Detection and Resolution of Data Inconsistencies, and Data Integration using Information Quality criteria.

Maria del Pilar Angeles

School of Mathematical and Computer Sciences, Heriot-Watt University,
Edinburgh,EH14 4AS
pilar@macs.hw.ac.uk

Lachlan M. Mackinnon

School of Mathematical and Computer Sciences, Heriot-Watt University,
Edinburgh,EH14 4AS
lachlan@macs.hw.ac.uk

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. 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.

Keywords: Databases, Heterogeneous, Data Quality, Information Quality, Semantic Integration, Information Integration.

Introduction

The development of Information Systems, network communications and the World Wide Web, has permitted 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 information quality.

Integration of schemas on existing databases into a global unified schema is an approach developed over 20 years ago, [BATINI86]. However information quality can not be guaranteed after integration, because data quality is dependent on the design of the information and its provenance [Wang96], [Bunemann98]. Even greater levels of inconsistency exist when information is retrieved from different information 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, but what is the available information in the shorter possible time.
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.

The aim of this paper is to establish the context and background on data quality for information retrieval from distributed heterogeneous systems with regard to the following issues:

- How the information quality has been modelled and measured.
- Which quality dimensions are more useful for query processing.
- Which quality criteria have been more determinant as user priorities for classification of the final result set.
- Discussion on how these approaches have been developed, and highlight new topics for further research.

This paper is organized as follows: in Section 2 the background on the establishment of information quality criteria, models and assessment is discussed. In Section 3 some issues are presented in order to help query answering and ranking of query results using priorities given by the user. Section 4 concludes this paper identifying further work to be carried out.

2 Background

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. Solving structural, syntactical and semantic heterogeneities between source and target data has been a complex problem for data integration for a number of years.

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.

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 information 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 information, such as accuracy, reliability, availability, timeliness and cost of information access.

Several approaches to solve inconsistency between databases have been implemented:

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

In the following section definitions of data quality models and some information quality measurement methods are presented.

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" [Strong97a]

Data quality has been characterized by quality criteria or dimensions such as accuracy, completeness, consistency and timeliness. [Wand96],[Motro98],[Gertz98a],[Naumann02], [Wand96],[Strong97],[Pipino02],[Naumann00].However there is no general agreement on data quality dimensions. [Wang95].

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. [Wang95].

An ontologically based approach was developed by Yair Wand et. al [Wand96], 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.)

	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 1: Data quality dimensions as related to the internal or external views [Wand96]

A different classification of data quality dimension was developed by Diane Strong et.al. [Strong97] 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.)

DQ Category	DQ concerns	Causes	DQ Dimensions
Intrinsic DQ	-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 DQ	Lack of computing resources Problems on privacy and confidentiality: Interpretability. Understandability. Data representation.	Systems difficult to access. Must protect confidentiality. Representational DQ dimensions are underlying causes of accessibility DQ problem.	Accessibility, Access Security
Contextual DQ	Operational Data production problems: Changing data consumers' needs. Distributed computing.	Incomplete data. Inconsistent representation. Inadequately defined or measured data. Data results that could not be properly aggregated.	Relevancy, Value Added, Timeliness, Completeness, Amount of Data
Representational DQ	Computerizing and data analyzing	Computerized data inaccessible because: Multiple specialists are needed to interpret data across multiple specialities. Limited capacities to summarize across image and text data.	Interpretability, Ease of understanding, Concise and Consistent representation Timeliness Amount of data

Table 2: Data quality classification based on data-consumer perspective [Strong97]

In the Total Data Quality Management (TDQM) [Wang98] are presented the concepts, principles and procedures as a methodology who defines the following life cycle: define, measure, analyze an improve information product as essential activities to ensure high quality, managing information as a product. As has been shown, there is no focus on multi database integration, nor data inconsistency detection nor database retrieval solutions. There are just definitions, and in the best cases, measurement of data quality aspects.

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

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 extend to which data is not missing and is not 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) a) schema b) column c) 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 extend 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 The time the data is actually used. Currency Volatility Sensitivity factor (subjective 0-1)	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 stop valid - 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 extend 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 More than one state of the IS matching a state of the real sys. 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 standard A Previous 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)
Appropriate amount of data	"The extent to which the volume of data is appropriate for the task at hand. Quantity being neither too little nor too much"	Pipino/Wang	#provided data pd #needed data nd	Min(pd/nd,nd/pd)

Table 3: Quality dimensions definitions, determinant factors and metrics by author.

The assessment methods for information quality criteria

Information Quality criteria has been classified in an assessment-oriented model [Naumann00], 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.)

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

Table 4: Classification of Data quality based on assessment class and source of metadata

The AIM Quality Methodology (AIMQ) [Yang02] 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.

In the following section we will present some approaches demonstrating how an information quality model, assessment methods and user priorities can help in the process of data integration.

3 Measuring Data Quality in Heterogeneous Databases

Database integration is divided 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. [Motro98]. 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.

- Data integration techniques have been developed based on data quality aspects [Gertz98a][Gertz98b] 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.

- The project MULTIPLEX [Motro98] 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. 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 a ranking of tuples in the answer space. The conflict resolution strategy, and the quality estimates are addressed by the multi database designer.

- An enhancement of the Multiplex system FUSIONPLEX [Anokhin01],[Anokhin03] 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 discarded based on corresponding feature values. Finally the mono-attribute tuples are joined back together resulting in single tuples.

- Information Quality Reasoning: Selection of data sources, and optimization of query planning by considering user priorities has been also addressed in [Naumann98],[Naumann99],[Naumann00] by the definition of a quality model and a quality assessment method under the following assumptions:

  - Query processing: Concerned with efficiently answering a user query to a single or multi database. In this context efficiency means speed.

  - 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.

  - 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. Such aspects are related through the establishment of information quality criteria, assessment methods and measure.

Commonly the information sources on the web are classified by counting the appearances of certain words and using statistics to determine the most "relevant sources".

In this approach information sources were selected by using Data Envelopment Analysis method (DEA)[Charnes78] 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:
a) Applying a fusion function per each quality criteria and find the best combination to query [Nauman98] such as fused time as the maximum of the individual scores or fused process as the sum of the individual prices.
b) 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),[Hwang81].
The completeness of the query result derived from different sources is approached in [Naumann03] 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.

In the Seminar "Data Quality on the Web" [Gertz04], it was established that it is essential to first concentrate on developing expressive data quality models, 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.

The main characteristics of each approach are presented in Table 5.

Quality Dimensions considered	Scope	Main characteristics.
Motro: Timeliness, Accuracy, cost, availability	Resolving Inconsistencies and using IQ features for the source data. Attribute Level of granularity	Metadata features: cost, accuracy availability, clearance with values from 0 to 1. SQL extended: USING for features, and WITH for weights on features. NULLS: restrictive (all comparations to nulls evaluate to false) Polyinstance (union of nonempty query fragments) Polytuple (clustering polyinstance using the key with just one attribute) Fusion by any, avg, etc. Joining the monoattribute tuples in a single tuple
Gertz: Accuracy, completeness, timeliness, consistency. Problems at integration level are rather data quality problems that data integrity, problems like outdated/accuracy/completeness There is no data integrity concept that covers these aspects during database integration. Data quality <> Data integrity.	Specific data quality goals.	Metadata : Information profiles where each quality statement contains: --temporal condition associated --Def. each constraint at global relation level --Integrity constraint in case of conflict between component databases. SQL Extended (with goal) The query processor decomposes global queries in sub queries such that the retrieved data all have the same quality Correctness and completeness can be considered as integrity constraints imposed on the metadata.
Naumann: Completeness, timeliness, accuracy, reliability	Sources selection by understandability, extent, availability, time and price criteria using DEA method. Ranking query plans using SAW method and completeness as quality criterion	Mediator representing the semantic knowledge Query Correspondence Assertions (QCA): Which are set-oriented equations between queries against one wrapper (local db level) and queries against one mediator. For a given user query against the mediator schema the mediator tries to find combinations (query plans) of QCAs that are semantically contained. Information sources and query plans achieve I! scores in each criterion.

Table 5: Different Data Quality/Data Integration approaches.

4 Conclusions

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, who 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 information come from different applications, distributed both physically and logically these errors multiply. In the field of Information Systems, this shortcoming has been realized and there have been developed frameworks and models of reference as standards, such as ISO 15504 [ISO/IEC TR 15504-98] and CMMI [Ahern03].

Here, the general objective is to establish good practices for software engineering and to be able to talk the same language during software processes, without importing 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 would propose a Data Quality Manager as the manager to establish communication between the process of integration of information, the user and the application, to deal with semantic heterogeneity problems, as part of the above mentioned framework.(see Figure 1.)

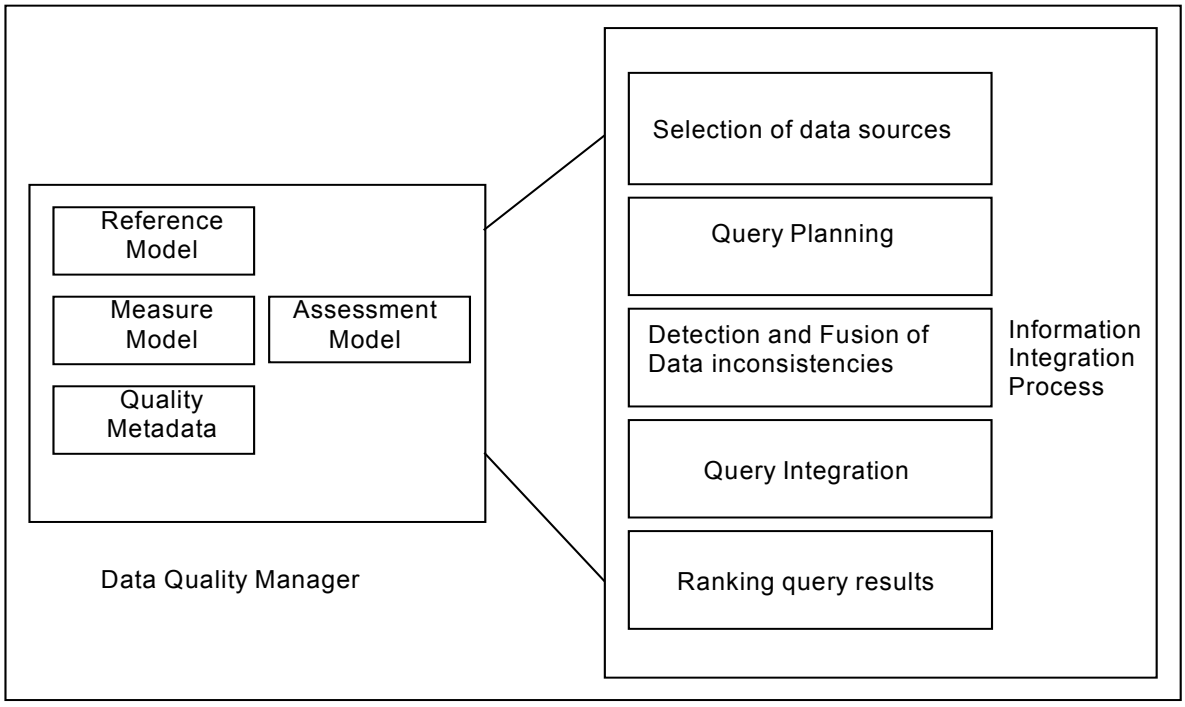

Figure 1: Data Quality Manager in the process of information integration.

4.- The Data Quality Manager will contain the following elements:
Reference Model: In this model will be defined the data quality criteria depending on data sources, users and application domain.
Measurement Model: It will contain the definition of the metrics to be used to measure data quality, also the definition of the data quality metadata 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 by the user and the application.
5.- The Data Quality Manager will establish the basis for taking decisions during the identification of information sources in heterogeneous systems, such that:

- Based on certain criteria of quality, to classify the sources of information depending on the application domain. This must be stored in metadata for every source of information (see Figure 1.A.)
- 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 1.B.)
- The query planning, considering data quality estimations to find the best combination for the execution plan (see Figure 1.C.)
- After the query execution, and detection of inconsistent data, data quality might be used to perform data fusion.
- Integration of the information sources ranking with the quality criteria estimated by the user (see Figure 1.D.)

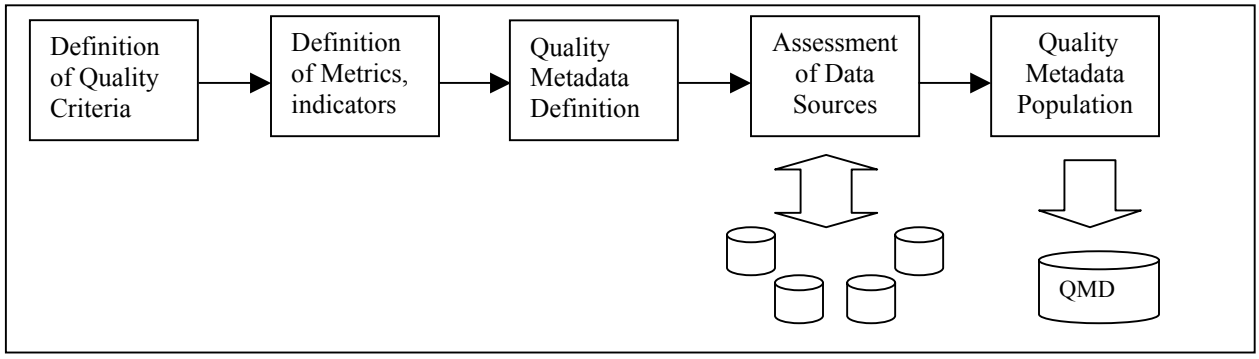

Figure 1.A: Data Quality Manager Components Definition

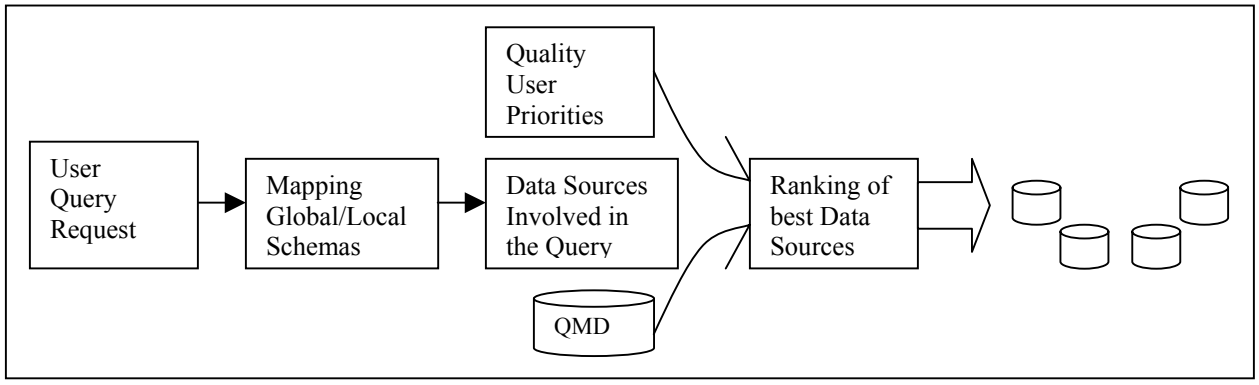

Figure 1.B: Selection of best data sources

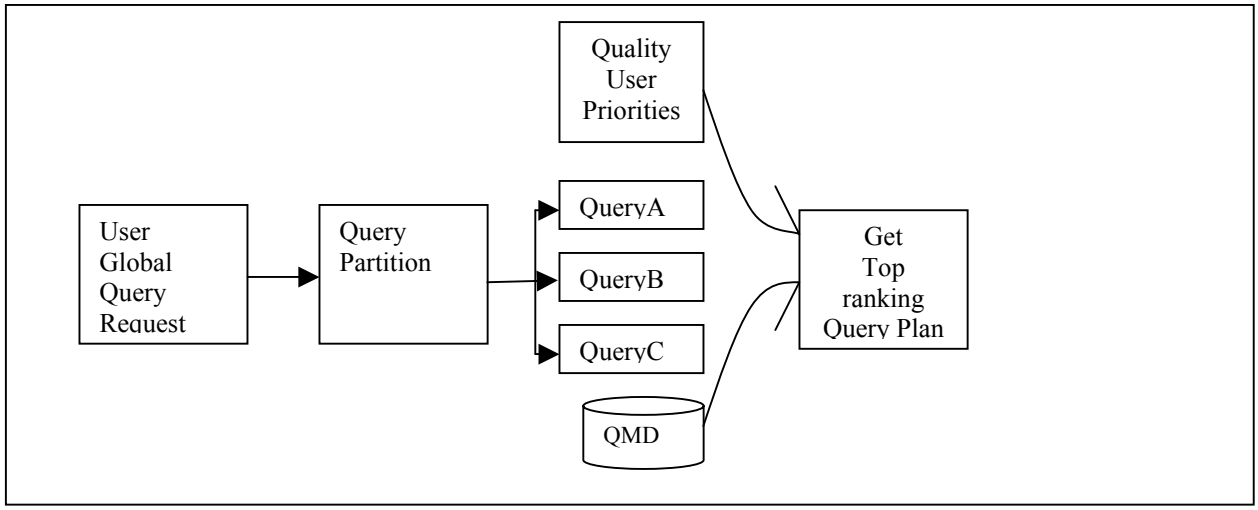

Figure 1.C: Query Planning

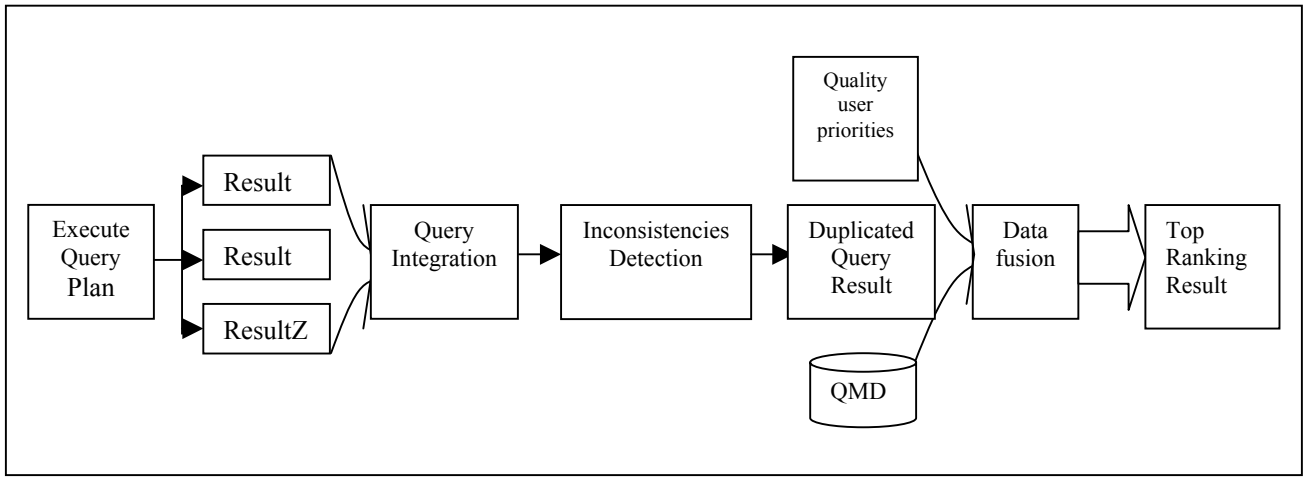

Figure 1.D: Ranking Query Results

References

[Anokhin01]
Anokhin P. and Motro A. (2001) "Data Integration: Inconsistency Detection and Resolution Based on Source Properties"
Proceedings of FMII 2001, 10th International Workshop on Foundations of Models for Information Integration,Viterbo, Italy

[Ahern03]
Ahern D.M., Clouse A., and Turner R.(2003) ”CMMI® Distilled: A Practical Introduction to Integrated Process Improvement”
Addison Wesley Professional, The SEI Series in Software Engineering.

[Anokhin03]
Anokhin P. and Motro A. (2003) "Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources"
Department of Information and Software Engineering, George Mason University, Technical Report ISE-TR-03-06

[Batini86]
Batini C., LenzeriniM. and Navathe S.B. (1986) "A comparative Analysis of Methodologies for Database Schema Integration"
ACM Computing Surveys18(4),pp.323-364

[Charnes78]
Charnes A., Cooper W., and Rhodes E. (1978) “Measuring the efficiency of decision making units”. European Journal of Operational Research , 429-444.

[Chrissis03]
Chrissis M.B., Konrad M. and Shrum S.(2003) “CMMI®: Guidelines for Process Integration and Product Improvement”
Addison Wesley Professional., The SEI Series in Software Engineering.

[Gertz98a]
Gertz M. and Schmitt I. (1998) "Data Integration Techniques based on Data Quality Aspects"
3rd National Workshop on Federal Databases, Magdeburg Germany

[Gertz98b]
Gertz M. (1998) "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

[Gertz04]
Gertz M. (2004) "Report on the Daugstuhl Seminar, Data Quality on the Web"
SIGMOD Record, Vol. 33, No. 1 March, 2004.

[Hwang81]
Hwang C.-L. and Yoon K. (1981) "Multiple Attribute Decision Making: methods and applications: a state-of-the-art survey"
Berlin ; Springer-Verlag.

[ISO/IEC TR 15504-98]
ISO/IEC Joint Technical Committee 1 (JTC1), Subcommittee 7 (SC7) Working Group 10 (WG10) page, there are nine parts of ISO 15504. 1998.

[Kon95]
Kon H.,Madrick, E. and Siegel Michael (1995) "Good answers from bad data"
Sloan WP#3868

[Kumar98]
Tayi G. Ballou D. and Guest Editors (1998) "Examining Data Quality"
Communications of the ACM,41(2),pp.54-57

[Leser00]
Leser U. and Naumann F. (2000) "Query Planning with Information Quality Bounds"
Proceedings of the 4th International Conference on Flexible Query Answering (FQAS00), Warsaw Poland

[Motro98]
Motro A. and Rakov I.(1998)"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, Editors), Roskilde, Denmark, Springer-Verlag, Berlin, Germany, pp. 298-307.

[Naumann98]
Naumann F. (1998) "Data Fusion and Data Quality"
Proceedings of the New Techniques & Technologies for Statistics Seminar, Surrent Italy

[Naumann99a]
Naumann F. (1999) "Quality-driven Integration of Heterogeneous Information Systems"
Proceedings of the 25th Very Large Data Bases Conference (VLDB99), Edinburgh, Scotland.

[Naumann99b]
Naumann F. and Roker C. "Do Metadata Models meet IQ Requirements"
Proceedings of the International Conference on Information Quality MIT Cambridge

[Naumann00]
Naumann F. and Roker C. (2000) "Assessment Methods for Information Quality Criteria"
Proceedings of the International Conference on Information Quality (IQ2000)Cambridge, MA

[Naumann01]
Naumann F. (2001) "From Databases to Information Systems-Information Quality Makes the Difference"
Proceedings of the International Conference on Information Quality (IQ2001)Cambridge, MA

[Naumann02a]
Naumann F. "Quality-Driven Query Answering for Integrated Information Systems"
Lecture Notes in Computer Sciences LNCS 2261, Springer Verlag, Heidelberg, 2002.

[Naumann02b]
Naumann F. and Haeussler M.(2002) "Declarative Data Merging with Conflict Resolution"
Proceedings of the International Conference on Information Quality (IQ2002) Cambridge, MA

[Naumann03]
Naumann F., Freytag J. and Leser U. (2003) "Completeness of Information Sources"
Workshop on Data Quality in Cooperative Information Systems (DQCIS2003) Cambridge, MA

[Pipino02]
Pipino L., Yang W. Lee and Richard Wang (2002)"Data Quality Assessment"
Communications of the ACM,44(4e),pp.211-218

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

[Pierce04]
Pierce E.(2004) "Assessing data quality with control matrices"
Communications of the ACM,47(2),pp.82-86

[Sheth90]
Sheth A. and Larson J. (1990) "Federated Database Systems for Managing Distributed Heterogeneous and Autonomous Databases"
ACM Computing Surveys 22(3),pp.184-236

[Strong97a]
Diane M. Strong, Yang W. Lee and Richard Y. Wang (1997)"Data Quality in Context"
Communications of the ACM,40(5),pp.103-110

[Strong97b]
Diane M. Strong, Yang W. Lee and Richard Y. Wang (1997) "10 Potholes in the Road to Information Quality"
Proceedings of IEEE,V.18(9162),pp.38-46

[Wand96]
Yair Wand and Richard Wang (1996) "Anchoring data quality dimensions in ontological foundations"
Communications of the ACM,39(11),pp.86-95

[Wang98]
Wang R. (1998)"A Product Perspective on Total Data Quality Management"
Communications of the ACM,41(2),pp.58-65

[Yang02]
Yang L., Strong D. and Wang R.(2002) "AIMQ: a methodology for information quality assessment"
Information and Management 40(2) pp. 133-146