DATA QUALITY EVALUATION IN AN E-BUSINESS ENVIRONMENT: A SURVEY

"Research-in-progress"

Soumaya Ben Hassine-Guetari
ERIC Laboratory, Lyon University
soumaya_ben_hassine@voila.fr

Abstract: Industrial databases and information systems are plagued by a plethora of quality issues especially in cooperative information systems under the use of external data files because of the lack of collaboration in internal integration processes of marketing and sales data which is still a biggest challenge facing businesses, and, above all, the lack of benchmarks and tools standardizing for the external files exchanges and handling.
In fact, the quality of exchanged data is essential for developing service-based applications and correctly performing cooperative activities such in Business-to-Business (B-to-B) marketing operations where issues are no longer limited to individual erroneous records but also to the insufficiency of customer knowledge and files features when integrating external data files.
In this paper, we describe data assessment methods needed to assess and maintain the quality of an e-business activity. The aim is to define the basis of a data quality assessment framework that is suitable for evaluating and enhancing an e-business information system.

Key Words: data quality issues, data evaluation tools, business-to-business applications, e-business application.

1- INTRODUCTION

Industrial databases and information systems are plagued by quality problems and, according to a study handled in 2007 by Lionel Schwartz, it encounters data loading and management issues. In fact, 62% of the companies refuse to formalize and standardize the data-gathering strategies tolerating, as a consequence, duplicated or incomplete data. Moreover, 27% of the companies perform data cleaning techniques once a year and 7% never handle data cleansing activities [29].
For instance, in business-to-business domain, companies are overwhelmed with issues related to the quality of their customer data and just don't know where to start," says Chad Gottesman, Extraprise CMO. "Unfortunately, data problems continue to grow, resulting in insufficient customer knowledge that leaves the B2B salesperson at a complete disadvantage when interacting with customers."[35]
In e-business applications, consisting in the conduct of business on the Internet and in particular servicing customers and collaborating with business partners [27], it is tedious to evaluate and maintain the quality level due to the increase in a company's interaction with its external environment (data vendor files).
In fact, the success of any marketing initiative basically depends on accurate and consolidated customer and prospect information, and performing quality benchmark provides accurate delivery of goods and services, increases target marketing effectiveness and enhances customer relationships through better service.

At the same time, data quality research and practice have noted a significant evolution during the last twenty years. In fact, in the first nineties, researches had focused on the definition of data quality evaluation criteria (especially accuracy) in the context of data values assessment in information systems. As a consequence, hundreds of dimensions have been defined leading to the multidimensional aspect of the data quality problem [27].
Nowadays, and, as a result of the shared architectures of information systems, this research domain is becoming more and more active. Proposed solutions are no longer punctual (focusing on data values) but involve as well the whole data workflow of the information system (data provenance, transformation, loading, mediation, processes quality, schema quality) affecting largely and indistinctly every application domain (database management, marketing, business intelligence, geospatial applications, etc).
In this paper, which is an introductory study of a project thesis focusing on the improvement of data quality in e-business applications, we describe our case study emphasizing on the main encountered issues in practice. We state, then, existing data quality evaluation's means and choose the appropriate ones in terms of their usability and adequateness in the context of the underlying applications.

2- CASE STUDY: CONTEXT DESCRIPTION

As in any data quality initiative it is critical to understand the context of the underlying project and the connections between data adequateness and accuracy, business processes and financial results, we will illustrate, in the following section, a detailed description of our application.

2.1- Description of the current application: principle and limitations

In this project, we are focusing on a specific case of e-business applications: a direct marketing Business-to-Business (B-to-B) project where we are interested in the assessment of the appropriateness and accuracy of the targeted prospects used in marketing campaigns' as well as the well identification of these prospects from a multi-sourced database. In fact, these prospects are gathered from numerous data sources:

- Official sources: national administrative data files. Example: SIRENE[1] file and BODACC[2] file utilized to validate prospects information, and, NPAI file, CHARADE file and ESTOCADE file utilized to detect obsolete data on the clients' files.
- Quasi-official sources: telephone files, etc
- Professional data providers (data vendors): Commback, Coface files, etc
- Other sources consisting of companies hiring or selling their contacts' databases aside from their main activity as a means of making profit and paying off the cost of their databases' management: JPG (office furnishing products), BERNARD (personal care products), etc

The major concern with the creation of a successful marketing campaign consists in selecting the *n* most appropriate contacts (*n* is set by the client aside from other prospects features as the targeted profession or business line) and requires an elaborated data analysis and customers' and clients' profiles.

In a practical point of view, the application deals with a prospects-counting platform based on a federated database of providers' data files. In fact, when a client provides his customers' file, a data append activity is undertaken. This append is almost done by comparison with the SIRENE file (reference database) using the unique identifier of French institutions and working persons: the SIRET identifier. Once the data file enriched, it is integrated into the federated database, the same data file could indicate the client's profile.

[1] The SIRENE file is provided by the INSEE (French national institute of statistics and economical studies). It describes the recent situations of the French institutions and companies.
[2] The BODACC file (Official bulletin of civil and commercial announcements) contains all legal announcements of the French institutions and companies.

Then, when the latter needs to run a marketing campaign on a specific population (for example: one thousand IT managers of companies with around a hundred employees), the system chooses amongst the contacts listed on the central database the population corresponding to the client's request. The selection criterion is based on the expertness of the system's user regarding, on one hand, the sources reputation and the accuracy and freshness of the data values, and on the other, the client's profile.
As a conclusion, the current application performs a non-automatic selection based on week criteria of information quality assessment.

2.2- Data Quality problems in the underlying B-to-B environment

When developing a B-to-B direct marketing campaign, it is of critical importance to target the prospects' companies and identify the correct individuals by function or title within the targeted companies in order to obtain the names of these individuals, and then maintain that information accurate over time.
But, such activity is laboring under numerous drawbacks. Some are inflicted by external factors (data files' providers) while the majority is related to the features and characteristics of the queried information system and its shortcoming quality.
In fact, working in a collaborative environment and supporting exchanged data with external companies increases the data replication and data inconsistency rates in the federated database where we can find numerous copies of the same information with variable quality. For instance, we can find 3 records of the same chief executive officers of a company with 3 different names provided by 3 different sources. The concern here is to identify the most reliable source giving the more accurate and up-to-date information.
Moreover, the majority of data vendors don't provide data management information required in the source selection phase when the system has to choose amongst duplicated record generated from different sources and containing some different items or conflicting information. Such data management information is related, for instance, to the data update information (update dates, records, methods, frequencies).
Furthermore, the lack of normalizations and benchmark for collaborative e-business tasks complicates the data integration and record linkage processes.

As a conclusion, and according to the application's specifications, data quality concerns are related to the following topics:

- Data provenance and origin: Also called as data lineage, data provenance is used to assess the data quality reliability of the different data sources, when working in a multi-sourced environment. Data provenance also helps determine the extent to which the quality of the data can be estimated in addition to the level of the authenticity of the data regarding the pedigree of provenance. In our project, an assessment model has to be handled in order to select the appropriate record regarding the customer profile and request.
- Data integration and record linkage with consideration to the sources quality and its records' prices. In fact, data integration has become a necessity in the sense that a poor integration can result in e-business chaos driving down efficiency and eroding return on investment (financial criteria defined later)[34].
- Query performances related to the counting task and prospects' selection
- Clients profiling needed in the prospects' selection phase to insure the success of the marketing campaign

Here is an overview of a B-to-B application in an e-business environment with some of the data quality issues and contexts:

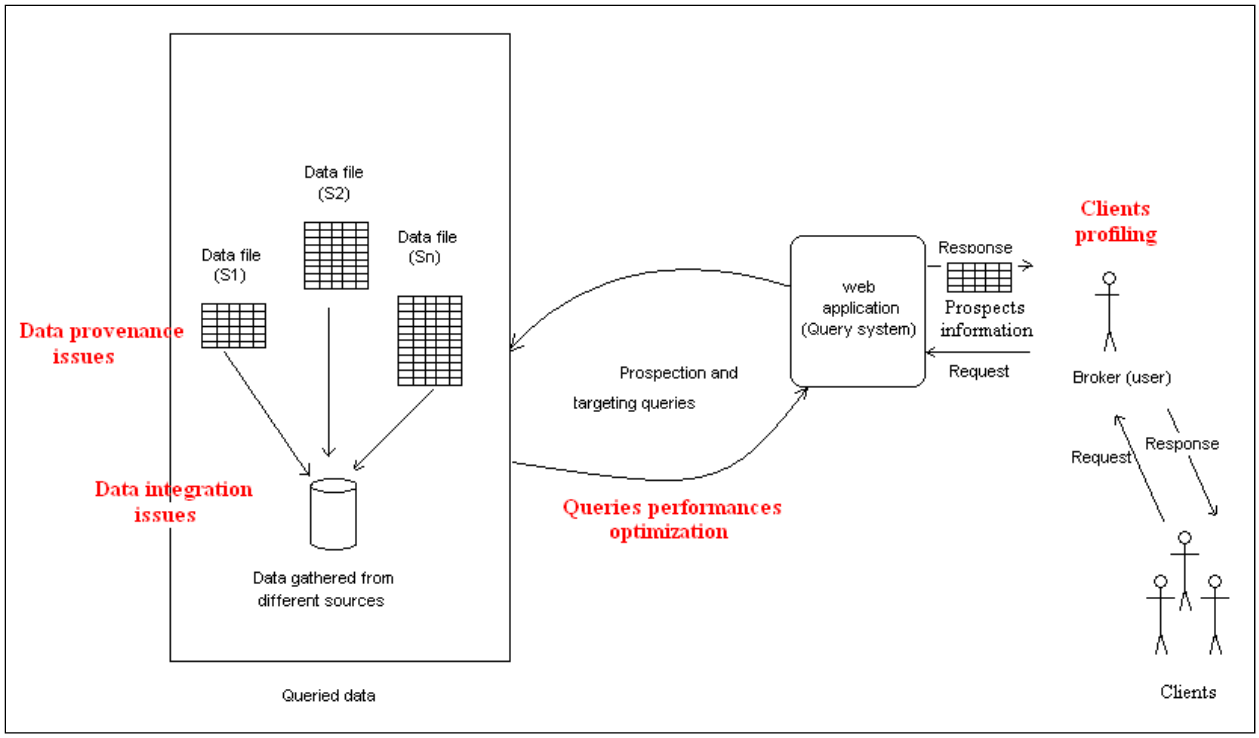

Figure1: conceptualized model of our B-to-B application

In this section, we described the context of the application and illustrated its main functionalities with a rough conceptualization model. An elaborated application model is to be defined in the following works as in this paper, we are rather interested in the data quality dimensions needed to compute and assess the appropriateness of the selected prospects to the clients' marketing campaign.
The following section details a brief state of the art of the data quality assessment tools and methods.

3- BACKGROUND: QUALITY ASSESSMENT

Information quality research has been carried out from the beginning of the fifties and evolved considerably where dealing with data quality has meant overcoming anomalous data and detecting the presence of errors.
Many solutions from various domains have been defined. For instance, in the information theory domain, the quality of the transmitted message was first evaluated by the Hamming complete error-detection and error-correcting encoding procedures, then, by the Shannon entropy formula, in order to assess the correctness (in sense of homogeneity) of transferred information through a noisy channel.
Then, many statistical techniques have been defined for what they named "Fault Diagnosis" such as applications of artificial intelligence techniques (artificial neural networks, fuzzy logic, etc) and data analysis techniques based on viewing diagnosis in terms of quality control (statistical process control (SPC), principal component analysis (PCA) for industrial process monitoring, matrix evaluation for digital models, etc).
Later, in the beginning of the nineties, information system management experts focused on the dimensional and subjective aspect of data quality defined by "the fitness for use" for a specific data set.
In the case of this survey, we concentrate on the latter data quality aspect where we distinguish the evaluation tools below:

- intrinsic quality of data records: data quality dimensions/data quality metrics
- ontological data quality dimensions and data quality metadata
- data warehouses data quality tools (integration schemas, ETL tools, data provenance, mediation query plans, etc)
- total data quality methodologies

More details are given in the following paragraphs.

3.1- Intrinsic quality of data records

Intrinsic quality of individual records comprises data quality dimensions and their associated metrics. In fact, improving data quality has always meant to information quality managers satisfying the customer needs, and the concern was to define the most suitable attributes or aspects. [27]

Many approaches have been defined for that purpose. For instance, a theoretical approach focusing on how data may become deficient during the data manufacturing process has been proposed by Wand and Wang in 1996. It has to deal with an ontological approach in which errors are defined as the inconsistencies between the real world system and the representation entities [25].

Another method, the intuitive approach, consists in using the perspicacity of the domain's experts. Besides, this approach is study dependant and is handled when the selection of the data quality attributes for any particular survey is based on the researchers' experience or intuitive understanding about what attributes are important.

While the two previous approaches focus on the product characteristics, other studies are based on the customer needs and "fitness for use" principle. This method analyzes data from customers' point of view to generate a set of adequate evaluation criteria. Concretely, it involves the users themselves into the quality metadata management process, by explicitly soliciting and exploiting their feedbacks. Unless this approach cannot be proven via fundamental principles, it was adapted in a plenty of studies and "fitness for use" has been the phrase defining the data quality concept [25,29,27,8]. The most famous one is the empirical study performed by Wang and Strong in 1996. Another example is the Berti-Equille's approach developed in 1999 that enhances the subjective aspect of data quality by adapting methods from cognitive science and human science domains. [3]

But, while data quality researchers provided us with a glut of dimensions, measuring data quality still remains an issue when evaluating databases data quality. The first concern is related to the maintainability and freshness of the measures values as they cannot be defined in a "one shot approach". Instead, they must evolve in a real-time architecture in order to maintain a high level of data quality in dynamic databases. The second concern is related to the use of the most adequate function as data quality is context dependant.

In this context, many studies have focused on making up the measurement functions. For instance, in 2004, Bouzeghoub and Peralta defined an evaluation framework for data freshness evaluation where they compare the numerous definitions and measurement approaches of this data quality dimension [18]. And later on, in 2006, Redman defined the term "measurement system" to refer to all activities that make up measurement. This system is based on a five-action process (business requirements, subject of measurement, measurement protocol, management action, and integration or evolution of the measurement system) and suggested a data accuracy measurement framework based on the customer-supplier model [20].

Some of the data quality dimensions and measurement functions are defined in Appendix1.

3.2- Ontological data quality dimensions and data quality metadata

Due to the concept dependency as well as the abundance of data quality dimensions, studies were oriented to conceptualize these criteria and create what we call "data quality metadata".

In fact, dealing with data quality with ontological methods supposes that the reality and the information model are represented in the same model. This allows using “closed loops semantics” to define “fitness for use” as leading to correct, executable decisions [10].
We distinguish technical and semantic metadata (ontology).
Technical metadata are defined by databases administrators or data warehouses teams and describe technical data tags helping reconcile data quality issues in its underlying environment (load date, update date, load cycle identifier, current flag indicator, operational system identifier, active operational system flag, confidence level indicator, etc). [15]
Semantic metadata or data quality ontology leads to a formalized and conceptualized description of the crucial evaluation aspects. For instance, Wang and Strong defined four categories of DQ dimensions:

1- intrinsic data quality related to the proper value of the data (accuracy, objectivity, etc)
2- contextual data quality highlighting the requirement that data quality must be considered within the context of the task at hand (relevancy, timeliness, completeness, etc)
3- representational data quality related to the information systems properties (interpretability, concise representation, etc)
4- accessibility data quality (access security)

Then, in 1999, Naumann, Leser and Freytag propose a classification of twenty two dimensions into three classes or metadata: one related to the user preferences, the second class concerns the query processing aspects and the last one is related to the data sources [17].
Besides, in [14], data quality dimensions are classified into sound, useful, dependable, and usable, according to their positioning in quadrants related to “product quality/service quality” and “conforms to specifications/meets or exceeds consumer expectations” coordinates. We distinguish then:

- completeness: when product quality is sound and comply to specifications (free for error, completeness, etc)
- usefulness: when product quality meets or exceeds customer needs (appropriate amount, relevancy, objectivity,etc)
- dependency: when service quality is conforming to specifications (timeliness and security)
- usability: when service quality meets or exceeds customer needs (accessibility, reputation, ease of operation, etc)

But, identifying data quality metadata has gone past the ontological domain to focus on data quality standards. For instance, Berti-Equille uses the concept of “data quality contract” to be used in a standardized user oriented application where a “data quality acceptability” and a set of dimensions and assessment criteria are pre-defined by the underlying user [4]. And a more concrete data evaluation standardization has been proposed by Batini in 2008 where he introduces a model which allows to uniformly define information quality dimensions related to heterogeneous types of information, such as structured data managed in databases, semi-structured and unstructured texts and images [1].

3.3- Data warehouses data quality tools

Dealing with data quality in data warehouses is basically due to the multi-sourced character of this architecture. In fact, the data warehouse is fed, daily, with an orders extract which comes, generally, from a heterogeneous source. Unfortunately, the data quality in that extract is poor as the source system does not perform much consistency checks and there are no data dictionaries. The data quality problems that need to be addressed handle then:

- data quality (value quality and format quality) where we can use classical data quality dimensions and metrics
- and, source quality, especially, source reputation, believability, objectivity, and reliability.

Generally, this concern is handled by ETL (Extract-Transform-Load) tools as well as data cleansing, data analysis, data standardization, and data validation processes.
In fact, according to [31], invoking data quality processes at ETL time is compelling. Converging includes:

- design time integration: where it's interesting to support a diversity of transformations, including those relevant to data quality
- execution time integration: where data quality processes are applied in the application that is generated and promoted to production
- metadata integration: where information is stored in a local meta data repository that is able to be interchanged thanks to bridges in a federated design and execution environment

In 2008, Fon Silvers, suggested a set of dimensions related to each ETL step, in the context of the data warehouse integration process [24]. He, thus, defined:

- the data model's related dimensions where data model identifies the main subject areas of the data warehouse. Dimensional data modelling focuses, then, on the business activities of an enterprise.
- the ETL extract dimensions dealing with source data validation
- the ETL transform ones performing inspection, cleansing and conforming source data to the needs of the data warehouse

Besides, to address the source quality evaluation, Fon Silvers evoked some source system analysis methods such as:

- data profiling: consisting in a static view of the enterprise through its data
- data flow diagram: which is a dynamic view of the enterprise through its data in motion
- data state diagram: which is a dynamic view of the enterprise through its data in motion and business relevance and meaning
- system of records: which is a discernment of the authoritative data within an enterprise

Always in a heterogeneous and multiple sourced information systems, Naumann, Leser and Freytag distinguish in [17] three data evaluation axes:

1- source-specific criteria, when one have to study the ease of understanding, reputation, timeliness and reliability of the data source
2- attribute-specific criteria focusing on the completeness and the amount of extracted data
3- query correspondence assertions criteria where the availability, price, representational consistency, response time, accuracy and relevancy measure of the query's response are crucial

Moreover, various studies have focused on the integration schema quality as a decisive criterion on the loaded and, eventually, manipulated data quality.

A summary of some schema assessment criteria is described in Appendix2 where we notice that data quality metrics definitions in a schema integration context are totally different from the ones related to other contexts.

3.4- Total data quality methodologies

This category of the data quality assessment methods consists in evaluating all the workflow of data in an information system. Data quality methodologies were addressed for each domain activity. In fact, in 1999, English presented the Total Quality Data Methodology (TQDM), initially conceived for data integration in a data warehousing context [9].

Then, in 2000, Shankaranarayan, Wang and Ziad identified the Total Data Quality Management approach (TDQM) that was initially conceived as a research activity widely used in several applications domains [23].

In 2006, Falorsi and Scannapieco defined the ISTAT approach that concerns inter-organizational information systems and was first specialized in controlling Address/Localization data items [32]

Besides, many frameworks have been identified. In 2002, [16] evoked the Data Quality Broker (DQB) as a solution of the poor data quality in cooperative information systems. The DQB is intended to ensure and propagate the highest quality data by a feedback mechanism that filters the normal interaction with the provider organization. The design of this solution is presented, in a multi-source environment, by specifying how the brokering and the improvement functions are provided and implemented through each distributed protocol of the organization.

Another framework is suggested in 2006 by Savla and Ninan based on a continuous and iterative framework that can help IT organizations control the data quality. That framework is based on the following steps: [21]

1- Detection of the data quality issues by continuously monitoring tools.
2- Correction of the data anomalies which is generally automatic but requiring user intervention in the case of numerous curative actions.
3- Measurement to entertain and enhance the level of resources' quality.
4- Learning to improve the ability to detect and correct anomalies by deploying additional tools and by reforming business processes.

As a conclusion, these data quality methodologies have as a common goal the definition of the more appropriate assessment tools for the task or activity at hand as well as the maintain and the supervision of the global quality of the information system.

We have stated in this section the scope of data quality assessment. We define, in the following, our perspective approach for the evaluation and enhancement of the B-to-B application described in the second section.

4- TOWARDS AN E-BUSINESS ORIENTED DATA QUALITY FRAMEWORK: PROJECT PERSPECTIVES

As described in the latter sections, our main purpose in this project focuses on the enhancement of the clients' marketing campaigns, and especially the improvement of the prospects selection process from the various data vendors' files.

For this sake, we intend to create a database to save all information related to the previous marketing campaigns such as the targeted records, their sources, the amount of data, the undeliverable mails rate, the erroneous phone numbers rate, the correct phone numbers, the cost of the campaigns (records prices and operations costs), etc.

The assessment of each campaign is measured by the ROI (Return On Investment) indicator defined by the following formula:

ROI = (Gain from investment – Cost of investment) / Cost of investment

According to the ROI values, we can appraise the success or the failure of the underlying marketing campaign, and, given the saved data fields, we can determine the inaccurate record and the inefficient data source.

Moreover, to insure a good quality of the clients files, data append processes are handled. Data append consists on "enhancing the customer and prospect file with additional data fields supplied by commercial data compilers". [12]

And, as most B-to-B marketers find the greatest value from knowing the size and industry of their customer and prospect companies, the enriched data fields can, for instance, be: the business status code (headquarters, subsidiary), geocode, enterprise indicator (SIRET), new business code, phone number, post office box, sales volume, etc.

To succeed in such an activity, many steps have to be undertaken:[12]

1. Cleaning up the underlying file (using data anomalies' detection and correction tools)
2. Creating a list of the data vendors (with the list of fields that each could provide)
3. Performing a test of several vendors in order to identify the best vendor for a given client
4. Running a quality evaluation consisting of:
    a. Match rate: defined as the number of the records identified as also appearing in the vendor's database, divided by the number of records the vendor received from you.
    b. Hit rate: defined as the number of matched records that also had the required fields, available for append, divided by the number of records the vendor was able to match against the vendor's database.

    c. Accuracy: expect a certain amount of incorrect data to be part of the append process.
    d. Price
5. Assessing a test results: based on the vendors' performance against the selection criteria, the selection decision should be fairly simple, as a combination of performance and price.

In our application, data append is nearly always done using the SIRENE file. In fact, such file is updated monthly and contains accurate information. Besides, the record matching with the SIRNE file is efficient as done on the unique SIRET identifier. This record linkage allows, aside from the data append, to validate the data values of the clients' file and, therefore, to assess the quality level of the file (or the source).

Furthermore, we intend to assess the quality level of a given data vendor file, we intend to compare the underlying file with its latest version in order to conclude about its updating frequency, and fields volatility.

Given the previous perspectives, we define four classes of quality concerns with the underlying dimensions and metrics. We will use both intuitive (expert's experience based) and empirical (final user's needs based) approaches in order to define the suitable dimensions and metrics for the data quality assessment task. In fact, experts' standpoint is crucial for the good functioning of the brokering system as the selection of the most accurate, complete and reliable information, while the users' point of views help in producing a successful data marketing campaign:

1. Source quality: where we are interested in the following dimensions
    a. Reputation: based on the experts' knowledge acquired by word of mouth
    b. Credibility: based on the previous campaigns' results (accuracy rate, NPAI rate, etc)
    c. Added value: when a data source provides a unique information (not found in others data files)
    d. Price: that depends on the given data fields, the files' specialization, the amount of data, etc.
    e. Appropriate amount of data
    f. Files freshness
2. Data quality: where we are interested in the following dimensions
    a. Accuracy: syntactic accuracy
    b. Timeliness in the sense of freshness of data
    c. Consistency
    d. Usability: that can be measured through the rate of undeliverable mails or erroneous addresses (mails and e-mails) or unusability because of legal concerns (opt-in e-mail addresses, stop mailing, etc)
3. Query system quality: where we are interested in:
    a. Accuracy
    b. Reliability of the query system (stable system, security)
    c. Accessibility
    d. Security of access
    e. Uniqueness in the selected prospects
    f. Understandability
    g. Availability of data
    h. Conviviality of the interface
    i. Consistency of the results
    j. ROI regarding the targeted prospects
4. User of the query system
    a. Clients' profile
    b. Data quality contract corresponding to the client's expectances towards the underlying marketing campaign

In this paper, we described our project context and defined a first approach to improve the marketing campaign; the purpose was to automate the process of the prospects selection as an e-business application. Obviously, Internet and e-business will add new complexities to the application.
In the following studies, we concentrate on the conceptualization of the general process.

5- CONCLUSION

This paper states the main problematic of a thesis project where we aim to put in place a data quality assessment tool for an e-business application. Concretely, our challenge consists in controlling some DQ features (accuracy, freshness completeness, volatility, source reputation, etc) in order to succeed a direct marketing campaign and make a satisfying ROI.
In this introductory paper, we defined our underlying application and described the main issues that are encountered in practice. We also stated the large outfit of data quality evaluation tools where we distinguish a "first-generation data quality approaches" that aim to detect and correct errors and anomalies; and are mainly interested on the accuracy and consistency issues; and a "second-generation data quality systems" that intend to prevent errors in newly-created data and consists of the analytical workflows and conceptualized frameworks [20].
In the following works, we are intending to develop the adaptive data quality broker based, on one hand, on the client's "data quality contract" (related to the clients' profile and expectances) and, on the other hand, on the experts (system users) "data quality acceptability".
From a technical point of view, this query system has to face data quality issues as well as the query execution delays that select the targeted prospects.

6- APPENDICES

6.1- Appendix1

An example of some data quality dimensions:

- Accuracy: referring to the correct representation of the real-life phenomenon [2]
- Timeliness: expressing how current data are for the task at hand. It's also expressed as the delay between a change of the real-world state and the resulting modification of the information system state [25]
- Completeness: describing the ability of an information system to represent every meaningful state of the represented real world system [25]
- Consistency: capturing the violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file
- Synchronization: concerns the synchronization between different time series concerning proper integration of data having different time stamps
- Value-added: Data are beneficial and provide advantages for their use
- Portability: The format can be applied to as a wide set of situations as possible [19]
- ...

An example of some common used dimensions metrics:

- Completeness: Percentage of null values [4]
- Freshness: Update frequency standard deviation= frequency of an object instance update since its creation / average frequency of the objects of the same granularity [4]

- Syntactic correctness: Percentage of format discordances , syntactical errors and misspellings [4]
- Accuracy: Outlier probability: the probability of being an outlier in the attribute domain (using the IQR: InterQuartile Range) [4]
- Uniqueness: Duplicates detection probability using AR and clustering algorithms that combine values that are similar or identical [4]
- Currency: Currency=Age+(Delivery time – Input time) [13]
- Volatility : Time period during which data remain valid
- Timeliness: $\left\{\max\left[0,\left(1-\frac{\text{Currency}}{\text{Volatility}}\right)\right]\right\}^{s}$ where *s* is a control factor depending on the sensibility of the ratio.
- Consistency: CODD integrity constraints (syntactic) or business constraints satisfaction (semantic)
- …

6.2- Appendix2

<table>
<tr><td colspan="3">Minimality [6]</td><td>The extent in which the schema is modeled without redundancies.
E.g.: 1 – (nb redundant schema elements / nb total schema elements)</td></tr>
<tr><td colspan="3">Schema completeness [6]</td><td>Comparison between the real world and the representation = % of the real world object modeled in the integrated schema that can be found in the sources.
E.g.: 1 – (nb incomplete items / nb total items)</td></tr>
<tr><td colspan="3">Type consistency [6]</td><td>The extent in which the attributes corresponding to a real world object is represented with the same data type across all schemas of a data integration system.
E.g.: 1 – (nb inconsistent schema elements / nb total schema elements)</td></tr>
<tr><td colspan="3">Column heterogeneity [7]</td><td>This measure seeks for DQ problems that can arise when merging data from different sources. The measure is based on a combination of information theory and data mining tools: it uses a soft clustering approach (based on Information Bottleneck Method) to spate the data values in distinguished groups; then uses the cluster entropy to compute the heterogeneity in each cluster.</td></tr>
<tr><td colspan="2">Source data Freshness [18]</td><td></td><td>The data freshness is linked to the data integration system and defined in the works of Peralta and Bouzeghoub at the source level.</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Frequently-changing data</td><td>Currency [22]</td><td>Time elapsed since data was extracted from the source.
Currency=query time – extraction time</td></tr>
<tr><td>Obsolescence [11]</td><td>Number of updates transactions (or operations) since the data extraction time.
Metrics: Log files, change detection techniques</td></tr>
</table>

		Freshness rate [5]	Percentage of tuples in the view that are up-to-date (have not been updated since extraction time) % of extracted tuples that are up-to-date
	Long-term changing data and stable data	Timeliness [17]	The time elapsed from the last update to a source. Timeliness=query time - **last update times**

The first 3 measures, that are more detailed in [6], deals with DQ issues basically associated to the representation of real world objects. The minimality score insures the optimization of the query executions.

7- BIBLIOGRAPHY

[1] Batini C., Barone D., Cabitza F., Ciocca G., Marini F., Pasi G., Schettini R. "Toward a Unified Model for Information Quality". *VLDB'08*. 2008.

[2] Batini C., Scannapieco M. "Data Quality : Concepts, methodologies and Techniques Data-centric systems and applications". Springer-Verlag. 2006. p161.

[3] Bert-Equille L. "La qualité des données et leur recommendation: modèle conceptuel, formalisation et application à la veille technologique". *PHD thesis*. 1999. pp171-181.

[4] Berti-Equille L. "Measuring and Constraining Data Quality with Analytic Workflows". *VLDB'08*. 2008.

[5] Cho J., Garcia-Molina H. "Synchronizing a database to improve freshness". *SIGMOD'00*. 2000.

[6] Da Conceição M., Batista M., Salgado A.C. "Information quality measurement in data integration schemas". *VLDB'07*. 2007.

[7] Dai B.T, Koudas N., Ooi B.C., Srivastava D., Venkatasubramanian S. "Column heterogenetity as a measure of data quality". *CeanDB'06*. 2006

[8] Devillers R., Bédard Y., Jeansoulin R., Moulin B. "Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data". *International Journal of Geographical Information Science*. 2007. pp261-282

[9] English, "Improving data warehouse and business information quality", Wiley & sons, 1999

[10] Frank A. U. "Data Quality Ontology: An ontology for imperfect knowledge". *Springer Berlin / Heidelberg*. 2007. pp406-420.

[11] Gal A. "Obsolescent materialized views in query processing of enterprise information systems". *CIKM'99*. 1999.

[12] Grossman B., Stevens R.P. "Enhancing your B-to-B database with data append", *4th part of the series Business-to-Business database marketing*, 2007

[13] Lee Y. W., Pipino L., Funk J.D., Wang R.Y."Journey to Data Quality". 2006

[14] Lee Y. W., Strong D. M., Kahn B. K., and Wang R. Y. "AIMQ: A Methodology for Information Quality Assessment". *Information and Management*. 2001.

[15] Marco D., Jennings M. "Implementing data quality through metadata". *An embarcadero technologies white paper*. 2006.

[16] Mecella1 M., Scannapieco M., Virgillito A., Baldoni R., Catarci T., Batini C. "Managing Data Quality in Cooperative Information Systems". *LNCS 2519*, pp. 486–502, 2002.

[17] Naumann F., Leser U., Freytag J.C. "Qualiy-driven integration of heterogeneous information systems". *VLDB'99*. 1999.

[18] Peralta V., Bouzeghoub M. "On the evaluation of data freshness in data integration systems". *BDA'04*. 2004.

[19] Redman T.C. "Data quality for the information age". 1996

[20] Redman T.C. "Measuring Data Accuracy". *Information quality*. 2005. pp21-36.

[21] Savla S., Ninan M. "Data quality control strategies". 2006.

[22] Segev A., Weiping F. "Currency-Based Updates to Distributed Materialized Views". *ICDE'90*. 1990.

[23] Shankaranarayan G., Wang R.Y., Ziad M. "Modeling the manufacture of an information product with IP-MAP". *ICID'00*. 2000.

[24] Silvers F. “Building and maintaining a Data Warehouse”. 2008.
[25] Wand Y., Wang R. Y. “Anchoring Data Quality Dimensions in Ontological Foundations”. 1996.
[27] Wang R.Y., and Strong D.M. “Beyond Accuracy: What Data Quality Means to Data Consumers.” *Journal of Management Information Systems 12.* 1996. pp5–33.

8- INTERNET REFERENCES

[27] http://searchcio.techtarget.com/sDefinition/0,,sid182_gci212026,00.html
[28] http://wiki.gbif.org/ecatwiki/wikka.php?wakka=DataQuality
[29] http://www.bligg.fr/note/Lionel-SCHWARTZ---management-et-projets-Systemes-d%27Information/B2B:-La-qualite-des-bases-de-donnees-fait-defaut/1821912.html
[30] http://www.census.gov/quality/P01-0_v1.3_Definition_of_Quality.pdf
[31] http://www.information-management.com/issues/20031101/7625-1.html
[32] http://www.istat.it/dati/pubbsci/contributi/Contr_anno2005.htm
[33] http://www.investorwords.com/4316/ROI.html
[34] http://users.southeasttech.com/Roger.Morris/CIS265/Wk10/EnterpriseIntegration.pdf
[35] http://www.destinationcrm.com/Articles/ReadArticle.aspx?ArticleID=47778