Detection and Resolution of Data Inconsistencies, and Data Integration using Data Quality Criteria

Pilar Angeles, Lachlan M. MacKinnon

Abstract — In the processes and optimization of information integration, such as query processing, query planning and hierarchical structuring of results to the user, we argue that user quality priorities, data inconsistencies and data quality differences among the participating sources have not been fully addressed. We propose the development of a Data Quality Manager (DQM) to establish communication between the process of integration of information, the user and the application, to deal with semantic heterogeneity and data quality. DQM will contain a Reference Model, a Measurement Model, and an Assessment Model to define the quality criteria, the metrics and the assessment methods. DQM will also help in query planning by considering data quality estimations to find the best combination for the execution plan. After query execution, and detection of inconsistent data, data quality might also be used to perform data inconsistency resolution. Integration and ranking of query results using quality criteria defined by the user will be an outcome of this process.

Index Terms — Data Quality, Heterogeneous Databases, Information Integration, Information Quality, Semantic Integration.

1 Introduction

The problems of data inconsistency in data integration have been widely discussed and researched for a number of years, and a large number of these have been resolved, as described in our own work [35], [36]. However, the combination of these solutions, and the resolution of the remaining issues, still remains an open issue. This has exacerbated as the development of Information Systems, network communications and the World Wide Web, has permitted widespread access to autonomous, distributed and heterogeneous data sources. An increasing number of databases, especially those published on the Web, are becoming available to external users. User requests are converted to queries over several data sources with different data quality, but the quality of the data sources utilised is not a feature of the process.

Integration of schemas on existing databases into a global unified schema is an approach developed over 20 years ago, [4]. However information quality can not be guaranteed after integration, because data quality is dependent on the design of the data and its provenance [31], [5]. Even greater levels of inconsistency exist when data is retrieved from different data sources.

On the other hand, different expectations exist on the quality of the information, depending on the user. A casual user on the Web does not expect complete and precise information [21], but close to his selection condition. A professional user expects accuracy and completeness of the information retrieved in order to make a decision irrespective of the time it could take to retrieve the data, although speed is still likely to be a lesser priority.

User priorities, data inconsistencies and data quality differences among the participating sources have not been fully addressed in the processes and optimizations of information integration, such as query processing, query planning and hierarchical structuring of results to the user.

- *P. Angeles, School of Mathematical and Computer Sciences, Heriot Watt University Edinburgh,U.K.EH14 4AS. E-mail: pilar@macs.hw.ac.uk.*
- *L.M. MacKinnon, School of Mathematical and Computer Sciences, Heriot Watt University Edinburgh,U.K.EH14 4AS. E-mail: lachlan@macs.hw.ac.uk.*

The aim of this paper is to establish the context and background on data quality for information retrieval and propose a Data Quality Manager to deal with data integration and data inconsistencies through the use of data quality properties.

This paper is organized as follows: in Section 2 the background on the establishment of data quality criteria, models and assessment is discussed. In Section 3 some issues are presented in order to help measuring data quality in Heterogeneous Databases. In Section 4 the elements of the Data Quality Manager are presented, and how it interacts with data integration and data fusion processes. Finally Section 5 concludes this paper identifying main points of this paper.

2 Background

2.1 Data Integration in Heterogeneous Database Systems

Data integration is the process of extracting and merging data from multiple heterogeneous sources to be loaded into an integrated information resource [4]. Solving structural, syntactical and semantic heterogeneities between source and target data has been a complex problem for data integration for a number of years [28], [4],[35], [36].

One solution to this problem has been developed through the use of a single global database schema that represents the integrated information with mappings from global schema to local schemas, where each query to the global schema is translated to queries to the local databases using these mappings [4]. The use of domain ontology, metadata, transformation rules, user, and system constraints have resolved the majority of the problems of domain mismatch associated with schematic integration and global schematic approaches. However, even when all the mappings, semantic and structure heterogeneity are solved in the global schema, consistency may not have been achieved, because the data provided by the sources may be mutually inconsistent. This problem has remained because it is impossible to capture all the information and avoid null values. At the same time, each autonomous component database deals with its own properties or domain constraints on data, such as accuracy, reliability, availability, timeliness and

cost of data access.

Several approaches to solve inconsistency between databases have been implemented:

1. By reconciliation of data, also known as data fusion: different values become just one using a fusion function (i.e. average, highest, and majority), depending on the data semantic [16].
2. On the basis of individual data properties: associated with each data source (i.e. cost of retrieving data, how recent is the data, level of authority associated with this source, or accuracy and completeness of data). These properties can be specified at different levels: the global schema design level, the query itself or in the profile of the user [2].

Some definitions of data quality criteria, metrics and measurement methods are presented in the following sections.

2.2 Data Quality (DQ) vs. Information Quality (IQ)

"High data quality has been defined as data that is fit for use by data consumers and is treated independent of the context in which data is produced and used" [29].

Data quality has been characterized by quality criteria or dimensions such as accuracy, completeness, consistency and timeliness [31], [16], [8], [22], [29], [25] and [20]. However there is no general agreement on data quality dimensions [32], [14].

There has not been a specific differentiation between IQ and DQ, because the terms data and information are often used synonymously. However, Data quality is related to accuracy and integrity and on the other hand, Information Quality is concern with data quality in context, and is related to how the information is produced and interpreted.

2.3 Data Quality Classifications

A definition of quality dimensions and a framework for analysis of data quality as a research area was first proposed by Richard Wang et al. [32]. An ontologically based approach was developed by Yair Wand et al. [31], this model analyzed data quality based on discrepancies between the representation mapping from real world (RW) to information system (IS) and vice versa, through design and operation activities involved in the construction of an information system as an internal view. A real world system is said to be properly represented if there exists an exhaustive mapping, and no two states in RW are mapped into the same state in IS. Four intrinsic data quality dimensions were identified: complete, unambiguous, meaningful and correct. Additionally mapping problems and data deficiency repairs were suggested. The analysis produced a classification of data quality dimensions as related to the internal or external views. Data Quality measurement method was not addressed (See table 1).

A different classification of data quality dimension was developed by Diane Strong et al. in [29] is based on a data-consumer perspective. Data quality categories were identified as intrinsic, accessibility, contextual and representational. Data quality measurement method was not addressed. Each category was directly addressed to different data quality dimensions (See table 2).

In Total Data Quality Management (TDQM) [33] the concepts, principles and procedures are presented as a methodology which defines the following life cycle: define, measure, analyze and improve data as essential activities to ensure high quality, managing data as a product. There is no focus on multi-database integration, or data inconsistency detection

TABLE 1
CLASSIFICATION BASED ON INTERNAL OR EXTERNAL VIEW [31]

	Dimensions
Internal view (design operation)	**Data- related:** Accuracy, reliability, timeliness, completeness, currency, consistency, precision **System-related:** Reliability
External view (use, value)	**Data-related:** Timeliness, relevance, content, importance, sufficiency, usableness, usefulness, clarity, conciseness, freedom of bias, informativeness, level of detail, quantitativeness, scope, interpretability, understandability **System-related:** Timeliness, flexibility, format, efficiency

TABLE 2

CLASSIFICATION BASED ON DATA-CONSUMER PERSPECTIVE [29]

DQ Category	**DQ concerns**	**Causes**	**DQ Dimensions**
Intrinsic	Mismatches among sources of the same data are common cause of intrinsic DQ concerns	Multiple sources of same data. Judgment involved in data production.	Accuracy Objectivity Believability Reputation
Accessibility	Lack of computing resources. Problems on privacy and confidentiality: Interpretability. Understandability. Data representation	Systems difficult to access. Must protect confidentiality. Representational DQ dimensions are causes of inaccessibility.	Accessibility Access Security
Contextual	Operational Data production problems: Changing data consumers needs. Distributed computing.	Incomplete data. Inconsistent representation. Inadequately defined or measured data. Data results not properly aggregated.	Relevancy Value Added Timeliness Completeness Amount of Data
Representational	Computerizing and data analyzing	Data inaccessible because: Multiple interpretations across multiple specialities and limited capacities to summarize across image.	Interpretability Ease of understanding Concise and Consistent representation Timeliness Amount of data

TABLE 3
QUALITY DIMENSIONS DEFINITIONS, DETERMINANT FACTORS AND METRICS BY AUTHOR [9], [10], [16], [25], [31].

Dimension	Concern	Author	Factors	Metric
Accuracy	"Inaccuracy implies that Information System (IS) represents a Real World (RW) state different from the one that should have been represented" "Whether the data available are the true values (correctness, precision accuracy or validity)" "The degree of correctness and precision with which real world data of interest to an application domain are represented in an information system.	Wand /Wang Motro/Rakov Gertz	RW/IS states Data values	
Precision	Ambiguity: Improper representation: multiple RW states mapped to the same IS state	Wand /Wang	RW/IS states	
Completeness	"Ability of an IS to represent every meaningful state of the represented real world system. Thus is not tied to data-related concepts such as attributes, variables, or values" "The extent to which data is not missing and does not have sufficient breadth and depth for the task at hand" "All values for a certain variable are recorded" "Whether all the data are available" "The degree to which all data relevant to an application domain have been recorded in an information system."	Wand/Wang Pipino/Wang Ballou Motro Gertz	RW/IS states Data model (table, row, attribute, classes) schema column population	1 – (#incomplete items / #total items)
Correctness	"The IS state may be mapped back into a meaningful state, the correct one" "The extend to which data is correct and reliable"	Wand/Wang Pipino/Wang	RW/IS states	*1 - (# errors / # total)*
Timeliness	"Whether the data is out of date, An availability of output on time" "The extent to which data is sufficiently up to date for the task at hand" The degree to which the recorded data are up-to-date"	Wand/Wang Pipino/Wang Gertz	Currency Volatility	*Max (0, 1 - (# currency / #volatility))*
Currency	"How fast the IS state is updated after the real world system changes." Age: of data, when first received by the system Delivery time: when data is delivered by the user Input time: When data is received by the system. "Whether the data are up to date, reflecting the most recent values"	Wand/Wang Pipino/Wang Motro	Age Delivery time Input time	Age + delivery time – input time
Volatility	"The rate of change of the real world." "Refers to the length of time data remains valid."	Wand/Wang Pipino/Wang	Time	Time data invalid - Time start valid
Consistency	"Refers to several aspects of data. In particular, to values of data inconsistency would mean that the representation mapping is one to many. This is not considered a deficiency." "The extent to which data is presented in the same format" as consistent representation "Often referred as integrity constraints state the proper relationships among different data elements" "The degree to which the data managed in an information system satisfy specified constraints and business rules."	Wand/Wang Pipino/Wang Motro Gertz	RW/IS states Values of data on Integrity constraints Data representation. Physical rep. data Values of data on Integrity constraints	1 – (#inconsistent / #total consistency checks)
Believability	"The extent to which data is regarded as true and credible"	Pipino/Wang	Source of data S Accepted stand. A Previ. experience P	Min(A,S,P)
Accessibility	"The extent to which data is available, or easily and quickly retrievable"	Pipino/Wang	Time request TR Time delivery TD Time no longer useful TN. Data path A. Structure B Path lengths C	*Max (0, 1 – (TR – TD / TR – TN))* Min (A,B,C)

or database retrieval solutions.

There are just definitions, and in the best cases, measurement of data quality aspects.

In table 3, the different quality dimension definitions are presented with the relevant factors on each dimension and the proposed metric by author.

2.4 The assessment methods for information quality criteria

Information Quality (IQ) criteria have been classified in an assessment-oriented model by F. Naumann in [20], where for each criterion an assessment method is identified.

In this classification the user, the data and the query process are considered as sources of information quality by themselves, (see Table 4.)

TABLE 4

AN ASSESSMENT-ORIENTED CLASSIFICATION [20]

Assessment Class / Source IQ of metadata	IQ Criterion	Assessment Method
Subject Criteria	Believability	User experience
	Concise represent.	User Sampling
	Interpretability	User sampling
	Relevancy	Continuous assessment
User	Reputation	User experience
	Understandability	User sampling
	Value-added	Continuous assessment
Object Criteria	Completeness	Continuous assessment
	Customer Support	Parsing, sampling
	Documentation	Parsing
	Objectivity	Expert input
Information/ Data	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert input
Process Criteria	Accuracy	Sampling, cleansing
	Amount of Data	Continuous assessment
	Availability	Continuous assessment
Query Process	Consistent repress.	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

The AIM Quality Methodology (AIMQ) [34] is a practical tool for assessing and benchmarking IQ organizations, with three components: PSP/IQ Model which presents a quality dimension classification by product quality and service quality using information consumer perspective, and consolidates the dimensions into four quadrants: sound, dependable, useful, and usable information, these quadrants are relevant to IQ improvement decisions. IQA instrument measures IQ for each IQ dimension. In a pilot study, using questionnaires answered by information collectors, information consumers, and IS professionals in six companies, these measures are average for the four quadrants and the scale used in assessing each item ranged from 0 "not at all" to 10 "completely" and the IQ Gap Analysis Techniques assess the information quality for each of the four quadrants. These gap assessments are the basis for focusing IQ improvement efforts. This methodology uses questionnaires as main measurement method, taking a very pragmatic approach regarding IQ.

In the following section we will present some approaches demonstrating how a data quality model, assessment methods and user priorities, based on the work discussed above, can help in the process of data integration.

3 MEASURING DATA QUALITY IN HETEROGENEOUS DATABASES

Database integration is divided by Motro and Rakov [16] in two main problems, intensional and extensional inconsistencies. Intensional are related to resolving the schematic differences between the component databases, this issue is also known as semantic heterogeneity. Extensional inconsistencies are related to reconciling the data differences among the participating databases [16]. Information integration is the process of merging multiple query results into a single response to the user. There are several important areas of related work to consider in the following approaches.

3.1 Data Integration Techniques based on Data Quality Aspects

Data integration techniques have been developed by Gertz [8], [9] based on data quality aspects within an object oriented data model, and data quality information stored in metadata. Quality aspects such as timeliness, accuracy and completeness were considered in the process of database integration. The main aspect was the assumption that quality of the data stored at different sites can be different and the quality varies over time. Query language extensions were necessary to support the specification of data quality goals for global queries and thus data integration. In the case of data conflicts between semantically equivalent objects, the object with best data quality must be chosen. If no conflicts exist between objects but their quality level is different, the integrated objects need to be grouped to allow the ranking of the results.

3.2 Multiplex

The project MULTIPLEX directed by Motro and Rakov [16], addressed the problem of extensional inconsistencies and a Data Quality Model for Relational Databases. MULTIPLEX was based on accuracy and completeness as quality criteria, this model assigned a quality specification for each instance of a relation, and these quality specifications were calculated by extending the relational algebra. The quality of answers was calculated by the measure of arbitrary queries from the overall quality specification of the database [16]. In the case of multiple sets of records as possible answers to one query, each set of records has an individual quality specification. A voting scheme, using probabilistic arguments, identifies the best set of records to provide a complete and sound answer and ranking of tuples in the answer space. The conflict resolution strategy, and the quality estimates are addressed by the multidatabase designer.

3.3 Fusionplex

An enhancement of the Multiplex system FUSIONPLEX [2], [3] stores information features or quality criteria scores in metadata, the considered quality dimensions are timestamp, accuracy, availability, clearance and cost of retrieval. Inconsistencies are resolved by data fusion, allowing the user to define data quality estimation on a vector of features weights, performance thresholds and a fusion function at attribute level, as required. This approach reconciles the conflicting values at attribute level using an intermediate result named polyinstance, which contains the inconsistencies. First the polyinstance is divided in polytuples, and using the feature weights and the threshold, members of each polytuple are discarded. Second each polytuple is separated into mono-attribute polytuples using the primary key, assuming that the same value of the primary key between databases refers to the same object but with different data values, and attribute values are dis-

carded based on corresponding feature values. Finally the mono-attribute tuples are joined back together resulting in single tuples.

3.4 Information Quality Reasoning

Information Quality reasoning is defined as the integration of information quality aspects, to the process of planning and optimizing queries against databases and information systems by F. Naumann in [21]. Such aspects are related through the establishment of information quality criteria, assessment methods and measure.

Selection of data sources, and optimization of query planning by considering user priorities has been also addressed in [21] by the definition of a quality model and a quality assessment method under the following assumptions:

1. Query processing: Concerned with efficiently answering a user query to a single or multi database. In this context efficiency means speed.
2. Query planning: Is concerned with finding the best possible answer given some cost or time constraint. Query planning involves regarding many query execution plans across different, autonomous sources that together form the complete result.

In this approach information sources were selected by using Data Envelopment Analysis method (DEA) [6], and the following quality dimensions: understandability, extent, availability, time and price, discarding sources with poor quality before executing the query.

However different sources have different quality scores and they must be fused to determine the best quality result, the quality fusion can be done in two ways 1) applying a fusion function per each quality criteria and find the best combination to query [17] or 2) computing the information quality score using different quality criteria such as availability, price, accuracy, completeness, amount response time for each plan and thus a ranking of the plans using Simple Additive Weighting method (SAW) explained in [11].

The completeness of the query result derived from different sources is approached in [24] considering the number of results (coverage) and the number of attribute values in the result (density). Completeness is calculated as the product between the density and the coverage of the corresponding set of information sources.

3.5 Data Quality on the Web

In this seminar, it was established that it is essential to first concentrate on developing expressive data quality models, and once such models are in place, develop tools that help users and IT managers to capture and analyze the state of data quality in an information system. [10].

4 Data Quality Manager

Databases have traditionally been considered to be sources of information that are precise and complete. However the design and implementation of such systems is carried out by human beings, whose are imperfect, so during the whole software life cycle errors occur that are reflected in the quality of both software and information. Furthermore, when these sources of data come from different applications, distributed both physically and logically, these errors multiply. In the field of Information Systems, this shortcoming has been realized and frameworks and models of reference have been developed as standards, such as ISO 15504 [12] and CMMI [1], [7].

Here, the general objective is to establish good practices for software engineering and to be able to talk the same language during software processes, no matter the architecture or implementation methodology. The same challenge need to be taken up in the Data Quality area, based on the following:

1. It is essential to identify a framework that establishes the models corresponding to the criteria of quality, methods of measurement, assessment and improvement, and considers the data quality life cycle.

This framework can be used as good practice during information system development, integration, capture and tracking of changes in data. Tracking changes should offer quality improvement and data cleaning based on a feedback provided by the same information system or a set of recommendations to the information manager, and will help to achieve self regulating systems.

2. This framework might be considered in heterogeneous systems, before, during and after the integration of information.

3. We propose a Data Quality Manager as the mechanism to establish communication between the user, the application and the process of integration of information, to deal with semantic heterogeneity problems, as part of the framework mentioned above (see Figure 1.)

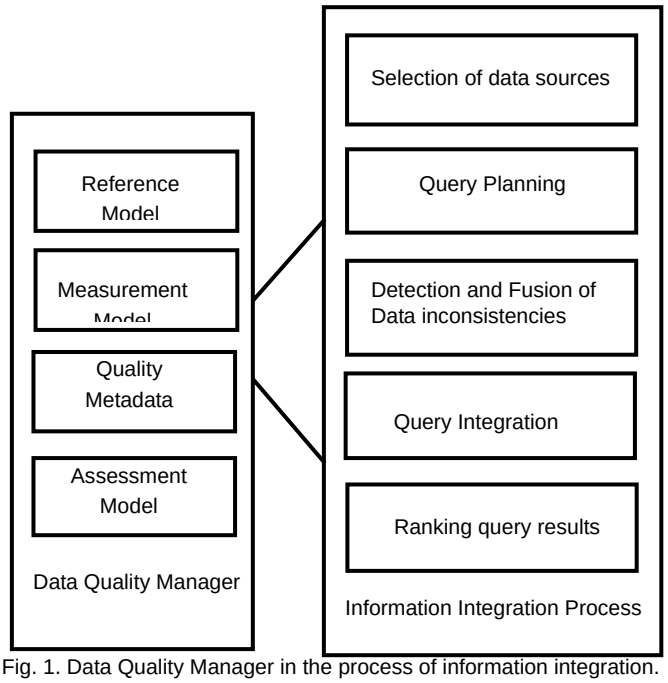

Fig. 1. Data Quality Manager in the process of information integration.

4. The Data Quality Manager will contain the following elements:

- Reference Model: In this model the data quality criteria will be defined depending on data sources, users and application domain.
- Measurement Model: This will contain the definition of the metrics to be used to measure data quality, also the definition of a quality metadata (QMD) and the specification of data quality requirements such as user profiles, query language.
- Assessment Model: The quality scores definition is essential to establish how the quality indicators are

going to be represented and interpreted.

5. The Data Quality Manager will establish the basis for taking decisions during the identification of data sources in heterogeneous systems, such that:

- To classify the sources of data based on certain criteria of quality, depending on the application domain. The scores must be stored in a metadata for every source of data (see Figure 2.)

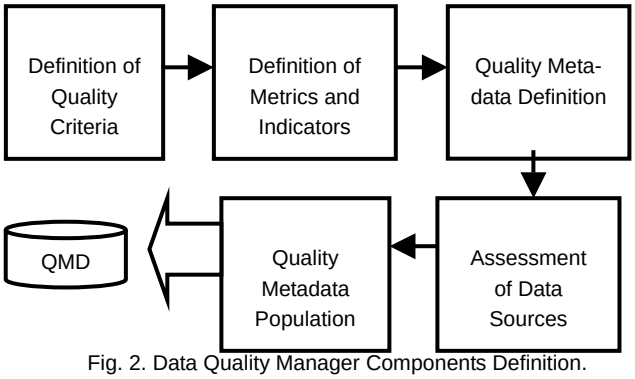

Fig. 2. Data Quality Manager Components Definition.

- The use of quality aspects previously stored in the metadata as a whole with the user priorities for the selection of the best sources of information before the execution of the queries, for example if the user prefers those sources of information that are more current with regard to those of major credibility (see Figure 3.)

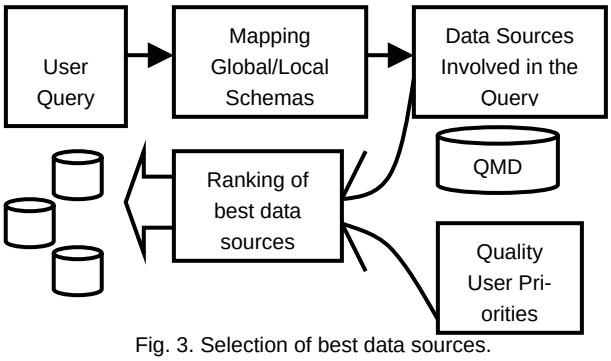

Fig. 3. Selection of best data sources.

- Help the query planning process by considering data quality estimations to find the best combination for the execution plan (see Figure 4.)

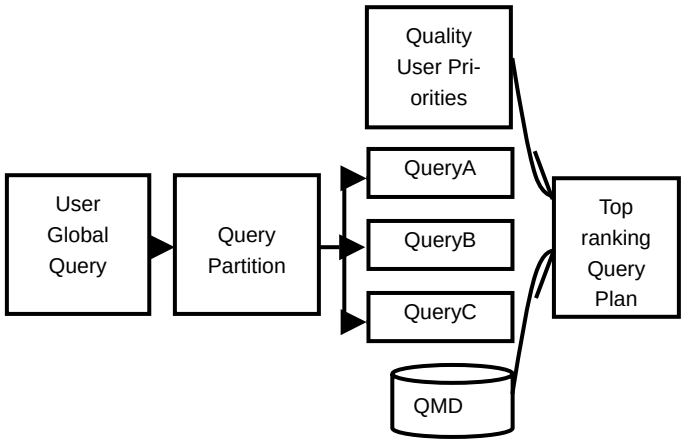

Fig. 4. Query Planning

- After query execution, and detection of inconsistent data, data quality might be used to perform data fusion (see Figure 5).

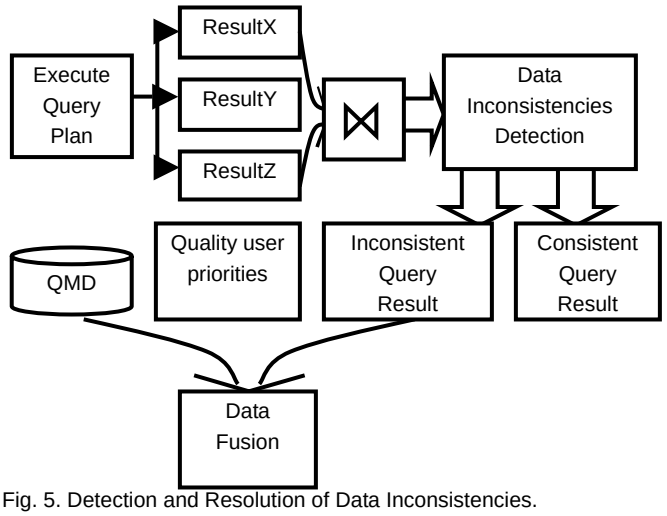

Fig. 5. Detection and Resolution of Data Inconsistencies.

- Integration of the information sources ranking with the quality criteria estimated by the user (see Figure 6)

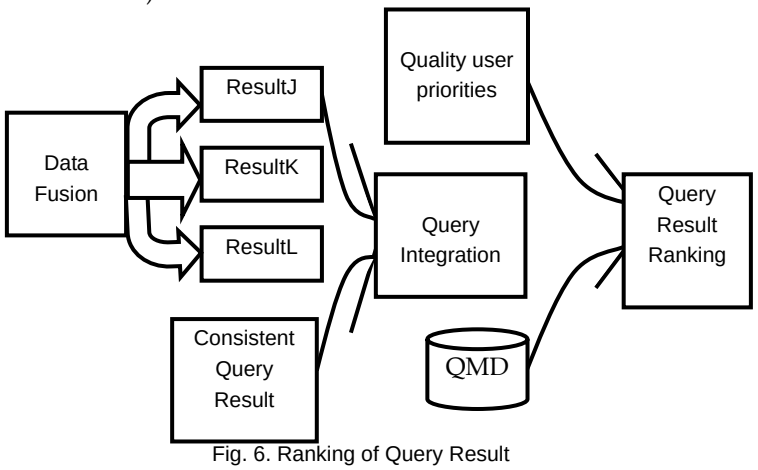

Fig. 6. Ranking of Query Result

CONCLUSION

We have shown that, although there has been considerable past work in the resolution of semantic heterogeneity in multi data source systems over a number of years, expressive data quality models and tools to utilise them remain to be developed [10]. The approach developed for Information Quality reasoning [21] provides some mechanisms for data source selection, but does not address many of the data quality factors identified in Table 3. Accordingly, we propose a Data Quality Manager as a framework to deal with data inconsistencies and lack of quality due to different sources; presenting a continuous process of data validation, such as definition of quality criteria, selection of best data sources, ranking of query plan, detection and fusion of data inconsistencies and ranking of query result considering quality of data sources and user expectations. This work is already under way and performance reporting of the tools developed will appear in the next twelve

months.

ACKNOWLEDGEMENT

This work was supported by financial funding from Consejo Nacional de Ciencia y Tecnologia CONACYT, Mexico.

REFERENCES

[1] D.M. Ahern, A. Clouse, and R. Turner, "CMMI® Distilled: A Practical Introduction to Integrated Process Improvement", *The SEI Series in Software Engineering*, Addison Wesley Professional, 2003.

[2] P. Anokhin and A. Motro, "Data Integration: Inconsistency Detection and Resolution Based on Source Properties", *Proc. of FMII 2001, 10th International Workshop on Foundations of Models for Information Integration.* Viterbo, Italy., 2001

[3] P. Anokhin and A. Motro, "Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources", Technical Report ISE-TR-03-06, Information and Software Engineering Dept., George Mason Univ., Fairfax, Virginia, 2003.

[4] C. Batini, M. Lenzerini and S.B. Navathe "A comparative Analysis of Methodologies for Database Schema Integration", *ACM Computing Surveys*, vol. 18, no. 4, pp. 323-364, 1986.

[5] P. Buneman, M. Liberman, C.J. Overton and V. Tannen, "Data Provenance", http://www.cis.upenn.edu/~wctan/DataProvenance, [(date information as accessed by the author citing the references, e.g. 17 Aug. 2004.)]

[6] A. Charnes, W. Cooper, and E. Rhodes. "Measuring the efficiency of decision making units", *European Journal of Operational Research*, pp. 429-444, 1978.

[7] M.B. Chrissis, M. Konrad and S. Shrum "CMMI®: Guidelines for Process Integration and Product Improvement", *The SEI Series in Software Engineering*, Addison Wesley Professional, 2003.

[8] M. Gertz and I. Schmitt, "Data Integration Techniques Based on Data Quality Aspects", *3rd National Workshop on Federal Databases*, Magdeburg, Germany, 1998.

[9] M. Gertz, "Managing Data Quality and Integrity in Federated Databases", *Second Annual IFIP TC-11 WG 11.5 Working Conference on Integrity and Internal Control in Information Systems*. Warrenton, Virginia, Kluwer Academic Publishers, 1998

[10] M. Gertz, "Report on the Daugstuhl Seminar, Data Quality on the Web", *SIGMOD Record*, Vol. 33, No. 1, Mar. 2004.

[11] C.L. Hwang and K. Yoon, "Multiple Attribute Decision Making: Methods and Applications: a state-of-the-art survey", Berlin; Springer-Verlag.

[12] ISO/IEC Joint Technical Committee 1 (JTC1), Subcommittee 7 (SC7) Working Group 10 (WG10) page, there are nine parts of ISO 15504. 1998.

[13] H. Kon , E. Madrick, and M. Siegel, "Good answers from bad data", Sloan WP#3868, 1995.

[14] G. Tayi, D. Ballou and Guest Editors, "Examining Data Quality", *Communications of the ACM*, vol. 41,no.2, pp.54-57, 1998.

[15] U. Leser and F. Naumann, "Query Planning with Information Quality Bounds", *Proceedings of the 4th International Conference on Flexible Query Answering, (FQAS00)*, Warsaw Poland, 2000.

[16] A. Motro and I. Rakov I, "Estimating the Quality of Databases", *Proceedings of FQAS 98: Third International Conference on Flexible Query Answering Systems*, T. Andreasen, H. Christiansen, and H.L. Larsen, ed., pp. 298-307. Roskilde, Den.mark, Springer-Verlag, Berlin, Germany, 1998.

[17] F. Naumann, "Data Fusion and Data Quality", *Proceedings of the New Techniques & Technologies for Statistics Seminar.* Surrent, Italy 1998.

[18] F. Naumann, "Quality-driven Integration of Heterogeneous Information Systems", *Proceedings of the 25th Very Large Data Bases Conference (VLDB99)*, Edinburgh, Scotland, 1999.

[19] F. Naumann and C. Roker, "Do Metadata Models meet IQ Requirements", *Proceedings of the International Conference on Information Quality*, MIT Cambridge, 1999.

[20] F. Naumann and C. Roker C., "Assessment Methods for Information Quality Criteria", *Proceedings of the International Conference on Information Quality (IQ2000)*, Cambridge, Mass., 2000.

[21] F. Naumann, "From Databases to Information Systems-Information Quality Makes the Difference", *Proceedings of the International Conference on Information Quality (IQ2001)*, Cambridge, Mass., 2001.

[22] F. Naumann, "Quality-Driven Query Answering for Integrated Information Systems", *Lecture Notes in Computer Sciences LNCS 2261*, Springer Verlag, Heidelberg, 2002.

[23] F. Naumann and M. Haeussler, "Declarative Data Merging with Conflict Resolution", *Proceedings of the International Conference on Information Quality (IQ2002)* Cambridge, Mass., 2002.

[24] F. Naumann, J. Freytag and U. Lesser, "Completeness of Information Sources", *Workshop on Data Quality in Cooperative Information Systems (DQCIS2003)*, Cambridge, Mass., 2003.

[25] L. Pipino, W.L. Yang and R. Wang, "Data Quality Assessment", *Communications of the ACM*,vol. 44 no. 4e, pp.211-218, 2002.

[26] [Parsian99] A. Parssian, S. Sumit and V. Jacob, "Assessing Data Quality for Information Products", *Proceeding of the 20th International Conference in Information Systems (ICIS1999)*, Charlotte, North Carolina USA, pp. 428-433, 1999.

[27] E. Pierce, "Assessing Data Quality with Control Matrices", *Communications of the ACM*, vol.47, no. 2, pp.82-86, 2004.

[28] A. Sheth and L. Larson, "Federated Database Systems for Managing Distributed Heterogeneous and Autonomous Databases", *ACM Computing Surveys*, vol. 22, no. 3, pp.184-236, 1990.

[29] D.M. Strong, W.L. Yang and R.Y. Wang, "Data Quality in Context", *Communications of the ACM*, vol. 40, no. 5, pp.103-110, 1997.

[30] D.M. Strong, W.L. Yang and R.Y. Wang, "10 Potholes in the Road to Information Quality", *Proceedings of IEEE*, vol.18, no. 9162, pp.38-46, 1997.

[31] Y. Wand and R. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations", *Communications of the ACM*, vol. 39, no. 11, pp.86-95, 1996.

[32] R.Y. Wang, V.C. Storey, and C.P. Firth, "A Framework for Analysis of Data Quality Research," *IEEE Trans. Knowledge and Data Eng.*,1995.

[33] R. Wang, "A Product Perspective on Total Data Quality Management", *Communications of the ACM*, vol. 41, no. 2, pp.58-65, 1998.

[34] L. Yang, D. Strong and R. Wang, "AIMQ: A Methodology for Information Quality Assessment", *Information and Management,* vol. 40, no. 2, pp. 133-146, 2002.

[35] L.M. MacKinnon, D.H. Marwick, H. Williams, "A Model for Query Decomposition and Answer Construction in Heterogeneous Database Systems"., *Journal of Intelligent Information Systems*, 1998.

[36] H. Williams, H.T. El-Khatib, L.M. MacKinnon, "A framework and test-suite for assessing approaches to resolving heterogeneity", Information and Software Technology, 2000.

Pilar Angeles obtained her first degree in computer engineering from the Universidad Nacional Autonoma de Mexico (UNAM), in 1993,
Diploma in Expert Systems from The Instituto Tecnologico Autonomo de Mexico (ITAM) in 1994, Diploma in Telematic Systems from ITAM in 1995, and M.Sc. in Computer Science, regarding Quality in Software Enginnering from the UNAM in 2000. Since 1989 she has been working on Technical Support for Databases in Casa de Bolsa Probursa, Nissan Mexicana, Software AG, Sybase de Mexico and e-Strategy Mexico. Recent research interests have included Data Quality and Heterogeneous Databases. She is a Funder Member of the "Quality in Software Engineering Mexican Association" (AMCIS).

Lachlan M. MacKinnon is Reader in Computer Science, and Director of Postgraduate Study in Computer Science, at Heriot-Watt University. He has a first degree in Computer Science, and a PhD in Intelligent Querying for Heterogeneous Databases. He researches and consults widely in Data, Information and Knowledge Technologies. He is a member of the IEEE, British Computer Society, ACM, AACE, immediate past Chair of the British

National Conference on Databases, and upcoming Chair of the British HCI Conference. He has over 50 conference and journal publications in this area.