HAL
archives-ouvertes.fr

Quality-Aware Integration and Warehousing of Genomic Data

Laure Berti-Équille, Fouzia Moussouni

▶ To cite this version:

Laure Berti-Équille, Fouzia Moussouni. Quality-Aware Integration and Warehousing of Genomic Data. ICIQ'05 - 10th International Conference on Information Quality, Nov 2005, Massachusetts Institute of Technology, Cambridge, MA, United States. pp.1-15. hal-01855920

HAL Id: hal-01855920
https://hal.inria.fr/hal-01855920

Submitted on 8 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

QUALITY-AWARE INTEGRATION AND WAREHOUSING OF GENOMIC DATA

Laure Berti-Equille
U. of Rennes, France
Laure.Berti-Equille@irisa.fr

Fouzia Moussouni
INSERM U522, Rennes, France
rmoussou@univ-rennes1.fr

Abstract: In human health and life sciences, researchers extensively collaborate with each other, sharing biomedical and genomic data and their experimental results. This necessitates dynamically integrating different databases or warehousing them into a single repository. Based on our past experience of building a data warehouse called *GEDAW* (*Gene Expression Data Warehouse*) that stores data on genes expressed in the liver during iron overload and liver pathologies, and also relevant information from public databanks (mostly in XML format), DNA chips home experiments and medical records, we present the lessons learned, the data quality issues in this context and the current solutions we propose for integrating and warehousing biomedical data. This paper provides a functional and modular architecture for data quality enhancement and awareness in the complex processes of integration and warehousing of biomedical data.

Key Words: Data Quality, Data Warehouse Quality, Biological and Genomic Data, Data Integration

1. INTRODUCTION

At the center of a growing interest due to the rapid emergence of new biotechnological platforms in human health and life sciences for high throughput investigations in genome, transcriptome and proteome, a tremendous amount of biomedical data is now produced and deposited by scientists in public Web resources and databanks. The management of these data is challenging, mainly because: *i)* data items are rich and heterogeneous: experiment details, raw data, scientific interpretations, images, literature, etc. *ii)* data items are distributed over many heterogeneous data sources rendering a complex integration, *iii)* data are often asynchronously replicated from one databank to another (with the consequence that the secondary copies of data are often not updated in conformance with the primary copies), *iv)* data are speculative and subject to errors and omissions, some results are world-widely published although the corresponding experiments are still on-going or are not validated yet by the scientific community, and *v)* biomedical knowledge is constantly morphing and in progress. For the comprehensive interpretation of one specific biological problem (or even a single gene expression measurement for instance), the consideration of the entire available knowledge is required (e.g., the gene sequence, tissue-specific expression, molecular function(s), biological processes, regulation mechanisms, expression in different pathological situations or other species, clinical follow-ups, bibliographic information, etc.). This necessarily leads to the upward trend for the development of data warehouses (or webhouses) as the keystones of the existing biomedical Laboratory Information Management Systems (*LIMS*). These systems aim at extensively integrating all the available information related to a specific topic or a complex question addressed by the biomedical researchers leading to new diagnostics and therapeutic tools.

Nevertheless detecting data quality problems (such as duplicates, errors, outliers, contradictions, inconsistencies, etc.), correcting, improving and ensuring biomedical information quality when data comes from various information sources with different degrees of quality and trust are very complex and challenging tasks mainly because of the high level of knowledge and domain expertise they require.
Maintaining traceability, freshness, non-duplication and consistency of very large bio-data volumes for integration purposes is one of the major scientific and technological challenges today for research communities in bioinformatics and information and database systems.
As a step in this direction, the contribution of this paper is threefold: first, we give an overview on data quality research and projects of multi-source information system architectures that "natively" capture and manage different aspects of data quality and also related work in bioinformatics (Section 2); secondly, we share the lessons learned from the development and maintenance of a data warehouse system used to study gene expression data and pathological disease information; in this context, we present data quality issues and current solutions we proposed (Section 3); finally, we propose a modular architecture for data quality enhancement and awareness in the processes of biomedical data integration and warehousing (Section 4). Section 5 gives concluding remarks and present our research perspectives.

2. RELATED WORK

2.1 Data Quality Research Overview

Data quality is a multidimensional, complex and morphing concept [10]. Since a decade, there has been a significant emergence of work in the area of information and data quality management initiated by several research communities [1] (statistics, database, information system, workflow and project management, knowledge engineering and discovery from databases), ranging from techniques in assessing information quality to building large-scale data integration systems over heterogeneous data sources or cooperative information systems [2]. Many data quality definitions, metrics, models and methodologies [49][40] or *Extraction-Transformation-Loading (ETL)* tools have been proposed by practitioners and academics (e.g., [13][14][41][48]) with the aim of tackling the following main classes of data quality problems: *i)* duplicate detection and record matching (also known as: record linkage, merge/purge problem [18], duplicate elimination [23][21][1], name disambiguation, entity resolution [3]), *ii)* instance conflict resolution using heuristics, domain-specific rules, data source selection [26] or data cleaning and *ETL* techniques [39], *iii)* missing values and incomplete data [42], and *iv)* staleness of data [5][46][9].

Several surveys and empirical studies showed the importance of quality in the design of information systems, in particular for data warehouse systems [12][43]. Many works in the fields of information systems and software engineering address the quality control and assessment for the information and for the processes which produce this information [6][37][4]. Several works have studied in detail some of the properties that influence given quality factors in concrete scenarios. For example concerning currency and freshness of data, in [8] the update frequency is studied for measuring data freshness in a caching context. Other works combine different properties or study the trade-off between them, for example how to combine different synchronization policies [45] or the trade-off between execution time and storage constraints [25]. Other works tackled the problem of the evaluation of data quality.

[1] Se the organization of international conferences such as the *International Conference on Information Quality* (*ICIQ* http://web.mit.edu/tdqm/www/iqc) at *Massachusetts Institute of Technology since 1996*, several international workshops such as *Data Quality in Cooperative Information Systems* (*DQCIS'03* in conjunction with *ICDT'03*) and the *International Workshop on Information Quality in Information Systems* (*IQIS'04* and *IQIS'05* in conjunction with *ACM SIGMOD*), and special issues of well-known journals such as *IEEE Transactions on Data and Knowledge Engineering* and *Communications of ACM*,

In [37], the authors present a set of quality dimensions and study various types of metrics and the ways of combining the values of quality indicators. In [35], various strategies to measure and combine values of quality are described. In [6], a methodology to determine the quality of information is presented with various ways of measuring and combining quality factors like freshness, accuracy and cost. The authors also present guidelines that exploit the quality of information to carry out the reverse engineering of the system, so as to improve the trade-off between information quality/cost. The problem of designing multi-source information systems (e.g., mediation systems, data warehouses, web portals) taking into account information about quality has also been addressed by several approaches that propose methodologies or techniques to select the data sources, by using metadata on their content and quality [33][34][16].

Three research projects dedicated to tackle data quality issues provide an enhanced functional architecture (respectively for database, data warehouse and cooperative information system) are worth mentioning. Recently, the **Trio** project (started in 2005) at Stanford University [50] is a new database system that manages not only data, but also the accuracy and lineage of the data. The goals of the Trio project are: *i)* to combine previous work on uncertain and fuzzy data into a simple and usable model; *ii)* to design a query language as an understandable extension to SQL; *iii)* to build a working system that augments conventional data management with both accuracy and lineage as an integral part of the data.

The European ESPRIT **DWQ** Project (*Data Warehouse Quality*) (1996-1999) developed techniques and tools to support the design and operation of data warehouses based on data quality factors. Starting from a definition of the basic data warehouse architecture and the relevant data quality issues, the DWQ project goal was to define a range of alternatives design and operational method for each of the main architecture components and quality factors. In [20][47] the authors have proposed an architectural framework for data warehouses and a repository of metadata which describes all the data warehouse components in a set of meta-models to which a quality meta-model is added, defining for each data warehouse meta-object the corresponding relevant quality dimensions and quality factors. Beside from this static definition of quality, they also provide an operational complement that is a methodology on how to use quality factors and to achieve user quality goals. This methodology is an extension of the *Goal-Question-Metric (GQM)* approach, which permits: *a*) to capture the inter-relationships between different quality factors and *b*) to organize them in order to fulfill specific quality goals.

The Italian **DaQuinCIS** project (2001-2003) was dedicated to cooperative information systems and proposed an integrated methodology that encompassed the definition of an *ad-hoc* distributed architecture and specific methodologies for data quality measurement and error correction techniques [44]. This specific methodology includes process- and data-based techniques used for data quality improvement in single information systems. The distributed architecture of DaQuinCIS system consisted of (*i*) the definition of the representation models for data quality information that flows between different cooperating organizations via cooperative systems (CIS) and (*ii*) the design of a middleware that offers data quality services to the single organizations.

In error-free data warehouses with perfectly clean data, knowledge discovery techniques (such as clustering, mining association rules or visualization) can be relevantly used as decision making processes to automatically derive new knowledge patterns and new concepts from data. Unfortunately, most of the time, these data are neither rigorously chosen from the various heterogeneous sources with different degrees of quality and trust, nor carefully controlled for quality. Deficiencies in data quality still are a burning issue in many application areas, and become acute for practical applications of knowledge discovery and data mining techniques [36]. Data preparation and data quality metadata are recommended but still insufficiently exploited for ensuring quality in data warehouses and for validating mining results and discovered knowledge [38].

2.2 Quality of Integrated Biological Data

In the context of biological databases and data warehouses, a survey of representative data integration systems is given in [21]. But the current solutions are based on data warehouse architecture (e.g., GIMS[2], DataFoundry[3]) or a federation approach with physical or virtual integration of data sources (e.g., TAMBIS[4], P/FDM[5], DiscoveryLink[6]) that are based on the union of the local schemas which have to be transformed to a uniform schema. In [11], Do and Rahm proposed a system called *GenMapper* for integrating biological and molecular annotations based on the semantic knowledge represented in cross-references. More specific to data quality in the biomedical context, other work has been recently proposed for the assessment and improvement of the quality of integrated biomedical data. In [28] the author 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[7]. 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.

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. Müller *et al.* [31] examined the production process of genome data and identified common types of data errors. Mining for patterns in contradictory biomedical data has been proposed [30], but data quality evaluation techniques are needed for structured, semi-structured or textual data before any biomedical mining applications.

3. QUALITY-AWARENESS FOR BIOMEDICAL DATA INTEGRATION AND WAREHOUSING

In life sciences, researchers extensively collaborate with each other, sharing biomedical and genomic data and their experimental results. This necessitates dynamically integrating different databases or warehousing them into a single repository. Overlapping data sources may be maintained in a controlled way, such as replication of data on different sites for load balancing or for security reasons. But uncontrolled overlaps are very frequent cases. Moreover, scientists need to know how reliable the data is if they are to base their research on it because pursuing incorrect theories and experiments costs time and money. The current solution to ensure data quality in the biomedical databanks is curation by human experts. The two main drawbacks are: *i)* data sources are autonomous and as a result, sources may provide excellent reliability in one specific area, but not in all data provided, and *ii)* curation is a manual process of data accreditation by specialists that slows the incorporation of data and that is not free from conflicts of interest. In this context, more automatic, impartial, and independent data quality evaluation techniques and tools are needed for structured, semi-structured and textual biomedical data.

---

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

3.1 Some Lessons Learned from Bio-Data Integration and Warehousing

Searching across heterogeneous distributed biological resources is increasingly difficult and time-consuming for biomedical researchers. Data describing genomic sequences are available in several public databanks via Internet: banks for nucleic acids (DNA, RNA), banks for protein (polypeptides, proteins) such as *SWISS-PROT*[8], generalist or specialized databanks such as *GenBank*[9], *EMBL*[10] (*European Molecular Biology Laboratory*) and *DDBJ*[11] (*DNA DataBank of Japan*). Each databank record describes a sequence with several annotations. Each record is also identified by a unique accession number and may be retrieved by key-words (see Figure 2 for examples). Annotations may include the description of the genomic sequence: its function, its size, the species for which it has been determined, the related scientific publications and the description of the regions constituting the sequence (codon start, codon stop, introns, exons, ORF, etc.). The project *GEDAW* (*Gene Expression Data Warehouse*) [15] has been developed by the *French 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. Relevant information from public databanks (mostly in XML format), micro-array data, DNA chips home experiments and medical records are integrated, stored and managed in *GEDAW* for analyzing gene expression measurements. *GEDAW* aims at studying *in silico* liver pathologies by using expression levels of genes in different physio-pathological situations enriched with annotations extracted from the variety of the scientific data sources, ontologies and standards in life science and medicine.

Designing a single global data warehouse schema (Figure 1) that integrates syntactically and semantically the whole heterogeneous life science data sources is a very challenging task. In the *GEDAW* context, we integrate structured and semi-structured data sources and we use a *Global As View* (GAV) schema mapping approach and a rule-based transformation process from a given source schema to the global schema of the data warehouse (see [15] for details)

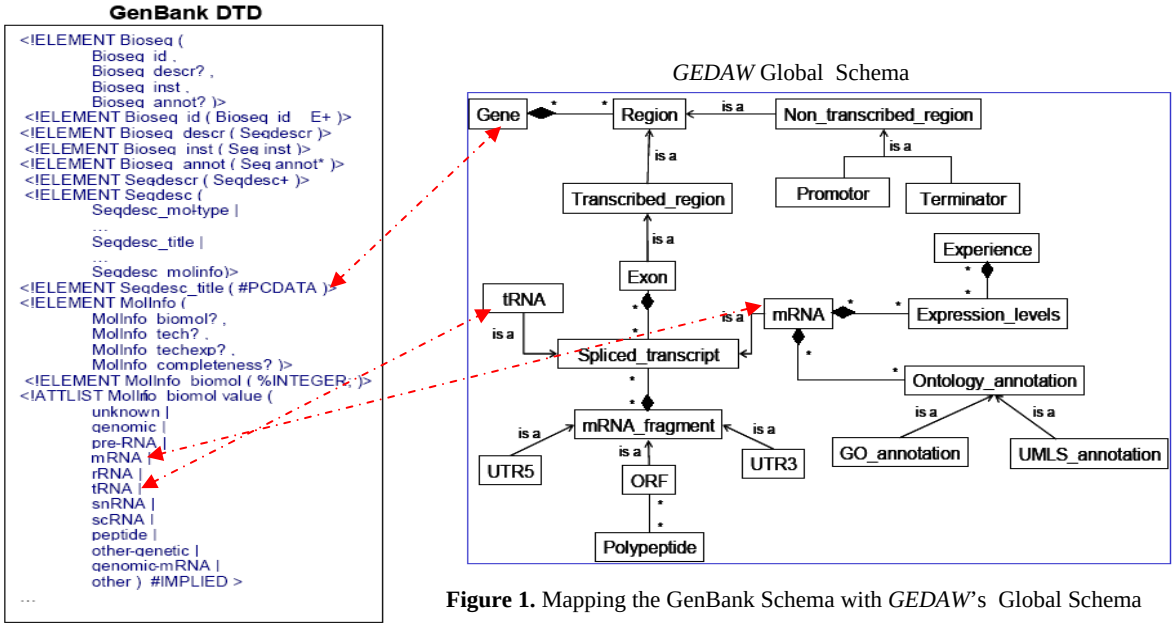

Figure 1. Mapping the GenBank Schema with *GEDAW*'s Global Schema

[8] *SWISS-PROT:* http://www.expasy.org/sprot
[9] *GenBank:* http://www.ncbi.nlm.nih.gov/Genbank
[10] *EMBL:* http://www.ebi.ac.uk/embl/
[11] *DDBJ:* http://www.ddbj.nig.ac.jp/

Figure 1 gives the 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 the *GEDAW* global schema).

3.2 Data Quality Issues and Proposed Solutions

The *GEDAW* input data sources are: *i) GenBank* for the genomic features of the genes (in XML format), *ii)* annotations derived from the biomedical ontologies and terminologies (such as *UMLS*[12], *MeSH*[13] and *GO*[14] also stored as XML documents), and *iii)* gene expression home measurements. Because gene expression data are massive (more than two thousands measures per experiment and hundreds of experiment per gene and per experimental conditions), the use of schema integration in our case – *i.e.*, the replication of the source schema in the warehouse - would highly burden the data warehouse.

By using a Global as View (*GAV)* mapping approach for integrating one data source at a time (e.g. in Figure 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 *LocusLink*[15] 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 interpretation differences among the records.

This is a typical problem of entity resolution and record linkage (see Section 3.2.1) 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 (see Section 3.2.2). Then, data are migrated and 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 (see Section 3.2.3).

---

[12] *Unified Medical Language System® (UMLS):* http://www.nlm.nih.gov/research/umls/
[13] *MeSH:* http://www.nlm.nih.gov/research/mesh
[14] *Gene Ontology™ (GO):* http://www.ontologos.org/IFF/Ontologies/Gene.html
[15] *LocusLink:* http://www.ncbi.nlm.nih.gov/LocusLink

3.2.1. Biological Entity Resolution and Record Linkage

Before data integration, the process of entity identification and record linkage can be performed using a sequence of increasingly sophisticated linkage techniques, described in the following, and also additional knowledge bases, ontologies and thesaurus (such as *UMLS Metathesaurus and MeSH-SR* vocabulary*)*, each operating on the set of records that were left unlinked in the previous phase:

1- Linkage based on exact key matching: *i.e.*, based on the gene names and the cross-referenced accession numbers (for instance between a gene from *Genew* and a protein in *SWISS-PROT*),
2- Linkage based on nearly exact key matching (*i.e.*, based on all the synonyms of a term and all the identifiers of a gene or gene product in *Genew*, the *UMLS Metathesaurus* and *MeSH-SR* and in the cluster of records proposed by *LocusLink*),
3- Probabilistic linkage based on the full set of comparable attributes (*i.e.*, based on the search for information about a gene or a gene product: the set of concepts related to this gene in the *Gene Ontology* (molecular function, biological process and cellular component) and the set of concepts related to the gene in *UMLS* and *MEDLINE*[16] abstracts including chemicals & drugs, anatomy, and disorders),
4- Search for erroneous links (false positives),
5- Analysis of residual data and final results for biological entity resolution.

3.2.2. Biomedical Data Scrubbing and Conflict Resolution

In order to define an appropriate data aggregation of all the available information items resulting from the first step of bio-entity resolution, data conflicts have to be resolved using rules for mapping the source records and conciliating different values recorded for a same concept. Mapping rules are defined to allow the data exchange from the public databanks into the *GEDAW* data warehouse. Apart from experimental data, public information items are automatically extracted by scripts using the *DTD* (*Document Type Definition*) of the data source translated into the *GEDAW* conceptual data model.
Three categories of mapping rules are proposed: 1) structural mapping rules, 2) semantic mapping rules and 3) cognitive mapping rules according to the different knowledge levels involved in the biological interpretation of data.

Structural mapping rules are defined at the schema level according to the *GEDAW* model by identifying the existing correspondences with relevant DTD elements (e.g., in Figure 1, the *Seqdesc_title* element in *GenBank* DTD is used to extract the name attribute of the gene and the *MolInfo_biomol* value its type of molecule with the appropriate structural mapping rules).

Semantic and cognitive mapping rules are used for data unification at the instance level: several rules may use available tools for determining analogies between homologous data (such as sequence alignment, for example). The result of the *BLAST* algorithm (*Basic Local Alignment Search Tool)* implemented as a set of similarity search programs allows considering that two genomic sequences match. The nomenclature provided by our previous work on *BioMeKE (Bio-Medical Knowledge Extraction system)* reported in [29] is also considerably used to conciliate duplicate records based on several ontologies, such as the *Unified Medical Language System® (UMLS)* covering the whole biomedical domain, and the *Gene Ontology™ (GO)* that focuses on genomics and other additional terminologies, as that provided by the *HUman Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC)* to resolve synonymy conflicts. More semantic mapping rules need to be built using this information for the integration process. For example, the *Locus-ID* is used to cluster submitted sequences associated to a same gene (with cross-referenced records in *LocusLink*) and the official gene name along with its aliases to relate different gene

[16] *MEDLINE/ PubMed*: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

appearances with different names in literature for example. These aliases are also stored in the data warehouse and used to tackle the mixed or split citation problems similar to those studied by [22] in Digital Libraries.

As an illustrative example presented in Figure 2, let us now consider three distinct records we may obtain from *GenBank* by querying the DNA sequence for the human gene HFE. A first record **1** identified by the accession number *AF204869* describes a partial gene sequence (size = 3043) of the HFE gene with no annotation but one relevant information item about the position of the promoter region (*1..3043*) in the "*misc_feature"* field. A second record **2** identified by the accession number *AF184234* describes a partial sequence (size = 772) of the protein precursor of HFE gene with a detailed but incomplete annotation. The third record **3** identified by the accession number *Z92910* describes the complete gene sequence (size = 12146) of the HFE gene with a complete annotation.

We need to integrate the information and to evaluate the quality of this three records because they are complementary regarding the biological topic of interest (*i.e.*, HFE human gene): the first record has a relevant data item that the other records do not have, the second record overlaps the third one regarding the gene sequence but provide more detailed annotation and the third record is complete regarding the gene sequence. This example shows the main quality criteria we use: *i.e.* completeness, relevancy and detail level of annotation.
In this example, using the *BLAST* algorithm for determining the sequence alignment between the two sequences of the records **2** and **3** shows 100% of alignment. This indicates that the sequence in both records **2** and **3** are perfectly identical and can be merged. The detailed annotation of record **2** can be concatenated with the more complete annotation of record **3** in the data warehouse.

Several cognitive mapping rules may be used in this example for conciliating data such as the position offset: in the record **3** the fourth exon is located at position 6494 and in the record **2** this same exon is located at the relative position 130, thus using overlapping information that identifies the same entities, we can deduce the position offset and use the following cognitive rule such as:
record(AF18423)/exon[number>=4]/position = record(Z92910)/exon[number >=4]/position - 6364

3.2.3. Quality Metrics and Metadata

As we previously mentioned, we identified several information quality criteria assigned to data extracted from biological databanks. We have classified them into three sets (see Table 1 for informal definitions):

- Bio-knowledge-based quality criteria such as originality, domain authority of the authors who submitted the sequence,
- Schema-based quality criteria such as local and global completeness, level of detail, and intra- and inter-record redundancy,
- Contextual quality criteria such as freshness, and consolidation degree.

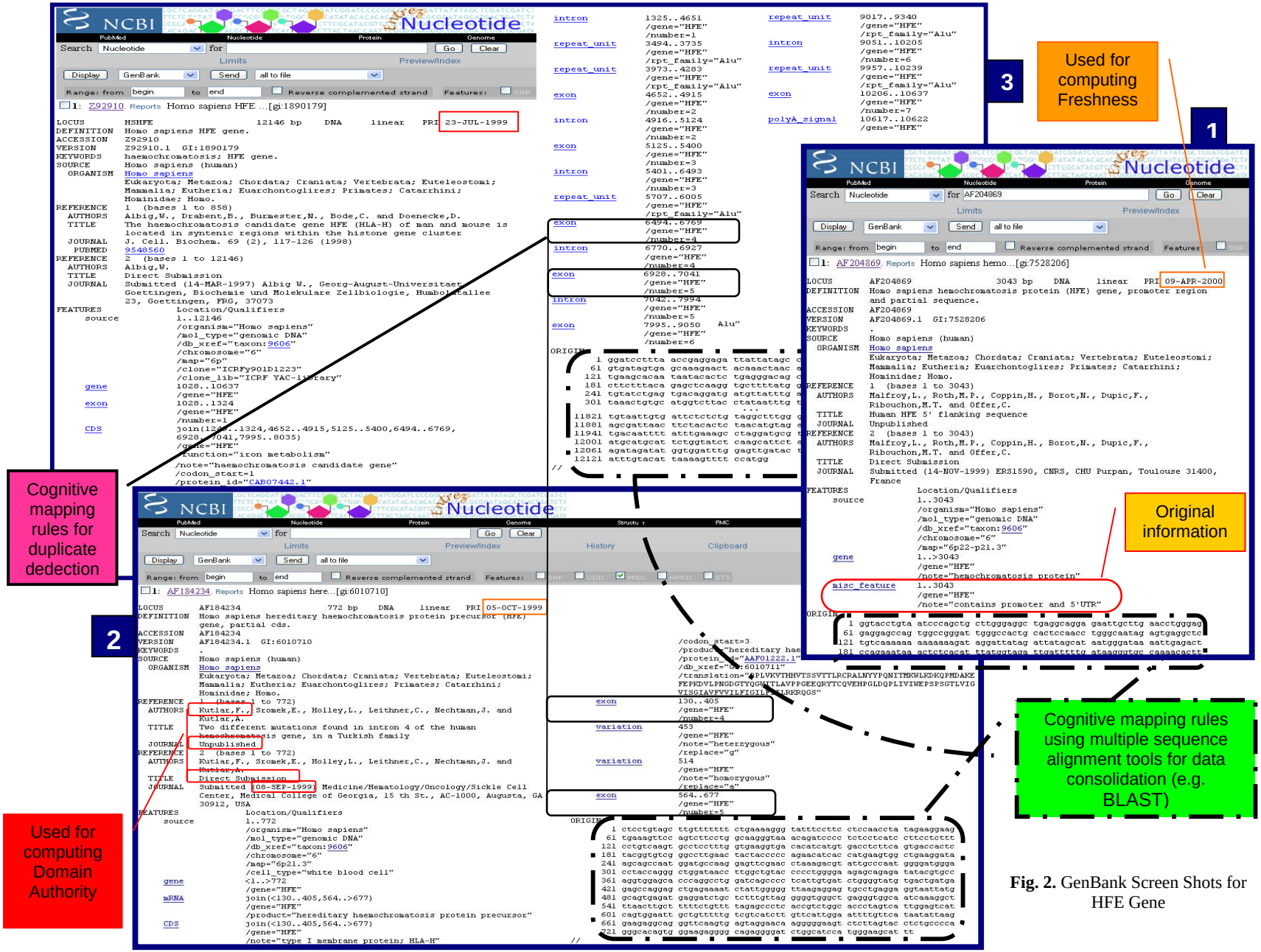

Fig. 2. GenBank Screen Shots for HFE Gene

Category	Quality Criterion	Target	Definition
Bio-Knowledge-based Quality Criteria	Originality	Data items and sub-items per record	Considering a set of records related to the same bio-entity (i.e., entity identification resolved), the originality of a data (sub-) item in a record set is defined by its occurrence frequency and its variability based on the normalized standard deviation of the edit distance between the considered strings.
	Domain Authority	Record	Domain authority is a grade in [0,1] that is computed depending on the status of the reference (*Published*, *Submitted, Unpublished*), the number of referenced submissions of the authors in the record and of the user-grade defined on the journal and authors reputations of the most recent reference of these authors.
Schema-based Quality Criteria	Local Completeness	Record	Local completeness is defined by the fraction of the number of items and sub-items with non null values on the total number of items and sub-items in the local data source schema (DTD).
	Global Completeness	Record	Global completeness is defined by the fraction of the number of items and sub-items with non null values provided by a source on the total number of items and sub-items in the global schema of the data warehouse.
	Level of Detail	Data items and sub-items per record	Level of detail is the number of sub-items per item described with non null values by a local source normalized by the total of possible sub-items in the data source schema.
	Intra-Record Redundancy	Record	Intra-record redundancy is defined by the fraction of items and sub-items in the record that are approximately the same based on the edit or q-grams distance functions or other semantic and cognitive rules
	Inter-Record Redundancy	Record Set of the same bio-entity	Inter-record redundancy is defined by the fraction of items and sub-items in the record set that are approximately the same based on edit or q-grams distance functions, BLAST or other sequence alignment techniques or other cognitive rules.
Contextual Quality Criteria	Freshness	Record	Freshness is defined by the difference between the current date and the publication date of the record
	Consolidation Degree	Data items and sub-items per record	Consolidation degree is defined by the number of inter-record redundancies and overlaps.

Table 1. *Proposed Quality Criteria for Biomedical Data*

Metadata are stored in XML files are linked to each XML record (identified by its accession number) used for data integration. Concerning the integration of genomic data, mapping are formalized and expressed by XPath and XSLT declarations. We proposed a simple XML quality metadata schema (Figure 3) that allows a flexible and extensible definition of the quality dimensions.

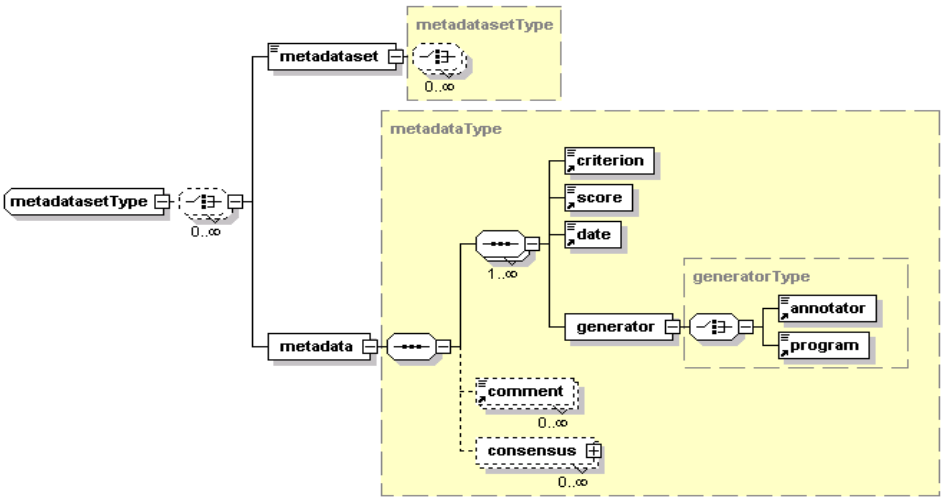

Figure 3. *Metadata XML Schema Representation*

As Figure 3 shows, the metadata type associated to a record (metadatasetType) can be a set of metadata (metadataset) or one metadata (metadata) which is composed of a criterion, a scoring value (in [0,1]) for the criterion that can be given by a human curator (annotator) or computed by a program (program), a creation date and a comment (comment). A consensus can be calculated for a given criterion and a date if several notations have been proposed by several data curators.

Considering the three records respectively identified by their accession number AF204869, Z92910 and AF184234, the originality of the sub-item *misc_feature* is 1 occurrence on 3 records with no variability,

the originality of the sub-item *KEYWORDS* is 3 occurrences on 3 records with a variability of 0.45. In the *REFERENCE* field of record Z92910, Albig *et al.* have published in *Journal of Cell. Biochem.* in 1998, and the domain authority of this record is 0.8; In the *REFERENCE* field of the record AF204869, Malfroy *et al.* have submitted a more recent internal paper in 1999 and their domain authority is 0.4. In the record AF204869, freshness of the data is 5 years and 3 months (from the 9th of April 2000). For the sake of brevity, we'll present in Figure 4 only an extract of the quality metadata file generated for the record AF204869.

```
<metadataset qid="q1_AF204869" dldref="AF204869">
<metadata mid="m1_AF204869">
        <criterion itemref="LOCUS_AF204869">Originality</criterion>
        <score>(3;1)</score>
        <date>Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>origin_pg.cc</program></generator>
</metadata>
<metadata mid="m2_AF204869">
        <criterion itemref="DEFINITION_AF204869">Originality</criterion>
        <score>(3;0.8)</score>
        <date>Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>origin_pg.cc</program></generator>
</metadata>
<metadata mid="m3_AF204869">
        <criterion itemref="ACCESSION_AF204869">Originality</criterion>
        <score>(3;1)</score>
        <date>Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>origin_pg.cc</program></generator>
</metadata>
<metadata mid="m4_AF204869">
        <criterion itemref="VERSION_AF204869">Originality</criterion>
        <score>(3;1)</score>
        <date>Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>origin_pg.cc</program></generator>
</metadata>
[---]
<metadata mid="m21_AF204869">
        <criterion itemref="RECORD_AF204869">Domain Authority</criterion>
        <score>0.45</score>
        <date>Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>dom_autho_pg1.cc</program></generator>
</metadata>
<metadata mid="m22_AF204869">
        <criterion itemref="RECORD_AF204869">Local Completeness</criterion>
        <score>0.40</score>
        <date> Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>loc_compl_pg1.cc</program></generator>
</metadata>
<metadata mid="m23_AF204869">
        <criterion itemref="RECORD_AF204869">Global Completeness</criterion>
        <score>0.6</score>
        <date> Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>glob_compl_pg2.cc</program></generator>
</metadata>
[---]
<metadata mid="m39_AF204869">
        <criterion itemref="FEATURES_AF204869">Level of Detail</criterion>
        <score>0.5</score>
        <date> Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>lev_det_pg2.cc</program></generator>
</metadata>
[---]
<metadata mid="m44_AF204869">
        <criterion itemref="RECORD_AF204869">Intra-Record Redundancy</criterion>
        <score>0.06</score>
        <date> Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>intra_rec_redun.cc</program></generator>
</metadata>
<metadata mid="m45_AF204869">
        <criterion itemref="HFE_Gene">Inter-Record Redundancy</criterion>
        <score>0.35</score>
        <date> Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>inter_rec_redun.cc </program></generator>
</metadata>
<metadata mid="m46_AF204869">
        <criterion itemref="RECORD_AF204869">Freshness</criterion>
        <score>5Y-2M-28D</score>
        <date> Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>fresh_pg.cc</program></generator>
</metadata>
[---]
<metadata mid="m66_AF204869">
        <criterion itemref="ORIGIN_AF204869">Consolidation Degree</criterion>
        <score>0.999</score>
        <date> Thu Jul 7 15:23:08 MET DST 2005</date>
        <generator><program>BLAST_call_consol_deg_pg.cc</program></generator>
</metadata>
```

Record Accession Number

Figure 4. *Example of Quality Metadata Associated to the Record AF204869*

4. A MODULAR ARCHITECTURE FOR QUALITY-PREVENTIVE INTEGRATION AND WAREHOUSING

From a functional perspective, different levels of periodic measurement and control can be implemented for ensuring a quality check-point grid upon the data warehouse system. They are represented as boxes with dotted-lines in Figure 5 from **A.** to **J.** and are different "modular" ways for implementing data quality check-points and mapping rules upon the data warehouse storage system. These modules consists of cleaning, reconciling, aggregating data and loading data into the data warehouse with appropriate *ETL* tools, record linking and mapping strategies. We used data pre- and post-validation programs before and after one-to-one mapping and massive data import in the data warehouse system (**H.** and **I.** in Figure 5). Record linking strategies previously described in Section 3.2.1 are included in the pre- and post-validation programs (in **H.** and **H'.**) before loading data. Mapping and various constraints can be implemented at different levels from the core of the data warehouse by constraints predicates, check assertions, triggers, and views with check option and stored procedures that will verify the integrity of loaded data but also update quality metadata files associated to each stored value with the identification of its source and corresponding quality. Others mapping and checking rules may be implemented at different places of the application programs: in the access module (**E.**), in the application program code (**F.**) or in the user interface (**G.**).

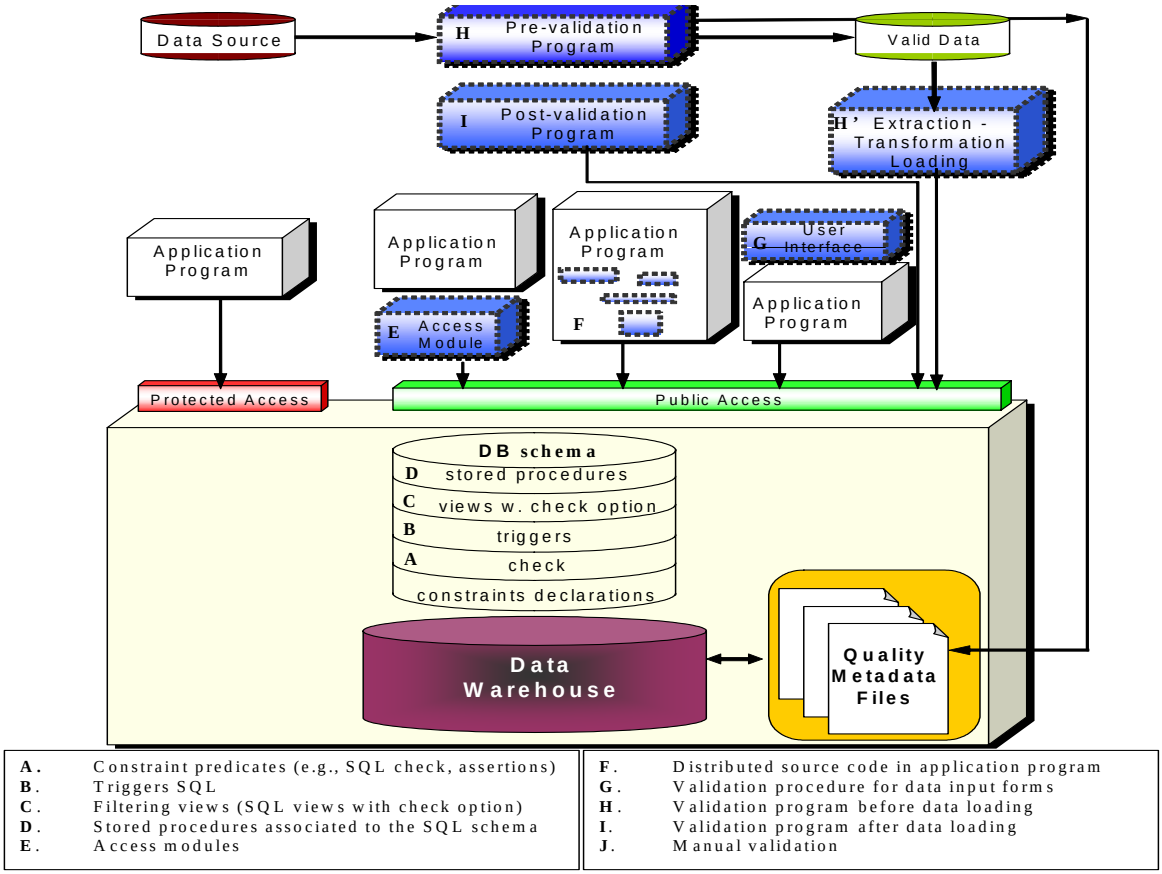

Figure 5. Different Levels for Controlling and Measuring Data Quality Before and After Loading Data in the DW

5. CONCLUSION AND PERSPECTIVES

Based on our past experience of building the biomedical data warehouse *GEDAW* (*Gene Expression Data Warehouse*) [15] [29] that stores all the relevant information on genes expressed in the liver during iron overload and liver pathologies (*i.e.*, records extracted from public databanks, data generated from DNA chips home experiments, data collected in hospitals and clinical institutions as medical records), we present some lessons learned, data quality issues in this context and current solutions we propose for quality-aware integrating and warehousing our biomedical data from a very programmatic and functional perspective. In this paper, we gave an overview of data quality related work relevant to our approach and also elements for data quality-awareness for the complex processes of integrating and warehousing biomedical data.

With regards to the limits of our warehousing approach, it is relevant as long as data integration from the heterogeneous sources in Biomedicine and their refreshment in the warehouse stay feasible automatically and with a reasonable performance. A filtering task is nevertheless performed by the expert on the delivered annotations before their storage in the warehouse by using multiple criteria, like the frequency information of the concept co-occurrences in Medline for instance. And we plan on continuing the development of *GEDAW* to extend the base of mapping and filtering cognitive rules and to complete the different levels of quality check-points previously described, which would allow us to validate the overall approach and demonstrate that the proposed quality metrics and functional architecture upon the data warehouse are respectively meaningful and really useful for our colleagues in biomedical Research.

REFERENCES

[1] Anathakrishna, R., Chaudhuri, S., Ganti, V., Eliminating Fuzzy Duplicates in Datawarehouses, *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] Bejelloun O., Garcia-Molina H., Su Q., Widom J., Swoosh: A Generic Approach to Entity Resolution, Tech. Rep., Stanford Database Group, March 2005. Available at http://dbpubs.stanford.edu/pub/2005-5

[4] Bobrowski M., Marré M., Yankelevich D., A Software Engineering View of Data Quality, *Proc. of the 2nd Int. Software Quality Week Europe (QWE'98)*, Brussels, Belgium, 1998.

[5] Bouzeghoub M., Peralta V., A Framework for Analysis of Data Freshness, *Proc. of the 1st. International Workshop on Information Quality in Information Systems (IQIS'04)*, Paris, France, 2004.

[6] Ballou D., Wang R., Pazer H., Tayi G., Modelling Information Manufacturing Systems to Determine Information Product Quality, *Management Science*, Vol. 44 (4), April 1998.

[7] Cho J., Garcia-Molina H., Synchronizing a Database to Improve Freshness, *Proc. of the 2000 ACM Intl. Conf. on Management of Data (SIGMOD'00)*, USA, 2000.

[8] Cho J., Garcia-Molina H., Estimating Frequency of Change, *ACM Trans. on Internet Technology (TOIT)*, Vol. 3 (3): 256-290, 2003.

[9] Cui Y., Widom J., Lineage Tracing for General Data Warehouse Transformation, *Proc. of Intl. Conf. on VLDB*, p. 471-480, 2001.

[10] Dasu T., Johnson T., *Exploratory Data Mining and Data Cleaning*, Wiley, 2003.

[11] 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.

[12] English L., Improving Data Warehouse and Business Information Quality, Wiley, 1998.

[13] Elfeky M.G., Verykios V.S., Elmagarmid A.K., Tailor: A Record Linkage Toolbox, *Proc. of the ICDE Conf.*, 2002.

[14] Galhardas H., Florescu D., Shasha D., and Simon E., Saita C., Declarative Data Cleaning: Language, Model and Algorithms, *Proc. of Intl. Conf. VLDB*, p.371-380, 2001.

[15] 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.

[16] Gertz M., Schmitt I., Data Integration Techniques based on Data Quality Aspects, *Proc. of the 3rd Workshop Föderierte Datenbanken*, pages 1–19, Magdeburg, December 1998.
[17] Gertz M., Managing Data Quality and Integrity in Federated Databases, *Proc. of the IFIP, Second Working Conference on Integrity and Internal Control in Information Systems*, pages 211–230, Kluwer, B.V., 1998.
[18] Hernandez M., Stolfo S., Real-world Data is Dirty: Data Cleansing and the Merge/Purge Problem, *Data Mining and Knowledge Discovery*, 2(1):9-37, 1998.
[19] Hull R., Zhou G., A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches, *Proc. of the 1996 ACM Int. Conf. on Management of Data (SIGMOD'96)*, Canada, 1996.
[20] Jarke M., Jeusfeld M. A., Quix C., Vassiliadis P., Architecture and Quality in Data Warehouses, *Proc. of Intl. Conf. CAiSE*, p. 93-113, 1998.
[21] Lacroix, Z., Critchlow, T., (Ed.), Bioinformatics: Managing Scientific Data, Morgan Kaufmann, 2003.
[22] Lee, D., Von, B.-W., Kang, J., Park, S., Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries, *Proc. of 2nd Intl. ACM Workshop on Information Quality in Information Systems (IQIS'05)*, Baltimore, USA, 2005.
[23] Low W.L., Lee M.L., Ling T.W., A Knowledge-Based Approach for Duplicate Elimination in Data Cleaning, *Information System*, Vol. 26 (8), 2001.
[24] Li W.S., Po O., Hsiung W.P., Selçuk Candan K., Agrawal D., Freshness-Driven Adaptive Caching for Dynamic Content Web Sites, *Data & Knowledge Engineering (DKE)*, Vol.47(2), 2003.
[25] Labrinidis A., Roussopoulos, N., Balancing Performance and Data Freshness in Web Database Servers, *Proc. of the 29th Int. Conf. on Very Large Data Bases (VLDB'03)*, Germany, 2003.
[26] Mihaila, G. A., Raschid, L., and Vidal, M. E., Source Selection and Ranking in the Web Semantics Architecture Using Quality of Data Metadata. *Advances in Computers*, 55, July 2001.
[27] Mannino M., Walter Z., A Framework for Data Warehouse Refresh Policies, *Technical Report CSIS-2004-001*, University of Colorado at Denver, 2004.
[28] 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.
[29] Marquet G., Burgun A., Moussouni F., Guérin E., Le Duff F., Loreal O., BioMeKe: an Ontology-Based Biomedical Knowledge Extraction System Devoted to Transcriptome Analysis, *Studies in Health Technology and Informatics*, vol. 95, p. 80-5, 2003.
[30] 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.
[31] 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.
[32] Naumann F., Leser U., Freytag J.-C., Quality-Driven Integration of Heterogeneous Information Systems, *Proc. of the 25th VLDB Conference*, Edinburgh, Scotland, 1999.
[33] Naumann F., Freytag J.C., Spiliopoulou M., Quality Driven Source Selection Using Data Envelope Analysis, *Proc. of the MIT Conf. on Information Quality (IQ'98)*, Cambridge, USA, 1998.
[34] Nie Z., Nambiar U., Vaddi S., Kambhampati S., Mining Coverage Statistics for Web source Selection in a Mediator, *Technical Report, ASU CSE TR 02-009*, 2002.
[35] Naumann F., Rolker C., Assessment Methods for Information Quality Criteria, *Proc. of the MIT Conf. on Information Quality (IQ'00)*, Cambridge, USA, 2000.
[36] Pearson R.K., Data Mining in Face of Contaminated and Incomplete Records, *Proc. of SIAM Intl. Conf. Data Mining*, 2002.
[37] Pipino L. L., Lee Y. W., Wang R. Y., Data Quality Assessment, *Communications of the ACM*, Vol. 45, No. 4, April 2002.
[38] Pyle D., Data Preparation for Data Mining, Morgan Kaufmann, 1999.
[39] Rahm E., Do H., Data Cleaning: Problems and Current Approaches, *IEEE Data Eng. Bull.* 23(4): 3-13, 2000.
[40] Redman T., Data quality: The Field Guide, Digital Press (Elsevier), 2001.
[41] Raman V., Hellerstein J. M., Potter's Wheel: an Interactive Data Cleaning System, *Proc. of Intl. Conf. VLDB*, 2001.
[42] Schafer J.L., Analysis of Incomplete Multivariate Data, Chapman & Hall, 1997.
[43] Shin B., An exploratory Investigation of System Success Factors in Data Warehousing, *Journal of the Association for Information Systems*, Vol. 4, 141-170, 2003.
[44] Scannapieco M., Virgillito A., Marchetti M., Mecella M., Baldoni R.. The DaQuinCIS Architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems, Information

Systems, Elsevier, 2004.
[45] Segev A., Weiping F., Currency-Based Updates to Distributed Materialized Views, *Proc. of the 6th Intl. Conf. on Data Engineering (ICDE'90)*, USA, 1990.
[46] Theodoratos D., Bouzeghoub M., Data Currency Quality Satisfaction in the Design of a Data Warehouse, *Special Issue on Design and Management of Data Warehouses, Int. J. Cooperative Inf. Syst.*, 10(3): 299-326, 2001.
[47] Vassiliadis P., Bouzeghoub M., Quix C. Towards Quality-Oriented Data Warehouse Usage and Evolution, *Proc. of Intl. Conf. CAiSE*, p. 164-179, 1999.
[48] Vassiliadis P., Vagena Z., Skiadopoulos S., Karayannidis N., ARKTOS: A Tool For Data Cleaning and Transformation in Data Warehouse Environments, *IEEE Data Eng. Bull.*, 23(4): 42-47, 2000.
[49] Wang R. Y., Journey to Data Quality, *Advances in Database Systems*, Vol. 23, Kluwer Academic Press, Boston, 2002.
[50] Widom J., Trio: A System for Integrated Management of Data, Accuracy, and Lineage. *Proc. of the Second Biennial Conference on Innovative Data Systems Research (CIDR '05)*, California, January 2005.