z92910">haemochromatosis; HFE gene</description>
</DESCRIPTION>
<SEQUENCE>
<sequence AC="AF184234" >ctcctgtagc ttgttttttt ... tgggaagcat tt </sequence>
<sequence AC="AF204869">ggtaactgta atccoagctg... cagagctggg gaa</sequence>
<sequence AC="Z92910"> ggatcottta acogaggaga... taaaagtttt ccatgg </sequence>
</SEQUENCE>
<mRNA>
<mRNA AC="AF184234" > <position>join(<130..405,564..>677)</position></mRNA>
<mRNA AC="Z92910" ><position>join(1249..1324,4652..4915,5125..5400,6494..6769,6928..7041, 7995..9095)</position>
</mRNA>
<PRIMARY_POLYPEPTIDE>
<primary_polypeptide AC="AF184234">hereditary haemochromatosis protein precursor</primary_polypeptide>
</PRIMARY_POLYPEPTIDE>
<PROTEIN>
<protein AC="AAF01222.1" AC="AF184234"><name>type I membrane protein; HLA-H</name></protein>
<protein AC="AF204869"><name>hemochromatosis protein</name></protein>
<protein AC="CAB07442.1">
</PROTEIN>
<PROMOTOR>
<promotor AC="AF204869"><position>1..>3043</position></promotor>
</PROMOTOR>
<EXON>
<exon number="1" AC="Z92910"><position>1038..1324</position></exon>
<exon number="2" AC="Z92910"><position>4652..4915</position></exon>
<exon number="3" AC="Z92910"><position>5125..5400</position></exon>
<exon number="4" AC="AF184234" ><position>130..405 </position> </exon>
<exon number="4" AC="Z92910"><position>6494..6769</position></exon>
<exon number="5" AC="AF184234"><position>564..677</position> </exon>
<exon number="5" AC="Z92910"><position>6928..7041</position></exon>
<exon number="6" AC="Z92910"><position>7995..9050</position></exon>
<exon number="7" AC="Z92910"><position>10206..10637</position></exon>
</EXON>
<INTRON>
<intron number="1" AC="Z92910"><position>1325..4651</position> </intron>
<intron number="2" AC="Z92910"><position>4916..5124</position> </intron>
<intron number="3" AC="Z92910"><position>5401..6493</position> </intron>
<intron number="4" AC="Z92910"><position> 6770..6927 </position></intron>
<intron number="5" AC="Z92910"><position> 7042..7994</position></intron>
<intron number="6" AC="Z92910"><position> 9051..10205</position></intron>
</INTRON>
</GENE>

Fig. 4. Multi-source record aggregating available information on gene HFE_Human

5.3 Source Quality Negotiation

In order to provide adequate level of quality, the mediator needs to include capabilities such as negotiation, monitoring and adaptation. These capabilities all require the expected and the provided quality levels to be explicitly specified. Quality dimensions can be specified statically at the time of source integration or dynamically at deployment or runtime. We characterize quality of data and quality of source along named dimensions (mentioned non-exhaustively in Table 2). Specifying abstractly quality dimensions with a name and a domain value gives a flexible approach for deciding which dimension should be provided and implemented for a given application. The transparency of user's global query processing is weakened by the fact that the user can specify quality contracts for the query results. Based on the matching of user's quality requirements and

the quality for each local source, source negotiation and data conciliation are generated dynamically by the global query processor of the mediator to ensure the retrieval of high quality data from the multiple local sources.

Quality Contract Type	Description
Availability	Time and way the source is accessible based on technical equipment and statistics (example of dimension : serverFailure)
Freshness	How up-to-date the information is (e.g. dataAge, lastUpdate, updateFrequency)
Accessibility	Estimation of waiting time for user seaching time and for request/response processing (including the time consumption per-query of the wrapper for translating, negotiating ...)
Security	Estimation of the number of corrupted data
Coverage	Estimation of the number of data for a specific information domain
Accuracy	Estimation of the number of data free-of-error
Completeness	Estimation of the number of missing data or null values (e.g. NbOfNullValue per object)
User satisfaction	User grade based on presentation of data results and ease of understanding and using

Table 2. Quality Contract Types

For our particular application, these quality dimensions may be specified in a contract type which represents quality dimensions, specifies the name, the domain and possibly user-defined ordering for each dimension. We can specify examples of source contract types such as : Availability, Freshness and Completeness. A contract is an instance of a contract type that represents a set of quality dimensions specifications for the source. A contract aggregates a number of constraints. Figure 5 presents three contract types and their instances.
Contract Conformance corresponds to constraint satisfaction for each dimension of each contract type.

Definition 1
We define a conformance score for each source S_j as a weighted function F on the i specified contract types and their k dimensions.
$\forall S_j, contractType_i, dimensionName_k,$

$$Conformance(S_j) = F(w_i, w_k, ContractType_i, dimensionName_k, S_j)$$

Each contract type may have particular weights w_i indicating the relative importance of the ith contract type for computing the conformance score of each source S_j (e.g. the contract type on `Freshness` is more important than the one on `Completeness`). Each dimension of a contract type may have particular weights w_k indicating the relative importance of the kth dimension of the contract (e.g. `dataAge` is more important than `updateFrequency`). The conformance scores range from 0 to 1. A multi-source object is built for each element of the mediator conceptual model. The next step is to select among the multi-source values those that best match the query and whose source best conforms the quality contract. Given a global query Q, the mediator computes the conformance score $Conformance(S, Q, o)$ of each source S. Source S responds with the objects that

```
type Availability = contract {                    S1_Availability = Availability contract {
serverFailure :enum{halt,initState,rolledBack};   serverFailure == initState;
numberOfFailures :decreasing number fail./month;  numberOfFailures <= 0.2 fail./month;
reliability : increasing number;};                reliability == 0.999; };

type Freshness = contract {                       S1_ Freshness = Freshness contract {
dataAge : number year,month,day;                  dataAge == 8 years, 11 months, 3 days;
lastUpdate : number day(s);                       lastUpdate == 52 days;
UpdateFrequency : number updates/month;};         updateFrequency == 25 updates/day;};

type Completeness = contract {                    S1_Completeness = Completeness contract {
NbOfNullValue : increasing number/Object;};       NbOfNullValue : 3/Object;}
```

Fig. 5. Example of quality source contract types and their instances

best match the query values. The query results contain the values for $o_i, \ldots, o$ for every object returned.

5.4 Source Conformance and Query Results Scoring

Each object o in the result for query Q is ranked according to a conformance score $Conformance(S, Q, o)$ of the source S and a conflict score $Conflict(S, Q, o)$. The values of conflict score range from 0 to 1.

Definition 2

The conflict score $Conflict(S, Q, o)$ of source S for query Q corresponds to the distance between the object o of S and q the queried object of Q. The distance $Dis(o, q)$ is the sum of data conflict importances :

$Dis(o, q) = \sum_{j=1}^{n} I_j$ with I_j the jth data conflict importance such as :

$$I_j = \begin{cases} 0 & \text{if there is no data conflict} \\ 0.1 & \text{if the conflict is weak} \\ 1 & \text{if the conflict is strong} \end{cases} \quad (1)$$

6 Conclusion

In order to facilitate the multi-source data integration in the context of distributed biological databanks, we propose a technique based on the notions of quality contract and source negotiation. Our approach is based on a standard wrapper-mediator architecture. A quality contract with a source allows to specify quality dimensions necessary to the mediator for data selection between several distributed applications. The selectivity of data sources is dynamically computed during contract negotiation we propose to associate within the global query processing and before data acquisition. The importance of semi-structured data conflicts is evaluated and the quality conformance of sources is scored : the integration of the data is carried out according to the quality of the data required by users. The originality of our approach with respect to data mediation and conciliation is to include specifications quality of source into the query processing. From the biological application point-of-view, we first introduce the notion

of biological integrated data quality and data recommendation for biologists. The complete toolkit for biological data cleaning and biologists' assistance for expressing mapping rules are under current development. Our final objective is to orientate the current standardization efforts to take into account the quality of biological data and to promote operational techniques and tools to evaluate and improve it.

References

[1] M. Carey, L. Haas, and P. Schwarz et al. Towards heterogeneous multimedia information systems: The GARLIC approach. In *RIDE-DOM*, pages 124–131, March 1995.

[2] S. Chawathe, H. Garcia-Molina, and J. Hammer et al. The TSIMMIS project: Integration of heterogeneous information sources. *IPSJ*, pages 7–18, October 1994.

[3] C. Chee, Y. Arens, C. Knoblock, and C. Hsu. Retrieving and integrating data from multiple information sources. *Intl. J. of Intelligent and Cooperative Information Systems*, 2(2):127–158, 1993.

[4] D. Clavanese, G. De Giacomo, and M. Lenzerini et al. Data integration in datawarehousing. Tech. Rep., 1997.

[5] S. Cluet, C. Delobel, J. Siméon, and K. Smaga. Your mediators need data conversion ! In *ACM SIGMOD Conf. on Management of Data*, pp. 177–188, 1998.

[6] M. Fernandez, D. Florescu, and J. Kang et al. Catching the boat with STRUDEL: Experiences with a web-site management system. In *ACM SIGMOD Conf. on Management of Data*, pp. 414–425, 1998.

[7] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative data cleaning : Language, model, and algorithms. Tech. Rep. RR-4149, INRIA, 2001.

[8] C. Goh, S. Madnick, and M. Siegel. Context Interchange : overcoming the challenges of the large-scale interoperable database systems in a dynamic environment. In *Proc. of CIKM'94*, pp. 337–346, 1994.

[9] M. Goodchild and R. Jeansoulin. *Data quality in geographic information : from error to uncertainty*. Hermès, 1998.

[10] W. Hou, Z. Zhang. Enhancing database correctness : a statistical approach. In *Proc. of ACM SIGMOD Conf. on Management of Data*, 1995.

[11] R. Hull. Managing semantic heterogeneity in databases: a theoretical prospective. In *Proc. of PODS'97*, pp. 51–61, 1997.

[12] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis. *Fundamentals of Data Warehouses*. Springer, 1998.

[13] S. H. Kan. *Metrics and models in software quality engineering*. Addison-Wesley, 1995.

[14] A. Y. Levy, D. Srivastava, and T. Kirk. Data model and query evaluation in global information system. *J. of Intelligent Information Systems*, 5(2):121–143, 1995.

[15] E.P. Lim, J. Srivastava, and S. Shekhar. Resolving attribute incompatibility in database integration : An evidential reasoning approach. In *Proc. of the 10th Intl. Conference on Data Engineering (ICDE'94)*, 1994.

[16] A. Monge, C. Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.

[17] F. Naumann, U. Leser. Quality-driven integration of heterogeneous information systems. In *Proc. of VLDB'99*, pp. 447–458, 1999.
[18] J. Ordille, A. Levy, and A. Rajaraman. Querying heterogeneous information sources using source descriptions. In *Proc. of VLDB'96*, pp. 251–262, 1996.
[19] Y. Papakonstantinou, H. Garcia-Molina, and J. Widom. Object exchange across heterogeneous information source. In *Proc. of ICDE'95*, pp. 251–260, 1995.
[20] T.C. Redman. *Data quality for the information age*. Artech House, 1996.
[21] J. Rothenberg. Metadata to support data quality and longevity. In *Proc. of IEEE Metadata Conf.*, 1996.
[22] F. Sadri. Reliability of answers to queries in relational databases. *IEEE TKDE*, 3(2):245–252, 1991.
[23] J. Schlimmer. Learning determinations and checking databases. In *Proc. of the AAAI-91 Workshop on KDD*, 1991.
[24] A. Sheth, C. Wood, and V. Kashyap. Q-data : Using deductive database technology to improve data quality. In *Proc. of ILPS'93*, pp. 23–56, 1993.
[25] D. Strong, Y. Lee, and R. Wang. Data quality in context. *Com. of the ACM*, 40(5):103–110, 1997.
[26] G. Tayi, D. Ballou. Examining data quality. *Com. of the ACM*, 41(2):54–57, 1998.
[27] R. Wang. A product perspective on Total Data Quality Management. *Com. of the ACM*, 41(2):58–65, 1998.
[28] R. Wang, S. Madnick. A polygen model for heterogeneous database systems : the source tagging perspective. In *Proc. of VLDB'90*, pp. 519–538, 1990.
[29] R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE TKDE*, 7(4):623–638, 1995.
[30] G. Wiederhold. Mediation in information systems. *ACM Computing Surveys*, 27(2):265–267, 1995.