# Information Quality Criteria for Peer Data Management Systems

Crishane Freire
Federal University of Pernambuco
50.732-970 Recife, PE, Brazil
+55 81 2126 8430
caf4@cin.ufpe.br

Bruno F. F. Souza
Federal University of Pernambuco
50.732-970 Recife, PE, Brazil
+55 81 2126 8430
bffs@cin.ufpe.br

Damires Souza
Federal Institute of Education,
Science and Technology of Paraiba
50.732-970 João Pessoa, PB Brazil
+55 81 3453 9213
dysf@cin.ufpe.br

Maria C. M. Batista
Federal Rural University of Pernambuco
52171-900 Recife, PE, Brazil
+55 81 2126 8430
ceca@deinfo.ufrpe.br

Ana Carolina Salgado
Federal University of Pernambuco
50.732-970 Recife, PE, Brazil
+55 81 2126 8430
acs@cin.ufpe.br

## ABSTRACT

Peer Data Management Systems (PDMSs) have been considered a natural extension to data integration systems. Although much work has been accomplished on PDMSs processes such as peer clustering and query answering, PDMSs still suffer from incomplete information quality control mechanisms to enhance these processes. We argue that Information Quality (IQ) may be a relevant discriminator in PDMSs environments. To help matters, in this work, we provide a review of IQ research related to PDMSs. To this end, we introduce an overview of IQ, describe existing PDMSs approaches and propose a set of IQ criteria to be used in PDMSs processes.

## 1. INTRODUCTION

Peer Data Management Systems (PDMS) came into the focus of research as a natural extension to distributed databases in the peer-to-peer (P2P) setting [11, 26, 44]. PDMSs are considered the result of blending the benefits of P2P networks, such as lack of a centralized control, with the richer semantics of a database [22]. They can be used for data exchanging, query answering and information sharing. For instance, in the areas of scientific research, the idea of setting up a PDMS to share research data among peers has already been widely discussed [11, 22].

A PDMS consists of a set of inter-related peers (data sources). Each peer has an associated schema within a domain of interest. However, PDMSs do not consider a single global schema. Instead, each peer represents an autonomous data source and exports either its entire data schema or only a portion of it. Such schema, named exported schema, represents the data to be shared with the other peers of the system. Between the exported schemas of two neighbor peers, mappings are generated. In general, queries submitted at a peer are answered with data residing at that peer and with data that is reached through mappings that are propagated over the network of peers.

Despite the significant amount of work in the development of their services (e.g., peer clustering), PDMSs still suffer from inadequate information quality control mechanisms to address, for instance, the management of quality of the data that are used and obtained as query answers as well as the quality of the mappings between peer schemas. Due to PDMS dynamic nature, IQ metrics should be evaluated on the fly. The objective of embodying IQ analysis in a PDMS is to provide improvements over the system's processes. As an illustration, processes such as query answering and peer clustering may be enhanced. .

In this sense, IQ may be a relevant discriminator in PDMSs environments. IQ is usually characterized via multiple dimensions or criteria, each of which captures a high-level aspect of quality. The role of each one is to assess and measure a specific IQ aspect [30]. Thus, quality metrics are used to measure a particular quality criterion [4].

As in any information system integrating data from autonomous and distributed sources, PDMSs are vulnerable to poor IQ in some aspects such as [5, 33, 44]: peer (data source), peer schema (or its representational model), mappings, data and query answers. For instance, regarding peers, their sources themselves might store data of poor quality or have a bad reputation. Regarding mappings, they may be considered as incorrect or incomplete or even not confident. In addition, a peer schema representation may be not sufficiently consistent or may be not considered as minimal (without redundancies) to the original source schema.

One possible usage of IQ criteria in PDMSs should be the assistance in query routing strategies. This can be done by considering IQ measures to find the best path to route queries through the network of neighbor peers. Furthermore, after query reformulation and execution, IQ might be used to enhance integration and ranking of query answer results.

The aim of this paper is to establish the context and background on IQ for PDMSs with regard to the following issues:


*VLDB '11*, August 24-30, 2008, Auckland, New Zealand.

- How the information quality has been classified and measured in general integration systems.
- How the information quality has been classified and measured in PDMSs.
- Which IQ criteria may be useful for PDMSs processes.
- Discussion on how these IQ criteria may be defined in order to be applied in PDMSs.

This paper is organized as follows: Section 2 explains the main aspects of IQ; Section 3 discusses IQ in the light of data integration; Section 4 discusses IQ in PDMSs; Section 5 indicates some IQ criteria chosen to be applied in PDMS. Finally, Section 6 points out some considerations and highlights important topics for further research.

## 2. INFORMATION QUALITY

IQ has become a critical aspect in organizations and, consequently, in Information Systems research [8, 32]. The notion of IQ has emerged during the past years and shows a steadily increasing interest. IQ is a multidimensional aspect and it is based on a set of dimensions or criteria. The role of each one is to assess and measure a specific IQ aspect [37, 41, 49]. All these IQ works assume that there exist some shared norms of quality, or quality expectations, and the ways of measuring the extent of meeting those norms and expectations. For our purposes, however, we will use the general definition of IQ – '*fitness for use*' - which encompasses the aspects of quality.

Following the definition of Keeney and Raiffa [39], the *measurability* of an IQ dimension is defined as the ability to assess the variation along a dimension within a reasonable cost. *Measuring* is defined as the process of mapping the attribute-level distributions of real-world entities to score values in an objective and systematic way. Accordingly, a *measure* is defined as a relation associating the attribute-level distributions of real-world entities or processes with numbers. We define a *measurement* as a score value characterizing a particular IQ attribute or criteria in an objective way.

It is important to distinguish the two concepts of Data Quality and Information Quality. Information quality (IQ) is a term to describe the quality of any element or content of information systems [41], not only the data. IQ assurance is the certainty that particular information meets some quality requirements. This leads us to think in a service-based perspective of quality which focuses on the information consumer's response to his/her task-based interactions with the information system. The use of the term *information* rather than *data* implies that the *use* and *delivery* of the data must be considered in any quality judgements, i.e. the quality of delivered data represents its *value* to information consumers [40]. Thus, we use the definition of *Information Quality* as a set of criteria to indicate the overall quality degree associated with the information in the system [27]. The term *Data Quality* is similar to the *accuracy* IQ criterion, i.e., only one characteristic or aspect of the Information Quality broader concept [25].

One of the best known quality dimensions classification is presented by Wang and Strong in [41]. They have conceived one of the first set of structured and classified quality dimensions which has been a strong reference for most of the studies later in this area. They empirically identified fifteen criteria. In order to achieve this, various attributes of quality were analyzed from the users' perspective. These were further grouped into four broad IQ classes: *intrinsic*, *contextual*, *representational*, and *accessibility*. The classes and their dimensions are explained as follows.

*Intrinsic Data Quality* denotes the quality of data itself. In this category, the following criteria are included:

- Believability: extent to which data is regarded as true and credible.
- Accuracy: the degree to which data are correct, reliable and certified free of error.
- Objectivity: the extent to which data are unbiased (unprejudiced) and impartial.
- Reputation: extent to which data are trusted or highly regarded in terms of their source or content.

*Contextual Data Quality* highlights the requirement that data quality must be considered within the context of a task at hand, i.e., data must be relevant, timely, complete and appropriate in terms of amount to add value. Thus, this category aggregates the following dimensions:

- Value-added: the degree to which data are beneficial and provide advantage from their use.
- Relevancy: the degree to which data are applicable and helpful for the task at hand.
- Timeliness: the extent to which the age of the data is appropriate for users needs.
- Completeness: the degree to which data are of sufficient breadth, depth and scope.
- Appropriate Amount of Data: the extent to which the data volume is appropriate.

The *Representational Data Quality* category is related to the format and the meaning of data. It includes the following dimensions:

- Interpretability: the extent to which data are in appropriate language and units and data definitions are clear.
- Ease of Understanding: the extent to which data are clear and without ambiguity and easily comprehended.
- Concise Representation: the extent to which data are compactly represented without being overwhelming.
- Consistent Representation: the degree to which data are always presented in the same format and are compatible with previous data.

*Accessibility Data Quality* defines if data are available or obtainable. It is concerned with the following criteria:

- Accessibility: the degree to which data are available or easily and quickly retrievable.
- Access Security: the extent to which access to data can be restricted and hence kept secure.

The classification of quality dimensions proposed by Wang and Strong [41] has guided a number of other classifications. There are some IQ criteria that can be used in data integration and PDMS environments. These criteria are presented in the next sections.

## 3. IQ IN DATA INTEGRATION SYSTEMS

A data integration system provides a unified view for users to submit queries over multiple autonomous data sources [1]. The queries are processed over a global schema that offers an integrated view of the data sources. There are some works on IQ issues in the setting of data integration scenarios, in particular using IQ in query formulation, processing (mediation) and optimization [9, 13, 15, 21, 24, 30, 36, 50].

There are some key points of the data integration system in which it is possible to insert IQ analysis: data source schema; mediation queries; integrated schema; source selection; query processing; data integration and data materialization [30]. Also, it is possible to think in several IQ criteria that can be associated with these data integration components.

### Data Source Schema

The data sources are autonomous and heterogeneous. Data sources can be added or can be temporarily or definitively unavailable from the integration system at any time. When a data source joins the data integration system, its exported schema describing all the information that will be shared in the data integration system may be available. There are some issues concerning the maintenance of a data source schema: first, if the source schema changes, the mediation schema must be updated to reflect source changes. Second, it is desirable that the data sources publish the most actual schema in order to keep the consistency between its contents and the information available to the data integration environment. Thus, it is interesting to evaluate the completeness of a source schema in terms of the integrated schema. To analyze IQ in data source schemas, we must consider the following IQ criteria: *schema completeness*, *consistency*, *reputation, availability* and *timeliness*.

### Mediation Queries

In Data integrations systems, schema mappings between the data sources and the global schema are defined [2]. Two approaches may be used: (i) global-as-view (GAV), where each object of the global schema is expressed as a view (i.e. a query) on the data sources and (ii) local-as-view (LAV), where mediation mappings are defined in an opposite way, i.e., each object in a given source is defined as a view on the global schema. Considering a data integration system with GAV schema mappings, a user query may be decomposed in mediation entities, thus it is desirable to choose at query execution time, which are the most adequate mediation queries to compute the entities involved in a given user query. The analysis of the mediation queries quality may be attached with the following criteria: *availability*, *timeliness* and *response time*.

### Integrated Schema

The *integrated schema* is an integrated view of the underlying data sources. The queries submitted to the data integration system are processed over this schema. The integrated schema is composed of a number of mediation entities and each one represents a real world concept obtained from data sources. The criteria we can use to measure the quality of integrated schema are *schema completeness*, *consistency* and *minimality*.

### Source Selection

One of the most common use of IQ criteria analysis in data integration is to guide the selection of data sources [15]. The system must be able to select the best data sources to answer to the user queries. It is possible to reduce computational cost of the user query by filtering out low quality sources based on some IQ criteria. The source selection involves the analysis of the following criteria: *reputation*, *data completeness*, *availability*, *timeliness* and *verifiability*.

### Query Answering

Data integration environments provide the execution of queries directly addressed to several autonomous, distributed and remote data sources. Thus, to analyze IQ in query answering, we establish the following criteria: *availability* and *response time*.

### Data Materialization

One of the main issues in data integration is the selective materialization process [29]. Some portions of data more intensively unavailable and static may be materialized in a data warehouse, and the more dynamic data will be accessed by virtual queries. The criteria related with data materialization are: *timeliness*, *response time*, *availability*, *reputation*, *verifiability*.

### Data Integration

At the instance level, integration problems include managing a mediation object as different objects with different attributes values coming from different sources and selecting a source when contradictory information is found in different data sources. Other problems include the conversion of a value coming from different sources and expressed with different content representation. We believe IQ criteria can play a crucial role in instance reconciliation, determining how to deal with similar or contradictory information. We have established the following criteria: *data completeness*, *timeliness* and *verifiability*.

The last step in data integration is the delivering of results to the user. Here, we can add an IQ criterion to enrich and facilitate the evaluation of the overall quality score of integrated query results: *accuracy*.

Table 1 summarizes, for each data integration process, what are the IQ criteria that must be assigned, evaluated and analyzed.

Some relevant works [13, 15, 21, 29, 31] are concerned with addressing IQ issues in data integration systems. In [15], the authors show a classification of IQ criteria with goals of query answering optimization in an integration system. They use the QCA concept to encode the mappings from source schemas to the mediation schema. The user queries are decomposed in QCA and the quality is evaluated by IQ criteria scores associated with these QCA. Also, they consider 22 criteria for analyzing the quality in query answering, but they do not take into account specific criteria for data source, mediation and user schemas. In such work, there are no references, for example, to the *minimality* criterion. The proposal discussed in [29] concerns a data integration system which uses IQ criteria for selectively materialize data into a local repository. In [31], the authors presented a proposal of investigating IQ in data integration schemas. This work has some specializations in [13] and [21]. The work presented in [13] uses the specifications of minimality and schema completeness for an expert schema. Wang [21] also uses the minimality and schema completeness criteria specification of [31] to define an ontology based quality framework that focuses on user requirements.

**Table 1. Data integration processes and IQ criteria [30]**

| System element | Relevant IQ criteria |
|---|---|
| Data source schema | *Schema Completeness, Consistency, Reputation, Availability, Timeliness* |
| Integrated schema | *Schema Completeness, Consistency, Minimality* |
| Mediation queries | *Availability, Timeliness, Response Time* |
| Source selection | *Reputation*, D*ata Completeness*, A*vailability*, T*imeliness*, V*erifiability* |
| Query processing | *Availability, Response Time* |
| Data materialization | *Timeliness, Response Time*, *Availability*, *Reputation*, *Verifiability* |
| Data integration | *Data Completeness, Timeliness, Verifiability, Accuracy* |

# 4. IQ IN PDMS

Peer data management systems (PDMSs) are a natural extension to data integration systems [10, 11]. On the other hand, PDMSs are different from traditional integration systems since they do not provide a mediator schema from which user queries are formulated. In such systems, data sources are stored at different peers and queries are submitted from their schemas. A peer knows about its neighboring peers by mappings, which help to translate queries and transform data. Queries submitted to one peer are answered by data residing at that peer and by data that is reached along paths of mappings through the network of peers. In this light, a peer can contain all or part of a global answer to a given query. In addition, other peers can provide the same answers with different levels of quality and cost.

Mainly due to semantic heterogeneity, research on PDMS has considered the use of ontologies [11, 18] as a way of providing a domain reference as well as to describe source schemas in a uniform notation. Carrying semantics for particular domains, ontologies are largely used for representing domain knowledge. Xiao [18] has introduced a new definition for the blending of PDMSs and ontologies' researches. In his work, such blending has led to the emergence of Ontology-based Peer Data Management Systems (OPDMS). In this work, we consider both PDMSs and OPDMSs.

An important characteristic of a PDMS lies in its dynamicity. Peers can join and leave the network at any time. Thus, the relations of trust, identification and classification of these sources of data become a great relevance factor. As a result, in PDMSs, the IQ of query answers depends not only on data quality of a particular data source (peer), but also on the quality of the mappings between neighbor peers [35]. Particularly, regarding the former, peers may store data of poor quality, and data may be outdated, erroneous, of dubious origin or incomplete. Regarding the latter, the mappings leading to the data can be incomplete or incorrect [38].

Particularly in PDMS, there are (at least) three kinds of runtime factors, which influence the answer to a given user query, and which also influence the quality of query answers [14, 20]:

- Network (dependent) variance: peers may change the data of their sources, change their schemas, redefine mappings, and new peers may join or leave the system at any time. Thus, the same query submitted to a given peer, but at different times, may yield different answers of different quality.
- Peer (dependent) variance: mappings are established differently from one peer to any other peer. Therefore, the same query submitted at the same time but, by different peers, will result in different query propagation graphs. Consequently, the query results may be different and of different quality.
- Query (dependent) variance: different queries submitted to the same peer and at the same time may result in different query propagation graphs, and, thereby, may produce different results and of different quality.

The majority of PDMSs approaches route and reformulate queries without concern on the mappings' IQ. Mappings in PDMSs may be defined by means of views, when dealing with view-based query rewriting, or by means of correspondences between peer schema elements, when dealing with query reformulation as a transformation of source concepts by target ones. Usually, these mappings (or correspondences) are determined between pairs of peers which have been semantically grouped. Due to heterogeneity, sometimes, concepts from a given peer may not have exact corresponding concepts in a target one. Nevertheless, these peer schemas (or ontologies) usually overlap at some semantic degree.

For these reasons, the IQ consideration in query answering for PDMS consisting of a large number of peers and mappings among them has been considered an important problem. Indeed, high quality level of query answering has been considered as the fact that data can flow among the peers preserving (at the best possible level of approximation) their soundness and completeness [14]. As a matter of fact, PDMS' query routing services should be able to rank the peers that can better contribute to a given query, according to IQ metrics. In this sense, in Section 4.1, we review the PDMS literature under the aspects of IQ. In Section 4.2, we compare some of these works in the light of the used IQ criteria.

## 4.1 RELATED WORK

The work of Zaihrayeu [20] makes a discussion about the application of quality criteria in P2P systems. It shows how the distributed subjective nature of P2P systems brings new dimensions to the quality assessment. For him, users cannot expect correct and complete query answers, but they accept incomplete and partially incorrect answers. Indeed, a given user query may not need the best possible answer, but simply need some answer. Such kind of answer has been called "good-enough" [20]. The idea is that "an answer will be good-enough when it will serve its purposes given the amount of effort made in computing it" [20]. Such notion has been provided as an extension to the definition of [17].

Zhuge et al. [19] present an automatic semantic link discovery method. To this end, it includes a semantic-based peer similarity measurement for efficient query routing, and schema mapping algorithms for query reformulation. If there is data inconsistency, the system uses a Quality of Peers (QoP) method. This method employs user-perceived quality scores such as the number of returned results, response time, traffic overhead, precision, recall to manage inconsistent data in returned data flows. The data returned by peers with higher QoP are considered more likely to

be consistent. Finally, the peer initiating the query will combine relevant data and then give a uniform view of the results to users and peers.

The work of Löser [3] discusses the concept of semantic overlay clusters (SOC) for super-peer networks. In this approach, it enables a controlled distribution of peers through clusters. To this end, it uses some components as follows: an information provider model, some clustering policies, matching engines and a model distribution engine. The information provider model provides an annotation schema designed to support the definition of semantic overlay clusters by local domain experts within the Edutella Network [48]. In a semantic overlay cluster environment, the model, composed by a set of attributes, is used in order to identify relevant information provider peers. The attributes are either extracted from the information provider peer automatically at runtime (Peer ID, Peer IP, Peer Domain, Completeness, Accuracy, Response Time, Amount of Data) or are manually defined by local domain experts (Peer Schema, Peer Name, Peer Description, Global Classification URI and Taxon Path). This work has recognized that reasoning about IQ has become one of the most important tasks when integrating information from autonomous data sources.

In the Chatty Web system [23], schema mapping quality measures are applied to queries. They are updated along with query execution in the network. Schema mapping quality is measured by syntactic and semantic similarity. Syntactic similarity refers to the extent of information lost from queries when attributes from one schema do not exist in another schema. Semantic similarity refers to the level of agreement on the meaning between schemas, and is measured by looking at the transformations a query suffers when expressed in terms of other schemas.

In [5], they present a solution for PDMS query reformulation that exploits completeness characteristics of mappings between peers. Their approach makes use of a decentralized strategy that guides peers in their decision about which mappings should be followed in order to reformulate queries. Such strategy makes use of statistics from both the peers own data and also from the mappings between neighboring peers. The objective is to decide whether it is worthwhile to send the query to that neighbor peer or whether the query plan should be pruned at this point. In other work, the authors use this approach over the System P [6, 7]. In System P, data completeness IQ is the only one dimension taken into account. The completeness model used in this PDMS is based on the strategies adopted at the previous described work [5, 7]. The model estimates both the result cardinality and the result richness, i.e., the number of returned non-null values. The authors highlight the need of more quality dimensions in the context of PDMSs.

The Humboldt Discoverer PDMS [38, 44] works with four layers: (i) the pdms one, composed by peers and mappings between them; (ii) the semantic dimension, which consists of ontologies and mappings between them; (iii) the Web dimension, where each peer maintains a concept store that indexes the peer's neighborhood on the schema level and (iv) the quality dimension, which influences the query answering in all dimensions. Some IQ criteria used by this work are concerned with the semantic dimension and with the web dimension. Regarding the former, the following criteria have been used: relevancy, intensional completeness and extensional completeness. Regarding the latter, the following ones have been considered: concept coverage, timeliness and peer count.

Karnstedt and his group [33] discuss the semantics of completeness for complex queries in P2P database systems. They also propose methods based on the notion of routing graphs for estimating the number of expected query answers. To this end, it presents an approach for estimating query completeness on peer level. They describe the estimation of completeness by observing the progress of query execution at peer level. To this end, they build a routing graph that represents the peers and connections a query travels during query answering. Each node in the graph, a routing point, represents one peer involved in query answering. Actually, the graph is a tree. The number of leafs in the tree is the number of replies that need to be estimated. The peer level approach does not count the received data items in the answer but the responding peers. The numbers of the responding peers and the expected number of responding peers are compared. Thus, they verify a linear correlation between the number of failed peers and the resulting miss of data from the expected answers.

GrouPeer is a system designed to enable accurate query evaluation through semantic overlay clustering [47]. It automatically creates and maintains semantic groups from relational P2P databases. In GrouPeer, individual nodes decide whether to answer the successively rewritten query or automatically rewrite its original version. The function rewriting can change depending on what a peer considers as a 'good' or 'bad' contribution to the rewriting. GrouPeer does not predefine a similarity threshold below which a query should not be rewritten. The authors recognize that the need (and quality standards) of each user about peer information is unique. Therefore, the similarity threshold according to which a query rewriting should be accepted or disregarded should be tuned by the user. In this sense, the user can tune the weights' values for all query elements. Nowadays, GrouPeer does not provide a mechanism to automatically evaluate obtained query answers.

The ESTEEM (Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents) is a community-based P2P platform for supporting semantic collaboration among a set of independent peers, without prior reciprocal knowledge and no predefined relationships [46]. The goal of ESTEEM is to provide an integrated platform for both data and service discovery/sharing in a community-based P2P environment. A distinguishing feature of ESTEEM is the use of semantic communities to explicitly give shape to the collective knowledge and expertise of peer groups with similar interests. To this end, it uses some Semantic Web techniques, concerning: i) shuffling-based communication, for supporting P2P interactions and the autonomous formation of semantic communities of peers; ii) semantic matchmaking, for enforcing data and service discovery at different levels of flexibility and granularity; iii) context management, for profiling the peer behavior and for filtering the available resources according to the peer current context and preferences; and iv) quality-aware data integration, for specifying different levels of peer/data reliability. In order to support trust and data quality, the Esteem platform adopts the DaQuinCIS [34] system, an architecture for managing data quality in cooperative information systems. This system was extended by introducing specific solutions targeted to work in a P2P environment. The ESTEEM also uses a routing-by-community mechanism which is an extension of H-Link [45]. The H-Link semantic routing

mechanism was developed in the framework of HELIOS peer-based system for knowledge sharing and evolution [43].

## 4.2 COMPARATIVE ANALYSIS

We have accomplished an analysis of these described approaches in terms of which IQ criteria have been used and what kind of assessment they have done. The works of Zaihrayeu [20] and Zhuge *et al*. [19] present a discussion regarding the importance of considering IQ criteria in PDMS, although they do not provide neither used IQ criteria nor quality metrics. The other works present some IQ criteria or some kind of quality factor (not really a criterion) and some brief indication about how they assess them. Most of them discuss IQ issues in terms of schema mappings and query answers. The comparison result concerning these works has been summarized in Table 2.

Regarding mappings, the works presented in [23, 47, 38] focus on the quality of schema mappings applied to queries. The first one [23] uses two quality factors (syntactic and semantic similarity) to assess the quality of attributes that are preserved in query reformulation. Similarly, the work presented in [47] provides a comparison between the submitted query with the set of concepts present in the peer schema, i.e., it tries to identify the similarity between what is being requested in the query and if the peer can answer it. The third work [38] uses quality criteria to measure the quality of data sources and mappings. It considers Timeliness (i.e., the freshness of the stored data), Peer Count (i.e., the number of peers known to the respective peer supporting a specific ontology) and Concept Coverage (i.e., the number of concepts supported by this peer relative to the number of concepts contained within the linked ontology) in order to discover relevant data sources for submitted queries. Nevertheless, although these works indicate the use of some mappings IQ criteria or some factor that assist quality identification, most of them do not give details about how they accomplish such tasks. An exception of this is the work of [38] which defines IQ criteria and presents some techniques to be used in order to assess them.

Regarding query answers, some IQ criteria are more clearly defined. The works presented by Löser *et al.* [3], Karnstedt et al. [33], Roth and Nauman [5, 6, 7], Heese *et al.* [38] and Montanelli [46] recognize the importance of completeness and accuracy of query answers. They try to measure completeness through different points of view, namely: (i) in [3], by considering the absolute number of available resources; (ii) in [33], by observing the progress of query execution at peer level; (iii) in [5], by including the use of coverage (i.e., the proportion of the size of a tuple set to the number of all tuples stored within the PDMS) and density (i.e., the arithmetic mean over all attributes occurring in a query); and (iv) in [38], by defining intensional and extensional completeness . They also indicate the use of some other criteria such as response time (i.e., the average delay in milliseconds between submission of a query and the reception of the response), amount of data (i.e., the size of the query result) and consistency (i.e., captures the violation of semantic rules defined over a set of data items ). Although, they provide these definitions, the main focus usually rely on query completeness.

Analyzing these works, we can observe that, although they have pointed out the importance of IQ criteria in their processes, the definition and assessment of such IQ criteria is still difficult to determine and manage. Reasons underlying this difficulty include the peers heterogeneity, the varying capabilities of the peers in terms of data, the dynamicity of the network and the degree the available domain knowledge is employed. Thereby, we verify the relevance of IQ research in PDMSs settings and the necessity of deeply investigating not only IQ criteria related to mappings and query answers, but also some other related to the data sources, their schemas and existing working data.

**Table 2. Comparative Analysis of Related Works**

| Work | IQ Criteria (or Quality Factors) |
|---|---|
| Zaihrayeu [20] | Not identified |
| Zhuge *et al*. [19] | Not identified |
| Löser *et al.* [3] | Completeness, accuracy, response time, amount of data |
| Aberer *et al.* [23] | Syntactic and semantic similarity of schema mappings. |
| Roth and Naumann [5] | Completeness as coverage + density |
| Humboldt Discoverer [38, 44] | Timeliness, Peer Count, Concept Coverage of mappings; relevancy, intensional completeness and extensional completeness of queries |
| Karnstedt *et. al* [33] | completeness |
| Kantere *et al.* [47] | Query structural similarity; Quality of received answers |
| Montanelli *et al.* [46] | Column completeness, format consistency, accuracy, internal consistency. |

# 5. CHOSEN IQ CRITERIA FOR PDMS

A considerable number of PDMSs approaches have pointed out the need of gathering and evaluating IQ in order to improve their services. Nevertheless, only few works have really used IQ in a complete approach, i.e., gathering IQ, measuring IQ according to some defined metrics and applied these obtained results to enhance the PDMS services. Most of the related works presented in Section 4 discusses IQ issues in terms of schema mappings [23, 38, 47] and query answers [3, 5, 33, 46].

Similarly as in data integration systems, we believe that, in PDMSs, there are some elements in which an IQ analysis is suitable: the peers; the schemas (represented as ontologies, in the case of OPDMSs); the schema mappings; the data and query answers. In this light, in this section, we present a set of quality criteria that may be used for PDMSs. In order to achieve this, several IQ criteria used in data integration systems and other ones identified from related works have been analyzed from a PDMS's point of view. These criteria have been further grouped into five broad IQ classes, according to the identified PDMS elements. The PDMS elements and their IQ criteria are presented in Table 3. Furthermore, we provide the definitions underlying these IQ criteria as well as how they can possibly be evaluated.

**Table 3 – PDMS Elements and corresponding IQ Criteria**

| PDMS Elements | Quality Criteria |
|---|---|
| Peer | Availability<br>Reputation<br>Access frequency |
| Peer Schema/Ontology | Completeness<br>Representational Consistency<br>Minimality |
| Mappings | Confidence<br>Relevancy |
| Data | Timeliness<br>Freshness<br>Trust<br>Relevancy |
| Query answer | Completeness<br>Accuracy<br>Relevancy |

The *Peer* element represents an autonomous data source that exports its entire data schema or only a portion of it. Each peer expresses and answers queries based on its exported schema. Regarding the *Peer* element, there are three IQ criteria that should be taken into account:

- Availability: concerns the verification of how often a data source itself is commonly available (and not only its schema). Since a peer may join and leave the network at any time, it is rather relevant to measure the degree of availability of a peer in a given time interval. Such criteria can be measured taking into account the statistics about the peer network connection. To this end, it's necessary to monitor the number of times a peer has been unavailable or even the percentage of time that the peer is accessible.
- Reputation: Wang [41] defines reputation as "the extent to which data are highly regarded in terms of their source or content". Thus, it concerns the degree to which the information of a source is in high standing. Reputation can be assessed by calculating the average of query answered and source information or a score reflecting users' preference.
- Access frequency: represents how often a peer is accessed in a given time interval. In order to assess this criteria, we should measure the ratio between the number of times that a peer has been accessed and the total number of queries submitted at the PDMS in a given time interval.

When a peer joins a PDMS, it exports its schema usually using a common unified notation. As mentioned in Section 4, in OPDMSs [12, 18], ontologies are used as peer schema representation. Thus, the *Peer Schema* or *Peer Ontology* elements represent the set of concepts shared by each peer, i.e., the shared knowledge provided by a peer to a PDMS. Regarding the peer schema/ontology, there are some IQ criteria that should be considered, as follows:

- Completeness: the degree to which entities and properties of the peer are not missing in the schema. The more data and knowledge a peer provides through its schema/ontology the more attractive it is to users. This criteria may be assessed by taking the ratio between the number of items the peer provides and the total number of existing items in the PDMS.
- Minimality: indicates the degree in which the peer schema is modeled compactly and without redundancies. In other words, the peer schema must be as precise and correct as possible. Its metrics may be obtained by the number of redundant concepts in proportion to the total number of schema concepts.
- Representational Consistency: is the "extent to which data are always presented in the same format and are compatible with previous similar data" [41]. In a PDMS which uses some kind of reference vocabulary (e.g., a domain ontology), it concerns the consistency of a schema in terms of such reference vocabulary. In these cases, usually representational consistency refers to measuring if schema elements are compatible with the elements found in the reference vocabulary. It can be assessed by applying some similarity measure between the schema and the vocabulary used as reference.

Mappings represent associations between peer schema/ontology elements. In a PDMS, usually, schema matching techniques are used to establish such mappings which are the basis for query reformulation. There are some IQ criteria that can be used related to mappings:

- Relevancy: refers to how often a given mapping has been used in a set of query reformulations during some time interval. It may be assessed by the percentage of usage in query reformulations during a period of time.
- Confidence: refers to the level of trust (or confidence) that has been associated to the mapping (or correspondence) at its creation time. The confidence level may also be increased or decreased according to user feedback.

The *Data* element represents the data which are stored in the peers. Defined IQ related to *Data* refers to the stored data, and not to the peer schema or peer ontology. They are defined as follows:

- Timeliness: refers to the data update frequency, i.e., how often data changes in a source. This criterion can be assessed by calculating the average (in days) of data update.
- Freshness: represents the time passed from the last data update to the current access date.
- Trust: regards the data trustworthy. The data trust can be assessed taking into account the peer's reputation and by averaging the number of answers accepted by users within a specific query.
- Relevancy: refers to the suitability of data to queries submitted by users. This criterion is subjective and user-dependent, since only the user can determine whether something is relevant or not. In order to assess such criteria, we should evaluate the user feedback regarding the query answers, i.e., it should be assessed considering the indicated relevance of produced query answers.

The *Query answer* element represents the system's answers to queries submitted by the users. In dynamic distributed settings such as PDMSs, answers to user queries, in most of the cases, are not supposed to be complete, as they usually are in other integration approaches. However, the answers shall be as close as possible to users' needs, and they shall reflect the current status of the environment. Usually in a PDMS, every peer $P_i$ maintains a

neighborhood $N(P_i)$ selected from the set of existing peers in the setting. A query management process allows to specify a user query at some peer $P_i$, and to compute it in a fully decentralized manner involving the set of relevant neighbor peers. There are some IQ criteria related to query answer that can be taken into account, as follows:

- Completeness: due to dynamicity, query answers in a PDMS may not be complete, considering its original definition (completeness is typically understood as the ratio of answer set size to the total amount of known data [33; 16]), which requires the knowledge of the total amount of data in the system and relies on the closed world assumption. Instead, peer schemas in the set of available peers have an open-world assumption [28; 42], i.e., the data returned by querying these peer schemas may be incomplete. In this light, completeness in PDMSs may be defined as the ratio between received results and the existing suitable data belonging to the available peers at query answering time. In order to assess completeness, firstly, it's necessary to identify the set of peers which can contribute to the submitted query. Then, if these relevant peers are available. After all, we should route the query to these peers and estimate the set of possible answers they can produce. At end, we should compare the received results with this estimated set of possible answers. Such process should generate statistics regarding the peer's contribution to the given query.
- Accuracy: refers to the degree to which answers to user queries (that are submitted in a PDMS and are reformulated according to existing mappings) conform to the user requirements relating to information precision. Query answers are supposed to be correct, reliable and certified free of error. Nevertheless, sometimes, query answers which are not an exact match, but a close match to the requirements specified at query submission time, can still serve the purpose of users. This may be considered depending on users preferences and on the dynamicity of the environment. In this case, accuracy should be measured by means of a user feedback.
- Relevancy: refers to the degree to which data are applicable and helpful for the query at hand. Relevancy should be measured according to the determined relevancy of the peers that have contributed with the answers as well as according to some kind of user feedback.

In general, PDMSs data management is inherently open world [33]: while answering a query, peers can fail, leave or join the network. We argue that estimating the completeness, relevancy and accuracy of query answers is a key aspect of reliable query answering in PDMS environments. Usually, an approximated, but prompt estimation may be satisfactory for the user and may be also considered satisfactory to be used in future tasks. In order to measure the best answers to the users, we could consider the following: (i) to assign a score to each individual answer by taking into account the number, relevance and reputation of the sources reporting the answer, (ii) to verify the prominence of the answer within the sources, and (iii) to aggregate the scores of produced similar answers.

# 6. CONCLUSIONS AND FURTHER WORK

Due to the ever increasing complexity of PDMSs services, managing Information Quality (IQ) is becoming more and more a necessity. In this sense, this work presented a review of IQ research related for PDMSs. First, we explained the importance of IQ and an usual classification of IQ aspects. Once the PDMSs are considered as an evolution of data integration systems (DIS), we discussed an IQ classification oriented to address DIS components. We presented a summary of relevant works concerned with the use of quality criteria in PDMSs and, finally, we have proposed and briefly described a list of IQ criteria we believe can be used for PDMSs processes.

We are currently investigating how a PDMS can be extended to support a dynamic information quality aware service. To better understand the role of each IQ aspect and to facilitate the running of such IQ oriented PDMS, we also intend to group the selected criteria in broader dimensions. Then, we intend to formally specify the selected PDMS IQ criteria and to experiment the criteria evaluation in an existing PDMS [10].

# 7. REFERENCES


[1] A. Halevy, A. Rajaraman and J. Ordille. Data Integration: the Teenage Years, *In Proceedings. of the 25th International Conference on Very Large Data Bases (VLDB)*, pages 9-16, Seoul, Korea, September 2006.

[2] A. Halevy. Theory of Answering Queries Using Views. *SIGMOD Record*, vol. 29, no.4, December 2000.

[3] A. Löser, M. Wolpers, W. Siberski, W. Nejdl. Semantic Overlay Clusters within Super-Peer Networks. In Proceedings of the International Workshop on Databases, Information Systems and Peer-to-Peer Computing (P2PDBIS). Berlin, Germany 2003.

[4] A. Marotta and R. Ruggia. Managing Source Quality Changes in Data Integration Systems. In proceedings of 2nd International Workshop on Data and Information Quality (DIQ'05), 2005.

[5] A. Roth and F. Naumann. Benefit and Cost of Query Answering in PDMS. In Proc. of the Int. Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P), 2005.

[6] .A. Roth and F. Naumann, System P : Completeness-driven Query Answering in Peer Data Management Systems. *Business, Technologie and web (BTW'07)*, pages 1-4, Aachen, Germany, 2007

[7] A. Roth, F. Naumann, T. H¨ ubner, and M. Schweigert. System P: Query Answering in PDMS under Limited Resources. In Proc. of the Workshop on Information Integration on the Web (IIWeb), 2006

[8] B. Stvilia, L. Gasser, M. B. Twidale and L. C. Smith. A Framework for Information Quality Assessment. In Journal of the American Society for Information Science and Technology, Vol. 58, N. 12, pages 1720-1733, October 2007.

[9] C. Aggarwal and P. Yu. A Survey of Uncertain Data Algorithms and Applications. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 5, May 2009.

[10] C. Pires. Ontology-Based Clustering in a Peer Data Management System. PhD thesis, Federal University of Pernambuco (UFPE), Recife, PE, Brazil, April 2009.

[11] D. Souza, C. E. Pires, Z. Kedad, P. C. A. R. Tedesco, A. C. Salgado. A Semantic-based Approach for Data Management in a P2P System. To be published in LNCS Transactions on Large-Scale Data- and Knowledge-Centered Systems, 2011.

[12] D. Souza. Using Semantics to Enhance Query Reformulation in Dynamic Distributed Environments. PhD Thesis, Federal University of Pernambuco (UFPE), Recife, PE, Brazil, April 2009.

[13] F. Duchateau and Z. Bellahsene. Measuring the Quality of an Integrated Schema. *In Conceptual Modeling – ER 2010*, Lecture Notes in Computer Science, 2010.

[14] F. Giunchiglia and I. Zaihrayeu I. Making peer databases interact - a vision for an architecture supporting data coordination. In Proceedings of the 6th International Workshop on Cooperative Information Agents (CIA), pages 18–35, Madrid (ES), 2002.

[15] F. Naumann and U. Leser. Quality-driven Integration of Heterogeneous Information Systems. *In Proceedings of the 25th International Conference on Very Large Databases (VLDB' 99)*. pages 447-458, Edinburgh, UK, September 1999.

[16] F. Naumann, J.-C. Freytag, and U. Leser. Completeness of integrated information sources. Information Systems, 29(7):583–615, 2004.

[17] H. Stuckenschmidt, F.V. Harmelen, and F. Giunchiglia. Query Processing in Ontology-Based Peer-to- Peer Systems, *Ontologies for Agents: Theory and Experiences*, Birkhauser, 2005.

[18] H. Xiao. Query processing for heterogeneous data integration using ontologies. PhD Thesis in Computer Science. University of Illinois at Chicago, 2006.

[19] H. Zhuge, J. Liu, L. Feng, X. Sun, and C. He, "Query Routing in a Peer-To-Peer Semantic Link Network," *Computational Intelligence*, vol. 21, N. 2, pages 197-216, May. 2005.

[20] I. Zaihrayeu. Towards Peer-to-Peer Information Management Systems. PhD Thesis, DIT – University of Trento, March 2006.

[21] J. A. Wang. Quality Framework for Data Integration. In Proceedings of the 27th British National Conference on Databases (BNCOD), 2010.

[22] J. Zhao. Schema Mediation and Query Processing in Peer Data Management Systems. Master Thesis, The University Of British Columbia, October 2006.

[23] K. Aberer, P. Cudre-Mauroux, and M. Hauswirth. The chatty web: emergent semantics through gossiping. In *WWW03 Conference Proceedings*, pages 197-206, May 2003.

[24] K. Keeton, P. Mehra and J. Wilkes. Do You Know Your IQ? : A Research Agenda for Information Quality in Systems. *ACM SIGMETRICS Performance Evaluation Review*, Vol. 37, Issue 3, December 2009.

[25] L. English. Seven Deadly Misconceptions about Information Quality. *DM Review Magazine, 1999*. Available at http://www.dmreview.com/issues/19990701/1239-1.html. Last access in May, 22nd 2011.

[26] L. G. A. Sung, N. Ahmed, R. Blanco, H. Li, M. A. Soliman, D. Hadaller. A Survey of Data Management in Peer-to-Peer Systems, Web Data Management,Winter, pages 1–50, 2005.

[27] L. L. Pipino, Y. Lee and R. Wang. Data Quality Assessment, *Communications of the ACM*, Vol. 45 N. 4, pages 211-218, April 2002.

[28] I. Tatarinov and A. Halevy. Efficient query reformulation in peer-data management systems. In Proceedings of the ACM International Conference on Management of Data (SIGMOD), pages 539-550, June 2004.

[29] M. C. Batista. Otimização de Acesso em um Sistema de Integração de Dados através do uso de Caching e Materialização de Dados, Master Thesis, Federal University of Pernambuco, 2003.

[30] M. C. Batista. Schema Quality Analysis in a Data Integration System. *PHD Thesis*, Federal University of Pernambuco, Brazil, June 2008.

[31] M. C. Batista and A.C. Salgado. Data Integration Schema Analysis: An Approach with Information Quality. *In Proceedings of the 12th International Conference on Information Quality (ICIQ)*, MIT, Massachusetts, USA, October 2007.

[32] M. Ge and M. Helfert. A Review of Information Quality Research - Develop a Research Agenda. *In Proceedings of the 12th International Conference on Information Quality (ICIQ)*, MIT, Massachusetts, USA November 2007.

[33] M. Karnstedt, K. Sattler, M. Haß, M. Hauswirth, B. Sapkota, R. Schmidt. Approximating query completeness by predicting the number of answers in DHT-based web applications. In Proceeding of the 10th ACM workshop on Web information and data management (WIDM'08), pages 71-78, California, USA, October, 2008.

[34] M. Scannapieco, A. Virgillito, C. Marchetti, M., Mecella, and R. Baldoni. The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. Information Systems, 29(7), pages 551–582, 2004.

[35] M. Yatskevich, F. Giunchiglia, F. McNeill, and P. Shvaiko. OpenKnowledge Deliverable 3.3: A methodology for ontology matching quality evaluation, 2006. Available at http://www.cisa.informatics.ed.ac.uk/OK/Deliverables/D3.3.pdf. Last access in May, 22nd 2011.

[36] P. Agrawal, O. Benjelloun, A. Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom. Trio: A System for Data, Uncertainty, and Lineage. In C. Aggarwal, editor, Managing and Mining Uncertain Data, Springer, September 2009.

[37] P. Angeles and L.MacKinnon. Quality Measurement and Assessment Models Including Data Provenance to Grade Data Sources. *In Conference on Computer Science and Information Systems*, pages. 101-118, Greece, June 2005.

[38] R. Heese, S. Herschel, F. Naumann, and A. Roth. Self-extending Peer Data Management. In Proceedings of The German Conference on Datenbanksysteme in Business, Technologie und Web, volume 65 ofLNI. GI, March 2005.

[39] R. Keeney and H. Raiffa. Decisions With Multiple Objectives: Preferences and Value Tradeoffs. John Wiley & Sons, New York, 1976.

[40] R. Price and G. Shanksa. Semiotic Information Quality Framework. *In Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS)*, pages 658-672, Prato, Italy, July 2004

[41] R. Y. Wang, D. M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, Vol. 12, N. 4, pages 5-33, 1996.

[42] S. Abiteboul and O. Duschka. Complexity of answering queries using materialized views. In Proc. of PODS, pages 254-263, Seattle, WA, June 1998.

[43] S. Castano, A. Ferrara, S. Montanelli, and D. Zucchelli, HELIOS: a general framework for ontology-based knowledge sharing and evolution in P2P system. In Proceedings of the *14th International Workshop on Database and Expert Systems Applications*, pages 597-603, 2003.

[44] S. Herschel and R. Heese, R. Humboldt Discoverer: A Semantic P2P index for PDMS. In Proceedings. of the International Workshop Data Integration and the Semantic Web, Porto, Portugal, 2005.

[45] S. Montanelli and S. Castano, "Semantically routing queries in peer-based systems: the H-Link approach," *The Knowledge Engineering Review*, vol. 23, pages. 51-72, Mar. 2008.

[46] S. Montanelli, D. Bianchini, C. Aiello, R. Baldoni, C. Bolchini, S. Bonomi, S. Castano, T. Catarci, V. Antonellis, A. Ferrara, M. Melchiori, E. Quintarelli, M. Scannapieco, F. a Schreiber, and L. Tanca, The ESTEEM platform: enabling P2P semantic collaboration through emerging collective knowledge, Journal of Intelligent Information Systems, Jun. 2010.

[47] V. Kantere, D. Tsoumakos, T. Sellis, N. Roussopoulos. GrouPeer: Dynamic clustering of P2P databases. *Information Systems*, v.34 n.1, pages 62-86, March 2009.

[48] W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, M. Nilsson, M. Palm, and T. Risch. EDUTELLA : A P2P Networking Infrastructure Based on RDF. In Proceedings of the eleventh International World Wide Web Conference (WWW'02), pages 604-615, Hawaii, USA: 2002.

[49] Y. W. Lee, D. M. Strong, B. K. Khan, and R. Y. Wang . AIMQ: A Methodology for Information Quality Assessment. *Information & Management*, Vol. 40, N. 2, pages. 133-146, December 2002.

[50] Z. Zhao. Data Quality-Oriented Data Integration in Peer-to-Peer System. In Proceedings of the 9th International Conference on Hybrid Intelligent Systems, August 2009.