QDex :
A database profiler for generic bio-data exploration and quality aware integration

F. Moussouni[1], L. Berti-Équille [2], G. Rozé[1], O. Loréal, E. Guérin[1]
[1] INSERM U522 CHU Pontchaillou, 35033 Rennes, France
[2] IRISA, Campus Universitaire de Beaulieu, 35042 Rennes, France
Corresponding author : fouzia.moussouni@univ-rennes.fr

Abstract: In human health and life sciences, researchers extensively collaborate with each other, sharing genomic, biomedical and experimental results. This necessitates dynamically integrating different databases into a single repository or a warehouse. The data integrated in these warehouses are extracted from various heterogeneous sources, having different degrees of quality and trust. Most of the time, they are neither rigorously chosen nor carefully controlled for data quality. Data preparation and data quality metadata are recommended but still insufficiently exploited for ensuring quality and validating the results of information retrieval or data mining techniques.
In a previous work, we built a data warehouse called GEDAW *(Gene Expression Data Warehouse)* that stores various information: data on genes expressed in the liver during iron overload and liver diseases, relevant information from public databanks (mostly in XML), DNA-chips home experiments and also medical records. Based on our past experience, this paper reports briefly on the lessons learned from biomedical data integration and data quality issues, and the solutions we propose to the numerous problems of schema evolution of both data sources and warehousing system. In this context, we present QDex, a Quality driven bio-Data Exploration tool, which provides a functional and modular architecture for database profiling and exploration, enabling users to set up query workflows and take advantage of data quality profiling metadata before the complex processes of data integration in the warehouse. An illustration with QDex Tool is shown afterwards.

Keywords: warehousing, metadata, bio-data integration, database profiling, bioinformatics, data quality

1. INTRODUCTION

In the context of modern life science, integrating resources is very challenging, mainly because biological objects are complex and spread in highly autonomous and evolving web resources. Biomedical web resources are extremely heterogeneous as they contain different kinds of data, have different structure and use different vocabularies to name same biological entities. Their information and knowledge contents are also partial and erroneous, morphing and in perpetual progress.

In spite of these barriers, we assist in bioinformatics to an explosion of data integration approaches to help biomedical researchers to interpret their results, test and generate new hypothesis. In high throughput biotechnologies data warehouse solutions encountered a great success in the last decades, due to constant needs to store locally, confront and enrich in-house data with web information for multiple possibilities of analyses.
A tremendous amount of data warehouse projects devoted to bioinformatics studies exists now in literature. These warehouses integrate data from various heterogeneous sources, having different degrees of quality and trust. Most of the time, the data are neither rigorously chosen nor carefully controlled for data quality. Data preparation and data quality metadata are recommended but still insufficiently exploited for ensuring quality and validating the results of information retrieval or data mining techniques [17]. Moreover, data are physically imported, transformed to match the warehouse schema which tends to change rapidly with user requirements, typically in Bioinformatics. In the case of materialised integration, data model modifications for adding new concepts in response to rapid evolving needs of biologists, lead to considerable updates of the warehouse schemas and their applications, complicating the warehouse maintainability.

Lessons learned from the problems of biomedical data sources integration and warehouse schema evolution are presented in this paper. The main data quality issues in this context with current solutions for warehousing and exploring biomedical data are shown [1,2]. An illustration is given using *QDex*, a Quality driven bio-Data Exploration tool that: *i)* provides a generic functional and modular architecture for database quality profiling and exploration, *ii)* takes advantage of data quality profiling metadata during the process of biomedical data integration in the warehouse and, *iii)* enables users to set up query workflows, store intermediate results or quality profiles, and refine their queries.

This paper is structured as follows: in Section 2, requirement analyses in bioinformatics and the limits of current data warehousing techniques with regards to data quality profiling are presented in the perspective of related work. In Section 3, an illustration with our experience in building a gene expression data warehousing system: system design, data curation, cleansing, analyses, and new insight on schema evolution, In Section 4, QDex architecture and functionalities to remediate to some of these limits are presented to provide database quality profiling and extraction of quality metadata, and *Section 6* concludes the paper.

2. RELATED WORK

2.1 Data integration issues at the structural level

High throughput biotechnologies, like transcriptome, generate thousands of expression levels on genes, measured in different physiopathological situations. Beyond the process of management, normalization and clustering, biologists need to give a biological, molecular and medical sense to these raw data. Expression levels need to be enriched with the multitude of data available publicly on expressed genes: nucleic sequences, chromosomal and cellular locations, biological processes, molecular function, associated pathologies, and associated pathways. Relevant information on genes must be integrated from public databanks and warehoused locally for multiple possibilities of analyses and data mining solutions.

In the context of biological data warehouses, a survey of representative data integration systems is given in [8]. Current solutions are mostly based on data warehouse architecture (e.g., GIMS[1], DataFoundry[2]) or a federation approach with physical or virtual integration of data sources (e.g., TAMBIS[3], P/FDM[4], DiscoveryLink[5]) that are based on the union of the local schemas which have to be transformed to a uniform schema. In [3], Do and Rahm proposed a system called GenMapper for integrating biological and molecular annotations based on the semantic knowledge represented in cross-references. Finally, BioMart [18], which is a query-oriented data integration system that can be applied to a single or multiple databases, is a heavily used data warehouse system in bioinformatics since it supports large scale querying of individual databases as well as query-chaining between them.

Major problems in the context of biomedical data integration come from heterogeneity, strong autonomy and rapid evolution of the data sources on the Web. A data warehouse is relevant as long as it adapts its structure, schemas and applications to the constantly growing knowledge on the bio-Web.

---

[1] *GIMS, http://www.cs.man.ac.uk/img/gims/*
[2] *DataFoundry, http://www.llnl.gov/CASC/datafoundry/*
[3] *TAMBIS, http://imgproj.cs.man.ac.uk/tambis/*
[4] *P/FDM, http://www.csd.abdn.ac.uk/~gjlk/mediator/*
[5] *DiscoveryLink, http://www.research.ibm.com/journal/sj/402/haas.html*

2.2 Bio-data quality Issues at the instance level

Recent advancement in biotechnology has produced massive amount of raw biological data which are accumulating at an exponential rate. Errors, redundancy and discrepancies are prevalent in the raw data, and there is a serious need for systematic approaches towards biological data cleaning. Biological databanks providers will not directly support data quality evaluations to the same degree since there is no equal motivation for them to and there are currently no standards for evaluating and comparing biomedical data quality. Little work has been done on biological data cleaning and it is usually carried out in proprietary or ad-hoc manner, sometimes even manual. Systematic processes are lacking. From among the few examples, Thanaraj uses in [14] stringent selection criteria to select 310 complete and unique records of Homo sapiens splice sites from the 4300 raw records in EMBL database.

Moreover, bio-entity identification is a complex problem in the biomedical domain, since the meaning of "entity" cannot be defined properly. In most applications, identical sequences of two genes in different organisms or even in different organs of the same organism are not treated as a single object since they can have different behaviours. In GENBANK data source for example, each sequence is treated as an entity in its own, since it was derived using a particular technique, has particular annotation, and could have individual errors.

Müller et al. [11] examined the production process of genome data and identified common types of data errors. Mining for patterns in contradictory biomedical data has been proposed in [10], but data quality evaluation techniques are needed for structured, semi-structured or textual data before any biomedical mining applications. Although rigorous elimination of data is effective in removing redundancy, it may result in loss of critical information. In another example, a sequence structure parser is used to find missing or inconsistent features in records using the constraints of gene structure [12]. The method is only limited to detecting violations of the gene structure.

More specific to data quality scoring in the biomedical context, [9] propose to extend the semi-structured model with useful quality measures that are *biologically-relevant*, *objective* (i.e., with no ambiguous interpretation when assessing the value of the quality measure), and *easy to compute*. Six criteria such as stability (i.e., magnitude of changes applied to a record), density (i.e., number of attributes and values describing a data item), time since last update, redundancy (i.e., fraction of redundant information contained in a data item and its sub-items), correctness (i.e., degree of confidence that the data represents true information), and usefulness (i.e., utility of a data item defined as a function combining density, correctness, and redundancy) are defined and stored as quality metadata for each record (XML file) of the genomic databank of RefSeq . The authors also propose algorithms for updating the scores of quality measures when navigating, inserting or updating/deleting a node in the semi-structured record.

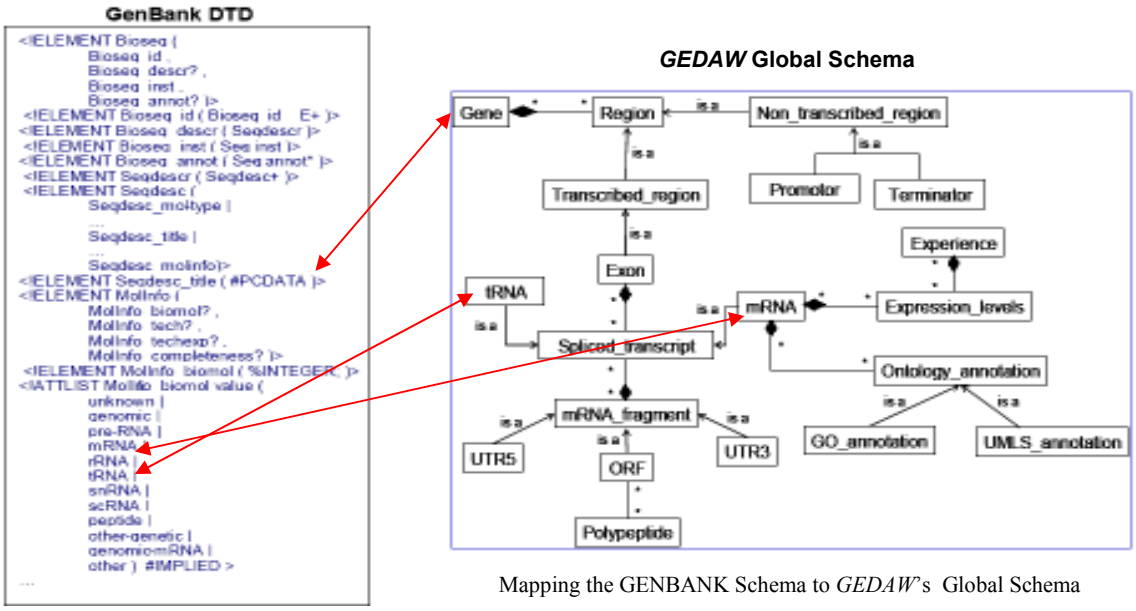
3. LESSONS LEARNED FROM BUILDING GEDAW

3.1 Database design, data integration, and application-driven workflow

The Gene Expression Data warehouse GEDAW [5] has been developed by the National Institute of Health Care and Medical Research (INSERM U522) to warehouse data on genes expressed in the liver during iron overload and liver pathologies. For interpreting gene expression measurements in different physiopathological situations in the liver, relevant

information from public databanks (mostly in XML format), micro-array data, DNA chips home experiments and medical records are integrated, stored and managed into GEDAW. GEDAW aims at studying in-silico liver pathologies by enriching expression levels of genes with data extracted from the variety of scientific data sources, ontologies and standards in life science and medicine including GO ontology [6] and UMLS [7].

Designing a single global data warehouse schema (Fig 1) that integrates syntactically and semantically the whole heterogeneous life science data sources is still challenging. In GEDAW context, we integrate structured and semi-structured data sources and use a Global As View (GAV) schema mapping approach and a rule-based transformation process from a source schema to the global schema of the data warehouse (see [4] for details).



Mapping the GENBANK Schema to *GEDAW*'s Global Schema

Fig 1: UML Class diagram representing the conceptual schema of GEDAW and some correspondences with the GENBANK DTD (e.g., Seqdes_title and Molinfo values will be extracted and migrated to the name and other description attributes of the class Gene in GEDAW schema).

With the overall integrated knowledge, the warehouse has provided an excellent analysis framework where enriched experimental data can be mined through various workflows combining successive analysis steps.

GEDAW supports several functions that consist of analyses on demand made on a group of genes of interest upon a database selection query with one or more criteria. These analyses correspond to APIs that use OQL (Object Query Language) and java to retrieve multiple information items about the genes. Some external analyses that correspond to external bioinformatics tools have been applied on subsets of integrated data on genes, as clustering for example. These two kinds of analyses have been combined to connect successive steps, thus forming a workflow.

One of them (Fig 2) has been designed according to the hypothesis that genes sharing an expression pattern should be associated. The strategy consists in selecting a group of genes that are associated with a same disease and a typical expression pattern (steps 1 and 2 in Fig 2), and then extrapolate this group to more genes involved in the disease (step 5) by searching for

expression pattern similarity (step 4). The genes are then characterized by studying the biological processes and the cellular components using integrated GO annotations (step 6).

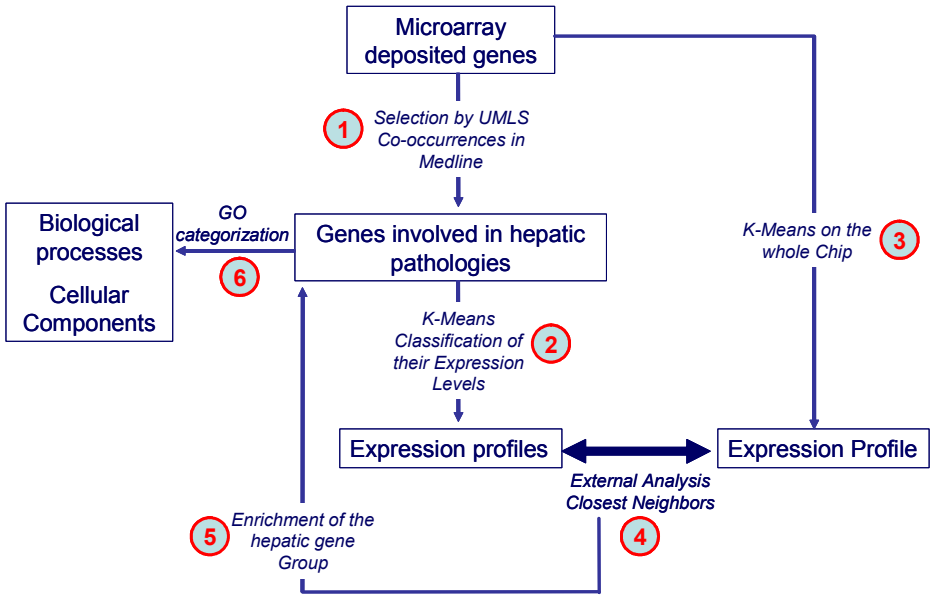


Fig 2: Combining Biomedical Information within an Expert Guided Workflow

This example, which is expert guided, has been used in order to extract new knowledge consisting of new gene associations to hepatic disorders [5]. The found genes are now biologically investigated by the expert for a better understanding of their involvement in the disease.

Requirement analysis from biologists and their associated workflows have been since rapidly evolving with a non-stop emergence on the Web of new complex data types like protein structures, gene interactions or metabolic pathways, urging to continuous evolution of the warehouse schema, contents and applications.

3.2 Bio-entity identification

By using GAV mapping approach for integrating one data source at a time in GEDAW (e.g. Fig 1 with GENBANK), we have minimized as much as possible the problem of identification of equivalent attributes. The problem of equivalent instances identification is still complex to address. This is due to general redundancy of bio-entities in life science even within a single source. Biological databanks may also have inconsistent values in equivalent attributes of records referring to the same real-world object. For example, there are more than 10 ID's records for the same DNA segment associated to human HFE gene in GENBANK! Obviously the same segment could be a clone, a marker or a genomic sequence.

Anyone is indeed able to submit biological information to public databanks with more or less formalized submission protocols that usually do not include names standardization or data quality controls. Erroneous data may be easily entered and cross-referenced. Even if some tools propose clusters of records (like EntryGene for GENBANK) which identify the same biological concept across different biological databanks for being semantically related, biologists still must validate the correctness of these clusters and resolve the differences of interpretation among the records.

This is a typical problem of entity resolution and record linkage that is augmented and made more complex due to the high-level of expertise and knowledge it requires (i.e., difficult to formalize and related to many different sub-disciplines of biology, chemistry, pharmacology, and medical sciences). After the step of bio-entity resolution, data are scrubbed and transformed to fit the global DW schema with the appropriate standardized format for values, so that the data meets all the validation rules that have been decided upon by the warehouse designer.

Problems that can arise during this step include null or missing data; violations of data type; non-uniform value formats; invalid data. The process of data cleansing and scrubbing is rule-based. Then, data are migrated, physically integrated and imported into the data warehouse. During and after data cleansing and migration, quality metadata are computed or updated in the data warehouse metadata repository by pre- and post- data validation programs.

4. DATABASE PROFILING FOR GENERIC DATA EXPLORATION AND QUALITY AWARE INTEGRATION

4.1 Database Profiling

Database Profiling is the process of analyzing a database to determine its structure and internal relationships. It consists mainly of identifying: *i)* the objects used and their attributes (contents and number), *ii)* relationships between objects with their different kinds of associations including aggregation and inheritance, and *iii)* the objects behaviour and their relative functions. Database profiling is then useful when managing data conversion and data cleanup projects.

Fig 3: The XMI document generated from GEDAW schema

The XMI *(XML Metadata Interchange)* document (see Fig 3) that collects metadata information on the objects of the database is generated on-demand for profiling GEDAW. It has been quite useful to face the syntactic heterogeneity of the evolving schemas of data sources and the warehousing system during its life cycle. A generic data exploration has been made possible by the development of QDex tools (Quality based Database Exploration) that parse the XMI document detailing the database structure (in terms of class, attributes, relationships, etc.) and generate a model-based interface to explore the multiple attributes on genes description stored in the warehouse.

4.2 Data Quality Profiling

Data quality profiling is the process of analyzing a database to identify and prioritize data quality problems. The results include simple summaries (counts, averages, percentages, etc.) describing for instance: completeness of datasets and the number of missing data records, the data freshness, and various data problems in existing records (e.g., outliers, duplicates, redundancies). During the process of data profiling, available data in the existing database are examined and statistics are being computed and gathered to track different summaries describing aspects of data quality. As a result, by providing QDex data profiling tools, one also provides data quality profiling tools.

A considerable amount of data quality research involves investigating and describing various categories of desirable attributes (or dimensions) of data quality. These lists commonly include accuracy, consistency, completeness, unicity (i.e., no duplicates), and freshness. Nearly 200 such terms have been identified in [15,16], regarding nature, definitions and measures of attributes.

Contradictory or ambiguous data is also a crucial problem as well, especially in bioinformatics where data are continuously speculative. Centralizing data in a warehouse is one of the initiatives one can take to ensure data validity.

Taking advantage of the stored XMI metadata information obtained by database profiling using the XMI document, QDex provides generic tools for bio-database exploration and data quality profiling. In developing QDex, we believe that profiling databases (both considering the structure of data sources and data warehouse) could be very useful for the integration process. Moreover, our work examines the extent of biological database profiling and proposes a way for flexibly building query workflows that follow the reasoning of biologists and assist them in the elaboration of their pioneer queries, including queries for data quality track.

5. QDEX USE CASE: APPLICATION TO GEDAW

5.1 Generic bio-data exploration

A global overview of QDex interface is given in Fig 4. Parts of the workflow that has been used to combine biological and medical knowledge to extract new knowledge on liver genes, has been flexibly reformulated using QDex GUI. The screen-shot below shows the central database of GEDAW as profiled using the XMI metadata document which gives an insight on the current warehouse schema. An overview of the extracted database profiling (classes and attributes) is browsed on the *Database Schema Viewer* frame. This includes the Gene, mRNA, ExpressionLevels, GOAnnotation and UMLSAnnotation classes. Based on these classes, the

user built by himself, scenarios of queries on the objects, using his criteria, in the *Query Maker* frame.

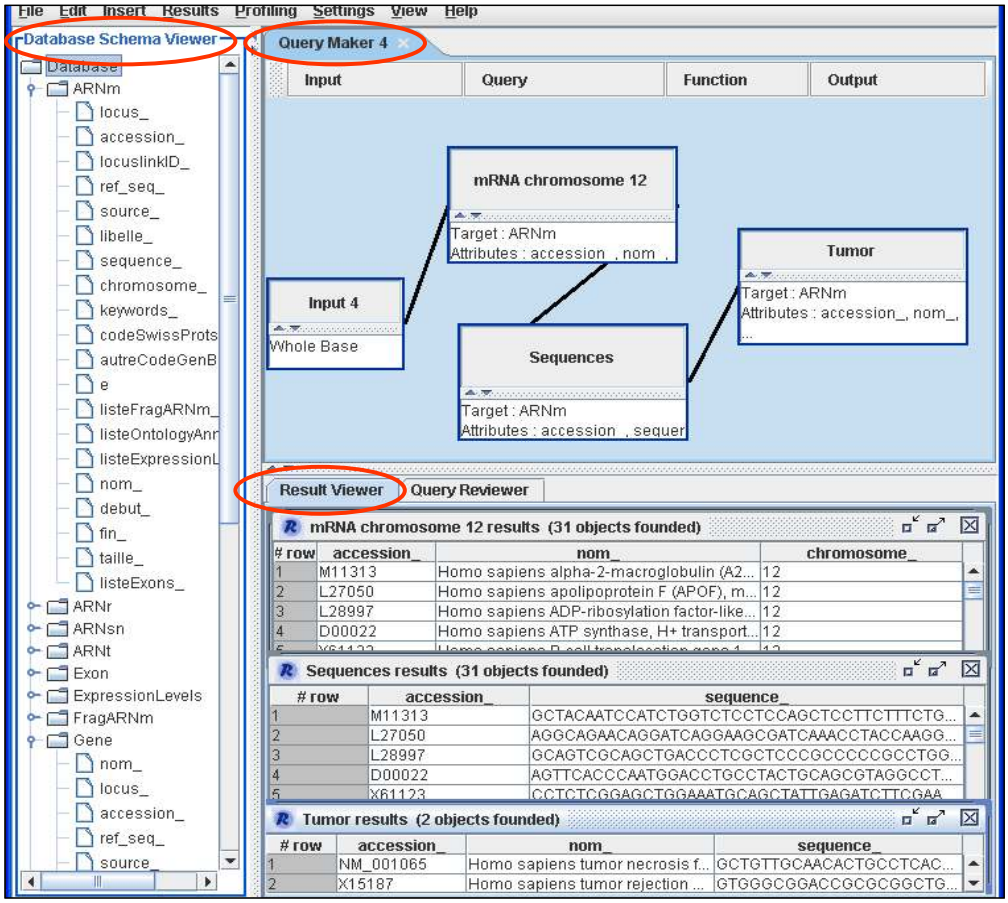


Fig 4: Global Overview of QDex interface

By having an immediate glance on his intermediate or final results browsed on the *Result Viewer*, the user may modify and re-execute his queries when needed. He is also able to save a workflow for ulterior reuse on different data, or export effective resulting data for a future use on external tools (clustering, spreadsheet, etc,). This interface makes QDex quite flexible and attractive for the biologist.

To construct the Liver Disease Associated Genes Group, the Genes of the array that are annotated by "liver disease" concept and its descendants in UMLS are selected using UMLSAnnotation Class (See Fig igure4). Corresponding mRNA or Gene names are browsed on the Result Viewer sub-frame. Using the query maker, the selected objects are refined using successive queries on the group by adding boxes on demand, to look for information on their sequences, their expression levels, and their annotations in Gene Ontology making a more exhaustive workflow.

5.2 Preliminary tools for bio-data quality track

The completeness dimension of the result of a query workflow is computed by counting the number of missing values of queried objects (see Fig 5). Actually, by using QDex, much more possibilities are offered to the user to compose various workflows on integrated objects in GEDAW.

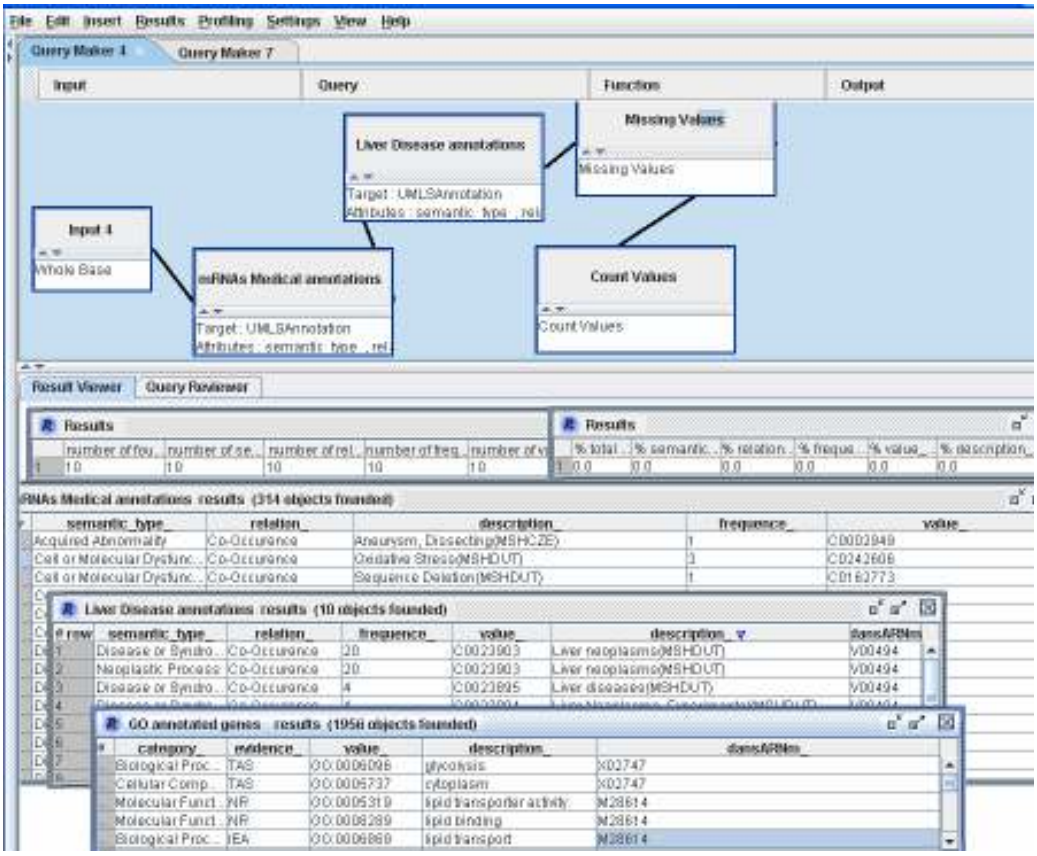


Fig 5: Preliminary tools for tracking completeness of biomedical data

The user can have indicators associated to the datasets or query results by specifying various useful metrics to describe the aspects of database or query result quality. QDex project being still in progress, more tools will be provided to the user for evaluating the quality of the data that are being explored including redundancy, freshness, and inconsistency (by checking user-defined or statistical constraints).

6. CONCLUSION

In this paper, we have presented a database profiling approach for designing a generic biomedical database exploration tool devoted to quality aware data integration and exploration. QDex has been applied to GEDAW: an object oriented data warehouse devoted to the study of high throughput gene expression levels in the domain of hepatology. Metadata extracted from the XMI document of GEDAW have been used to provide a generic interface that supports tools for convivial building of query workflows using multiple profiled attributes on the genes and preliminary tools for data quality track. By developing QDex, data are supposed already being integrated. Using QDex, the user has the ability to make a clearer view of the database content and quality. As we have mentioned, QDex is under ongoing development and our perspectives are to keep on taking advantage of the extracted metadata information, and to provide more tools (such as a quality metric library) to be gradually integrated to the interface in order to evaluate the quality of the data that are being explored. Our main objective is to cover the main data quality dimensions by providing predefined analytical functions whose results (as computed indicators) will describe various aspects of consistency, accuracy, unicity, and freshness of data. Another important aspect of our future work is linked to data quality

problems detection and concerns the design of pragmatic tools to help the expert to cleanse erroneous (or low quality) data within the QDex interface.

Finally, the original advantage of QDex resides in the fact that it can be generalized to any database schema outside bioinformatics. More specifically, we intend to apply QDex to the expected version of GEDAW which is being upgraded. This is for storing more actual bioinformatics data, like graph structures for gene pathways and system biology studies of genes expression profiles on the scale of a pangenomic DNA-Chip.

References

1. Anathakrishna, R., Chaudhuri, S., Ganti, V., Eliminating Fuzzy Duplicates in Data warehouses, *Proc. of Intl. Conf. VLDB*, 2002.
2. Batini C., Catarci T. and Scannapiceco M., A Survey of Data Quality Issues in Cooperative Information Systems, *Tutorial presented at the Intl. Conf. on Conceptual Modeling (ER)*, 2004.
3. Do, H.-H. and Rahm, E., Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach, *Proc. of the Intl. Conf. EDBT'04*, Heraklion, Greece, Springer LNCS, 2004.
4. Guérin E., Marquet G., Burgun A., Loréal O., Berti-Equille L., Leser U., Moussouni F., Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW, *Proc. of the 2nd Intl. Workshop on Data Integration in the Life Science (DILS)*, San Diego, 2005.
5. Guérin E., Marquet G., Chabalier J., Troadec M.B., Guguen-Guillouzo C., Loréal O., Burgun A., Moussouni F., Combining biomedical knowledge and transcriptomic data to extract new knowledge on genes, Journal of Integrative Bioinformatics, 3(2) 2006.
6. Harris MA et. al. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D258-61.
7. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
8. Lacroix, Z., Critchlow, T., (Ed.), Bioinformatics: Managing Scientific Data, Morgan Kaufmann, 2003.
9. Martinez A., Hammer, J. Making Quality Count in Biological Data Sources. *Proc. of the 2nd Intl. ACM Workshop on Information Quality in Information Systems (IQIS 2005)*, USA, June 2004.
10. Müller H., Leser U., Freytag J.-C., Mining for Patterns in Contradictory Data. *Proc. of the 1rst Intl. ACM Workshop on Information Quality in Information Systems (IQIS 2004)*, p. 51-58, France, June 2004.
11. Müller, H., Naumann, F., Freytag J.-C. Data Quality in Genome Databases. *Proc. of Conference on Information Quality* (ICIQ'03), p. 269-284, MIT, USA, 2003.
12. Overton,C.G. and Haas,J. Case-Based Reasoning Driven Gene Annotation. *Computational Methods in MolecularBiology.* Elsevier Science, 1998.
13. Rahm E., Do H., Data Cleaning: Problems and Current Approaches, *IEEE Data Eng. Bull.* 23(4): 3-13, 2000.
14. Thanaraj, T.A. A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures. *Nucleic Acids Res. 27*(13), 2627-2637, 1999.
15. Wang R. Y., Journey to Data Quality, *Advances in Database Systems*, Vol. 23, Kluwer Academic Press, Boston, 2002.
16. Wang, R., Kon, H. & Madnick, S. (1993), Data Quality Requirements Analysis and Modelling, Ninth International Conference of Data Engineering, Vienna, Austria. Article
17. Pyle D., Data Preparation for Data Mining, Morgan Kaufmann, 1999.
18. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W, BioMart and BioConductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005 Aug 15; 21(16):3439-40.