A Comprehensive Data Quality Methodology for Web and Structured Data

Carlo Batini, Federico Cabitza
Universit degli Studi di Milano Bicocca
Piazza dell'Ateneo Nuovo, 1
20126, Milano, Italy {batini, cabitza}@disco.unimib.it

Cinzia Cappiello, Chiara Francalanci
Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio 34/5
20133 Milano, Italy
{cappiello, francalanci}@elet.polimi.it

Abstract

Measuring and improving data quality in an organization or in a group of interacting organizations is a complex task. Several methodologies have been developed in the past providing a basis for the definition of a complete data quality program applying assessment and improvement techniques in order to guarantee high data quality levels. Since the main limitation of existing approaches is their specialization on specific issues or contexts, this paper presents the Comprehensive Data Quality (CDQ) methodology that aims at integrating and enhancing the phases, techniques and tools proposed by previous approaches. CDQ methodology is conceived to be at the same time complete, flexible and simple to apply. Completeness is achieved by considering existing techniques and tools and integrating them in a framework that can work in both intra and inter organizational contexts, and can be applied to all types of data. The methodology is flexible since it supports the user in the selection of the most suitable techniques and tools within each phase and in any context. Finally, CDQ is simple since it is organized in phases and each phase is char-acterized by a specific goal and techniques to apply. The methodology is explained by means of a running example.

1. Introduction

Assessing and improving data quality (DQ) is still a complex task, especially in modern organizations where data are ubiquitous and diverse. Data are processed inside organizations, but they can also be provided to, as well as taken from, other organizations and, hence, affect the quality of organizational services. Consequently, researchers have collected clear evidence that poor data quality can have a negative impact on customer satisfaction and, thus, competitiveness [22, 12]. There is a general agreement in the literature on the business value of DQ and on the cost implications of data errors for both internal and external users.

Researchers and practitioners have proposed several methodologies supporting the selection and implementation of the most effective and economical DQ measurement and improvement techniques. Existing DQ methodologies can be classified according to different criteria. Methodologies can be either general-purpose or specialized: while the former encompass a wide range of phases and quality dimensions and provide a number of domain-independent techniques, the latter focus on specific data quality tasks (e.g. either measurement or assessment), on a specific type of data (e.g. a set of relational tables or a set of Web pages), and on a specific application domain. Methodologies can also refer to a single organizational information system or to inter organizational information systems and, thus, focus on a group of organizations that are supposed to cooperate towards a common goal.

In the literature, the main general-purpose methodologies are the TDQM methodology [27, 26], initially a research methodology and currently a full-fledged consulting approach, and the TIQM methodology [12, 23], designed as a tool for organizational managers. The Total Data Quality Management (TDQM) methodology [28] can be seen as an extension to data of the Total Quality Management approach [24], originally proposed for manufacturing products. The core concept of TDQM is the "Information Product" (IP), a model that considers data as a particular output of manufacturing processes. In fact, this methodology uses suitable tools such as IP-Map [26] and control matrices [19] in order to diagram and analyze the process by which an information product is manufactured. In details, TDQM proposes four phases to manage a generic IP along its life cycle: definition, measurement, analysis, and improvement. These phases are executed iteratively in order to implement a continuous quality improvement program. In the definition phase, data quality dimensions are identified and quality requirements are specified along selected dimensions. The measurement phase produces the actual quality values, to be compared against equirements. The analysis phase identifies the

1-4244-0682-X/06/$20.00 ©2006 IEEE

causes of quality problems, which can be either anomalies in business processes or database errors. In the improvement phase, quality improvement activities are implemented.

The TIQM methodology [12, 13] has been initially conceived for data warehouse projects, but its broad scope and level of detail characterize it as general purpose. In particular, the main contributions of TIQM are a set of specific techniques for cost-benefit analysis and a general managerial perspective. The cost-benefit analysis is conducted by analyzing non-quality information costs and classifying them into three categories [12]: *process failure costs*, arising when poor quality information causes process errors (e.g., inaccurate mailing addresses cause wrong deliveries); *information scrap and rework*, occurring when poor quality information causes rework and additional repair activities (e.g., correcting a mailing address in a database and resending a letter to the correct address); *loss and missed opportunity costs*, corresponding to loss of profits due to poor information quality (e.g., a lower hit-rate of marketing campaigns due to inaccurate mailing addresses). Due to its managerial perspective, TIQM provides detailed indications on how to apply the methodology by building consensus within an organization and guidelines for monitoring the effectiveness of data quality improvement initiatives.

A few inter-organizational methodologies have been proposed in the literature, typically designed for cooperating public institutions (c.f. the British Better Regulation Executive (BRE)[1]) and emphasizing the social implications of data quality. Inter-organizational domains are usually more complex since processes and services are interrelated, competencies and functionalities are distributed across different institutions and interoperability may not be guaranteed. This greater complexity makes the assessment phase more difficult, as reconstructing the state of all data sets and flows may not be straightforward. For example, the Istat methodology [8], designed for the Italian Public Administration, distinguishes an initial analysis of inter-organizational processes, data and data flows and a subsequent identification and correction of errors within local databases owned by distributed agencies.

This paper presents a Comprehensive Data Quality methodology, called *CDQ*, that aims at integrating the phases, techniques, and tools proposed by previous approaches. This integration is required to overcome the following limitations of existing methodologies. First of all, previous approaches do not tie quality improvement to explicit quality targets set by organizational management according to contingent requirements and cost constraints. These targets can be achieved with multiple quality improvement processes. The literature does not provide support to select the optimal quality improvement process that maximizes benefits within given budget limits. In the CDQ methodology the cost-benefit analysis is intensively used in different steps to define the data quality targets on the basis of the available funds and to guide the selection of the most suitable improvement process. Second, previous approaches do not emphasize the initial requirements elicitation phase. Implicitly, they assume that contextual knowledge has been previously gathered and modelled. The focus is on how to reach total data quality without providing indications as to how to use contextual knowledge. The CDQ methodology aims at providing precise guidelines and techniques to analyze the business contexts along all relevant aspects related to DQ issues. However, it must be noted that existing methodologies represent the outcome of their authors' experience with DQ management. They have been accurately tested over extended periods of time. This bottom-up inference of methodological guidelines from practice represents an empirical validation, but is also the cause for the specialization of existing approaches on specific issues or contexts. Conversely, our approach is aimed at generality, is top-down, but has undergone only partial testing [5]. The integration of previous methodologies is performed by:(i) organizing methodological phases within a complete framework, which can work in both intra and inter organizational contexts; (ii) supporting the selection of the most suitable techniques and tools within each phase; (iii) extending existing techniques and tools to work with different types of data, by adopting the widely accepted distinction among:

a. structured data i.e. data for which a formal schema is defined in terms of type, format, relationships, and constraints (e.g., relational tables in any DBMS).

b. semi structured data also called "schemaless" or "self-describing" data [3, 10], i.e. data that are either partly structured [11, 3] or have a descriptive rather than prescriptive schema [2]. Any web page or XML file can be considered an instance of semi-structured data, where unstructured data are provided within a precise structure defined by tags and anchors.

c. unstructured data, i.e. any sequence of symbols, typically coded in natural language with no (explicit) schema or structure (e.g, text documents, pictures, audio, video and binary files).

In the following, we use the term "data" to denote any type of information, either structured, semi-structured or unstructured. Since data quality techniques can change significantly according to the type of data, our methodology supports the selection of DQ techniques and tools by taking into account the type of data handled by different processes. Note that the importance of unstructured data can be real-ized by considering that almost 85% of all organizational data and up to the 40% of employees' work is spent on unstructured

[1] http://www.cabinetoffice.gov.uk/regulation/

documents [9]. These considerations are supported by a quantification of costs: according to an ICD survey, a 1000-employee enterprise loses a minimum of $6 million a year in the time spent by workers searching for and not finding needed information [21, 1]. These problems are expected to increase with the continuous growth of unstructured data in any organizational domain (e.g., emails, web pages). Unfortunately, existing methodologies mostly focus on structured data [12, 22, 28, 8]. A fundamental goal of CDQ is to support the implementation of a DQ program for all types of organizational data. The following section describes the CDQ methodology that is an attempt to address the issues discussed above.

2. The methodology

The main rationale underlying this methodology is to achieve a reasonable balance between completeness and feasibility. To explain the methodology, we refer to the cooperative scenario described below.

> Cooperative Information Systems (CIS) are defined as systems involving multiple organizations that interact to reach a common goal. The methodology will be explained by referring to Government to Business CISs. In general, businesses must interact with several agencies in order to request administrative services. In our scenario, we consider three agencies: the Social Security Agency, the Accident Insurance Agency, and the Chamber of Commerce. These agencies interact with businesses and among each other in different situations. For example, when a new business starts or an existing business is closed, it is necessary to inform the Chamber of Commerce. In turn, the Chamber of Commerce provides general business information to the Accident Insurance Agency in case of casualty. Furthermore, in the interaction with businesses, agencies manage both agency-specific information, such as tax reports, and common information, such as addresses, legal forms, and milestone dates (e.g. start-up date).

The methodology consists of three main phases:

1. State reconstruction where all the relationships among organizational units, processes, services, data sources, and data are reconstructed and documented.
2. Assessment where DQ dimensions are measured and assessed in order to set new data quality targets.
3. Choice of the optimal improvement process where improvement activities are selected by evaluating their cost/benefit ratio.

Each of these phases is decomposed into more detailed steps as described in the next sections. With respect to previous literature, we emphasize state reconstruction activities, since we assume that methodological activities start from a clean slate. This adds to the generality of our approach, which can be applied in both new and well-known contexts.

The different types of knowledge in the CDQ methodology and their mutual relationships are shown in Figure 1. The main concepts of Figure 1 are:

1. The organization or the set of organizations involved in data processing, with its/their organizational structure and norms.
2. The most relevant business processes performed by the organization and their interaction within overall macro-processes producing services or goods for both internal and external clients.
3. The quality of processes, macro-processes and services, e.g. the execution time of a process, the usability of a service, the correctness and completeness of a piece of information.
4. The internal data sources, corresponding to all databases and data flows of interest belonging to the organization.
5. The external data sources, corresponding to all databases and data flows of interest that are outside of or-ganizational boundaries.
6. The data quality dimensions along which the quality of data is assessed.
7. The data quality activities that can be performed in order to improve the quality of data. An ordered set of DQ activities defines an improvement process.
8. Costs and benefits of data and processes. Two cost categories are considered: costs due to poor DQ and costs of the quality improvement process. Benefits (savings and/or increased revenues) are supposed to result from the use of higher-quality data.

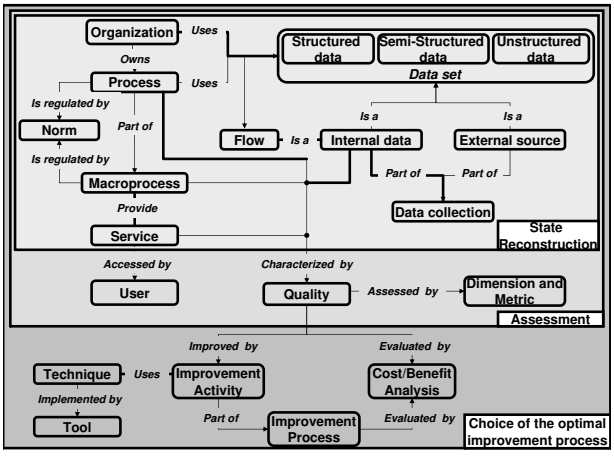

Figure 1. Knowledge involved in the CDQ methodology

Note that internal and external data sources can provide structured, semi-structured or unstructured data. As a consequence, a general-purpose methodology for DQ should

receive comprehensive input information on the organization or set of cooperating organizations it is applied to. The input/output structure of the CDQ methodology is shown in Figure 2.

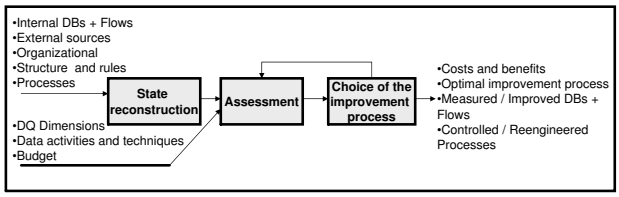

Figure 2. Inputs and outputs of the CDQ methodology

2.1 State reconstruction

In the state reconstruction phase, the organizational context modelled in terms of organizational units, processes, and rules, is related to the internal and external data involved in the production and delivery of goods and services. This phase provides a complete picture of data providers and users, of the data flows among them and of the use of data.

2.1.1 Focus on data

In the first step of this phase, we reconstruct the state of data and their use. This knowledge can be represented by means of two matrices: the Data/OrganizationalUnit matrix and the DataFlow/OrganizationalUnit matrix, modelling the relationships represented in Figure 1 linking organizations to data and data flows. Each cell of the Data/OrganizationalUnit matrix specifies whether an organizational unit either creates (i.e., owns) or just uses a set of data. Each cell of the DataFlow/OrganizationalUnit matrix specifies whether an organizational unit is either a user or a provider of a data flow. In case of structured data, both matrices can be refined to different levels of detail according to organizational requirements. For example, if data are structured, the Data/OrganizationalUnit matrix can be refined to the database or to the table level. The database level is typically adequate in most contexts.

For all external data, the ExternalDataSource/Data matrix must be reconstructed to specify the source of external data. A data source can supply multiple data sets (e.g. geographical addresses, current laws and regulations) and the same data set can be provided by multiple external sources. A price is also associated with each cell of the matrix, when available. Price information is used in the cost-benefit analysis to guide the source selection process.

> In our reference scenario, the Data/OrganizationalUnit matrix contains the Social Security registry, the Accident Insurance registry, and the Chamber of Commerce registry which are created and used by corresponding agencies. Note that each agency has its own registry of businesses and no shared database exists. Concerning data flows, each agency receives service requests from businesses and conveys corresponding service information to businesses. For service requests, the DataFlow/OrganizationalUnit matrix assigns the ”User” role to all agencies and the ”Provider” role to businesses. For service information, agencies assume the ”Provider” role while businesses become ”Users”. In an e-Government scenario, legal regulations are fundamental. This data set can be provided by different external data sources that are also identified in this step. Note that legal regulations represent a typical instance of either semistructured or unstructured data, depending on whether they are numbered and organized into predefined sections, or textual.

2.1.2 Focus on processes

The second step is focused on the relationship between organizational units and processes (see Figure 1). For each process, owner and contributing units are identified and specified in the Process/OrganizationalUnit matrix. This matrix supports the assignment of the responsibilities for the improvement activities proposed in the next methodological phase. The Process/ExternalData matrix is also reconstructed to represent the relationship between outsourced data and internal processes which use and possibly transform external data according to organizational requirements.

> In our scenario, three processes using external data can be identified: the process updating the information on businesses, the process updating the information on the branches of governmental agencies and the process updating information on the interactions between businesses and agencies. All agencies are involved in these processes.

2.1.3 Focus on macro-processes, services and norms

In the last step of the state reconstruction phase, organizational macro-processes are modelled in terms of elementary processes (see Figure 1). For each macro-process, that is a set of processes cooperating in the production and delivery of the same services, it is necessary to identify:

- the set of processes that cooperate in the production of services and the structure of their workflows modelled at different levels of detail depending on organizational requirements (e.g., a simple sequence);
- the services produced by each macro-process together with information on their clients and on the organization responsible for the macro-process;
- the norms that discipline the high-level structure of the process.

In order to represent this information, the Macro-process/ServiceNormProcess matrix is built.

As stated before, each interaction between businesses and agencies is disciplined by a law or by organizational rules. For example, as regards the macro-process tracking a variation of business information and updating corresponding data, the Macro-process/ServiceNormProcess matrix specifies:

- the processes: the insertion of the information in the database, the provision of a receipt to the business and additional notifications if inconsistencies occur.
- the service: information variation;
- the norms: each agency has to be informed within 60 days after the corresponding event;

This phase of the CDQ Methodology contributes to provide a complete vision of organizational processes and, hence, supports decisions on quality improvement activities. In particular, the integrated view of norms and processes is useful if the organization adopts a process-driven improvement approach [22].

2.2 Assessment

The assessment phase consists of two steps.

2.2.1 Problem identification

This step involves both internal and external users in order to identify the most relevant DQ issues as perceived by all clients of organizational processes. Users are interviewed and general DQ issues are gathered. These interviews highlight quality problems, although they do not relate these shortcomings to their causes. Interviews with users are also the key to understand the consequences of poor DQ on the work of internal users and on the satisfaction of external users. The outcome of this first step of the assessment phase is the basis for the analysis of the processes identified in the reconstruction phase and the identification of the causes of poor DQ. For example, internal users may complain about the inaccuracy of the responses to their queries. A likely cause is the replication of data in multiple databases that are not integrated, but are all used by applications at different times along the same process. Similarly, external users may feel dissatisfied whenever they receive incorrect, incomplete, or out-of-date information or, more generally, whenever they are provided with information that does not satisfy their requirements. This quality issue can be either a service problem (e.g. wrong market segmentation), or a data problem (e.g. incomplete databases or errors in data updates, and so on).

In our reference scenario, the results of interviews may highlight the following data quality issues: duplicate objects in individual databases, non-matching objects in the three duplicated registries of businesses, delays in the registration of updates and unavailability of updates in in force laws.
The first and second issues are related to the accuracy quality dimension while the third and the fourth issues are related to the currency dimension.

Since improvement activities are complex and expensive, identifying the precise set of data and processes that raise quality problems is a key factor for the effectiveness of the CDQ methodology.

2.2.2 DQ measurement

The second step aims at obtaining a quantitative evaluation of quality issues. For this purpose, it is necessary to select a subset of relevant DQ dimensions and related metrics. Then, quality metrics must be applied to data and data flows in order to provide a quantitative evaluation of the quality issues identified in the previous step.

Non matching objects in different databases can be classified as a case of inaccuracy and can be measured with the percentage of non-matching objects. The late delivery of an information to the final user can be classified as a lack of currency and measured as the time interval between the latest update and the time the user receives the data. At the end of the measurement process, we should be able to fill a table such as Table 1 by identifying the data sets responsible for poor quality through statistical analysis.

Table 1. Databases, quality dimensions and metrics

Quality dimension/Database	Duplicate objects	Matching objects	Accuracy	Currency
Social Security DB	5%		98%	3 months delay
Accident Insurance DB	8%		95%	5 months delay
Chamber of Commerce DB	1%		98%	10 months delay
The three databases		98%		

The identification of DQ dimensions and metrics varies with the type of data. For structured and semi-structured data, the quality of data is usually measured by means of quality dimensions such as accuracy, completeness, and currency since they are context independent and associated with consolidated assessment algorithms [7, 22, 17]. Anyway, in the literature, it is also possible to find contributions that provide general guidelines and indications for the assessment of both objective and subjective data quality dimensions [18, 20].

For unstructured data, assessment techniques are less consolidated. For example, it is inherently difficult to evaluate the accuracy and completeness of a text. The evaluation of currency is instead easier and more common. The frequency of updates can be easily measured for a text and related to specific benchmarks. The quality of unstructured data is widely analyzed in the archival domain. Archivists have the goal of preserving relevant digital and paper documents. They assess the relevance of a document by evaluating its quality by means of several dimensions such as *condition* (for non-digital documents), that refers to the physical suitability of the document for scanning, or *originality*, that refers to the reliability of the data source [15].

2.3 Choice of the optimal improvement process

The choice of the optimal improvement process phase is composed of five steps.

2.3.1 DQ requirement definition

Based on actual quality values, DQ_{ij}, associated with the i-th data set and the j-th quality dimension, the organization has to set target quality values DQ^*_{ij}, to be reached through the improvement process. DQ targets are defined by performing a *process-oriented* and a *cost-oriented analysis*, as described in the following.

The *process-oriented analysis* is heavily based on the information collected in the reconstruction phase, in particular data, processes, organizational units and their interrelations (see State Reconstruction phase). In order to define target quality values, it is necessary to fix the expected y-th process performance index $index_{xy}$ along each x-th process p_x. Considering the data d_i involved in the process, corresponding data quality dimensions DQ_{ij} to be improved can be identified and target values $DQ\_proc^*_{ij}$ can be determined. In order to relate process performance indexes with data quality measures, linear correlation indexes α_{xyij} are used:

$$DQ\_proc^*_{ij} = \alpha_{xyij} * index_{xy}$$

If multiple quality dimensions have to be improved, it is also possible to weigh each dimension based on the priorities of final users assessed in the previous phase. For the sake of simplicity, we do not consider the effects of possible correlations between DQ dimensions.

In the *cost-oriented analysis*, it is necessary to define the economic effort that the organization can afford in the DQ improvement process. A fundamental hurdle is that costs and benefits are difficult to estimate *ex ante*. We refer to *non-quality costs* as the costs associated with poor data quality and, consequently, with all the activities necessary to correct errors and re-execute tasks. Instead, quality costs are associated with the activities and resources necessary in the improvement project. Non-quality costs are represented in Figure 3. Note that non-quality costs (C_{ij}) are associated with actual data quality value DQ_{ij} (see (1) in Figure 3). Also note that each DQ dimension is characterized by a quality and a non-quality cost curve. For the sake of simplicity, all graphs will refer to a single dimension.

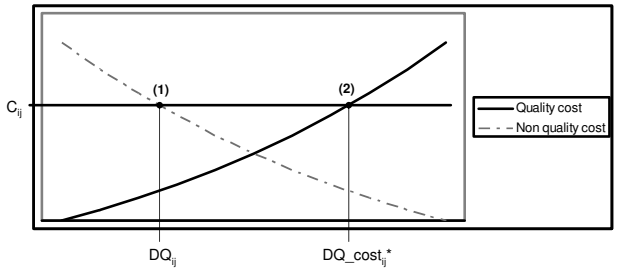

Figure 3. Quality and non-quality cost curves: definition of data quality targets

Non-quality costs can be considered as a potential saving and represent tangible benefits of quality improvement. In the worst case scenario, the benefits of the improvement process will be limited to the savings from non-quality costs. Additional tangible and intangible benefits can be achieved in higher-performance scenarios. Under this assumption, the sum of the costs C_{ij} along all the data sets and all the dimensions can be considered as the minimum budget allocated to the improvement process. From the quality cost curve, the minimum value of target quality $DQ\_cost^*_{ij}$ can be derived as the intersection of C_{ij} with the benchmark curve, accordingly (see (2) in Figure 3).

It must be noted that the quality cost curve depends on the improvement techniques that are implemented. Since the most suitable improvement techniques are not defined at this step, benchmark curves are used. These curves report the average cost associated with a sample of improvement projects for a specific quality dimension. Quality and non-quality cost curves are comparable if they refer to the same time interval, projected into the future for quality costs and projected into the past for non-quality costs.

The results of the process-oriented and cost-oriented analyses are then compared. In the first iteration of the methodology, the choice of the lowest target value is recommended for all quality dimensions: $if\ \ DQ\ proc^*_{ij} > DQ\ cost^*_{ij} \Longrightarrow\ DQ^*_{ij} = DQ\ cost^*_{ij}\ else\ \ DQ^*_{ij} = DQ\ proc^*_{ij}$ These targets represent a qualitative balance between the "100 per cent quality" (ideal and unreachable target) and the actual situation.

> The cost-oriented analysis shows that agencies spend about 10 million euro a year to correct and reconcile misaligned records. Furthermore, b ecause most tax

fraud prevention techniques rely on cross-referencing records over different agencies, misalignments may result in undetected tax frauds. At the end of the evaluation process, we should be able to set target values and fill a table similar to Table 2.

Table 2. Databases and target quality values

Quality dimension/Database	Duplicate objects	Matching objects	Accuracy	Currency
Social Security DB	1%		99%	3-4 days delay
Accident Insurance DB	1%		99%	3-4 days delay
Chambers of Commerce DB	0,3%		99%	2-3 days delay
The three databases		97%		

2.3.2 DQ improvement activity selection

The goal of this step is to understand which process-driven and which data-driven activities lead to the most effective results of the quality improvement process, given the targets set in the previous step. Concerning process-driven activities, a typical business process reengineering activity (see [14, 25], for a comprehensive discussion) is composed of three activities:

- Map and analyze the *as-is* process.
- Design the *to-be* process, according to one or multiple alternatives.
- Implement the reengineered process and continuously improve.

A business process reengineering activity is applied to the processes creating and transforming data. For example, in the case of structured data, if the main problem is an inconsistency among different overlapping sources, synchronization activities should be reengineered. As regards unstructured data, such as web pages, if the main problem is the publication of out-of-date information, the definition of the expiry date should become a mandatory activity and expired page should be removed with dedicated activities. In this way, the web page creation and maintenance process will be reengineered to satisfy temporal validity targets.

> A process-oriented improvement solution is to offer services through an infrastructure common to all agencies. Back-end activities must become cooperative and a notification infrastructure that guarantees the synchronization among update events should be implemented. One of the three agencies (e.g. the Chamber of Commerce) should be responsible for the collection of updated information from businesses and for the notification of updates to the correct agencies.

The selection of data-driven activities starts from the analysis of quality issues and corresponding causes provided by the Problem Identification phase. In the following, a few cases are discussed:

1. Let us assume that a certain data flow is characterized by very poor quality. We may perform a source selection activity on the data conveyed by the data flow. The goal of a *source selection* activity is to replace the actual source with one or more data sources that provide the requested information. Source selection criteria and techniques can be found in [16] and [17]. In [6] a linear integer programming technique is proposed to optimize the choice of sources by maximizing their cost to quality overall ratio. In [5] optimization and negotiation mechanisms are proposed to build the most suitable data set by integrating data from different providers.
2. If a relational table has low accuracy and we know that a different source could guarantee higher accuracy for a subset of all data, an *object identification* activity should be performed to merge data from different sources.
3. If a table characterized by low completeness is mainly used for statistical applications, an *error correction* (or data imputation) activity can be performed to change null values into valid values, without changing the statistical distribution of values [29].

The data-driven improvement activities listed above are general and valid for all types of data. For unstructured data derived from structured data, a specific data-driven improvement technique is *data profiling*. This technique is used in order to relate a text file to a database schema by discovering recurring patterns inside the free text [4].

> An object identification activities must be performed on the three databases. This involves two redesign alternatives (a) the creation of a central database integrating all types of information on businesses that is currently managed by three separate registries; (b) the creation of a light central database, called *Identifiers database* whose records identify corresponding business records managed by individual agencies. The first solution cannot be implemented, both for privacy reasons and for the preservation of agencies' autonomy. Furthermore, in the considered scenario, it is necessary to have a complete knowledge of in force laws. In order to have high quality information about current regulations, the agencies could refer to external sources and choose the most suitable one through a source selection activity.

At the end of this step, we should be able to produce an *Activity/Data matrix*, reporting a cross for all data sets or dataflows to which the activity applies. Note that a data set can include different types of data. For example, a web page

Data sets Data flows/ Activities	Type	Social Security DataBase	Accident Assurance Database	Chamber Of Commerce Database	Data flows Among agencies	The new Identifiers database	External Source
Object Identification for Stock data	Data driven	X	X	X			
Process Reengineering on update Processes	Process driven	X	X	X	X	X	
Source Selection	Data driven						X

Figure 4. The Activity/Data Matrix

can contain both structured and unstructured data. Some data in the page can be extracted by querying a database, but free text may also be published on the same page.

The Activity/Data Matrix of the reference scenario is shown in Figure 4.

The choice of the most suitable improvement techniques can be supported by specific heuristics. For instance, a high frequency of updates of a source represents an indication for process-driven improvement techniques. The orientation of an organization towards short- or long-term projects also provides indications: if the goal is a short-term project, data-driven improvement techniques should be selected.

2.3.3 Choice of the optimal technique for DQ activities This phase focuses on the selection of the best technique and tool for each activity in the *Activity/Data matrix* produced in the previous step. It is necessary to analyze the techniques implemented by commercial quality improvement tools and compare their costs and technical characteristics. The choice of a technique is influenced by the availability of corresponding affordable tools. With reference to the object identification activity, many commercial and open source tools include statistical techniques, tools adopting empirical and knowledge-based techniques are less common. If a tool can be personalized, it can be chosen and then adapted to organizational requirements. For instance, we could modify a "one shot" probabilistic technique by using a sequence of increasingly sophisticated linkage techniques, such as exact key matching or nearly exact key matching.

2.3.4 Choice of improvement processes In this phase, it is necessary to link crossings in the Activity/Data matrix by identifying candidate improvement processes. A property of a candidate improvement process is completeness, i.e. the inclusion of all databases and data flows involved in the improvement program. Usually, two or three candidate improvement processes are sufficient to cover all relevant choices.

In the reference scenario, a single improvement process is available (see Figure 5) where the number shows the

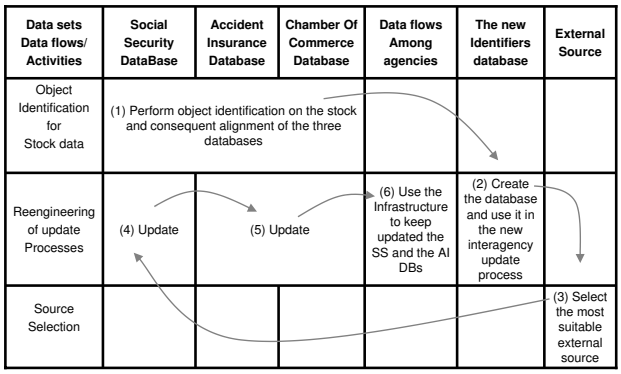

Figure 5. A sample improvement process

sequence of activity execution), including the reengineering of the update process, the object identification on the stock, and the selection of external source that provides correct in force laws. These activities have to be synchronized when the new system is deployed. Note that some activities could be performed in a parallel way since independent of each other (i.e. activity 2 and activity 3). Other possibilities, such as performing data integration, have been excluded in the DQ Improvement Activity Selection phase. Note that we do not need a periodic object identification, since business process reengineering permanently aligns the information owned by the three agencies.

2.3.5 Evaluation of improvement processes The costs of candidate improvement processes are compared in this step. It may happen that a business process reengineering activity enables a more efficient object identification activity and, therefore, should be executed first. Costs and benefits should be calculated accordingly. Costs should include technology costs due to hardware, software, and personnel. Costs should be compared with the net savings stemming from non-quality costs, that is the C_{ij} term
calculated in the sixth step. If costs are equal to savings, the most suitable solution has been found, since it allows it to reach the target quality values within minimum budget limits. The identified improvement solution might have additional intangible benefits that further justify the investment on quality. In case costs C^*_{ij} are lower, more ambitious target quality level DQ^{**}_{ij} can be identified by iterating the methodology from step 6.

If costs C^*_{ij} are higher than target values (see Figure 7), benchmark curves are applied and less ambitious target quality level DQ^{**}_{ij} must be set. Once again, the methodology must be iterated from step 6.

3 Conclusions

This paper proposes a methodology for DQ measurement and improvement. The DQ management strategy is

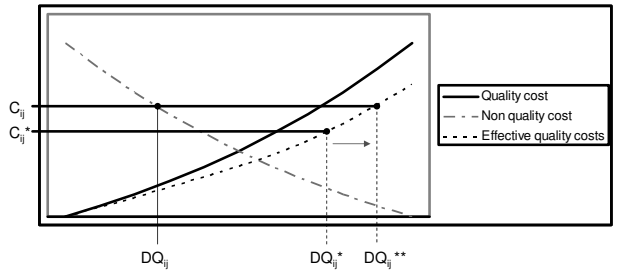

Figure 6. Evaluation of more ambitious quality targets

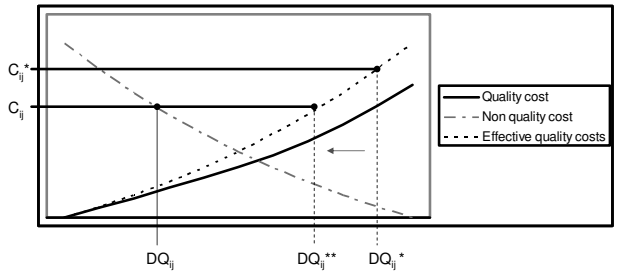

Figure 7. Evaluation of less ambitious quality targets

selected by evaluating quality and non-quality costs and by analyzing the trade-offs between alternative improvement techniques and ad-hoc amendment techniques. Future work will focus on individual phases of the methodology in order to perform in-depth analysis and provide more precise guidelines. A more complex cost-benefit model will also be developed, including benchmarks and quantitative evaluation techniques.

4. References

[1] The high cost of not finding information. Technical Report IDC 29127, International Data Corporation, April 2003.

[2] S. Abiteboul. Querying semi-structured data. In *Proc. of ICDT*, pages 1–18, 1997.

[3] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers, 2000.

[4] P. Aiken. *Data Reverse Engineering*. McGraw Hill, 1996.

[5] D. Ardagna, C. Cappiello, M. Comuzzi, C. Francalanci, and B. Pernici. A broker for selecting and provisioning high quality syndicated data. In *Proc. of the Tenth International Conference on Information Quality*, pages 262–279, 2005.

[6] A. Avenali, P. Bertolazzi, C. Batini, and P. Missier. A formulation of the data quality optimization problem in cooperative information systems. In *Proc. of the CAiSE workshops, Data and Information quality (DIQ '04)*, pages 49–63, 2004.

[7] D. Ballou, R. Wang, H. Pazer, and G. Tayi. Modelling information manufacturing systems to determine information product quality. *Management Science*, 44(4):462–533, 1998.

[8] M. Bertoletti, P. Missier, M. Scannapieco, P. Aimetti, and C. Batini. The service to businesses project: Improving government-to-business relationships in italy. In *Proc. of EGOV'03*, pages 468–471, 2003.

[9] R. Blumberg and S. Atre. The problem with unstructured data. *DM Review*, February 2003. SourceMedia Publishing.

[10] P. Buneman. Semistructured data. In *Proc. of PODS*, 1997.

[11] D. Calvanese, D. D. Giacomo, and M. Lenzerini. Modeling and querying semi-structured data. *Networking and Information Systems Journal*, 2(2):253–273, 1999.

[12] L. P. English. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc., New York, NY, USA, 1999.

[13] L. P. English. *Information and Database Quality*, chapter Total Quality data Management (TQdM), Methodology for Information Quality Improvement. Kluwer, 2002.

[14] M. Hammer and J. Champy. *Rengineering the Corporation: a Manifesto for Business Revolution*. 2001.

[15] H. Krawczyk and B. Wiszniewski. Visual gqm approach to quality-driven development of electronic documents. In *Proc. of the Second International Workshop on Web Document Analysis, co-located with ICDAR2003*, 2003.

[16] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*. LNCS 2261, 2002.

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

[18] F. Naumann and C. Rolker. Assessment methods for information quality criteria. In *Proc. of the Conference on Information Quality*, pages 148–162, 2000.

[19] E. M. Pierce. Assessing data quality with control matrices. *Communications of ACM*, 47(2):82–86, 2004.

[20] L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communication of the ACM*, 45(4):211–218, 2002.

[21] R. Rao. From unstructured data to actionable intelligence. *IT Professional*, 535(6):29–35, November/December 2003.

[22] T. Redman. *Data Quality in the Information Age*. Artech House, Boston, 1996.

[23] T. Redman. *Data Quality The field guide*. The Digital Press, 2001.

[24] J. E. Ross. *Total Quality Management: Text, Cases and Reading*. CRC Press.

[25] M. S, W. L., and H. C. S. Business process re-engineering : a consolidated methodology. In *Proc. of the 4th annual International Conference on Industrial Engineering Theory, Applications and Practise, San Antonio, Texax, USA*, 1999.

[26] G. Shankaranarayan, R. Y. Wang, and M. Ziad. Modeling the manufacture of an information product with ip-map. In *Proc. of the 5th International Conference on Information Quality*, Massachusetts Institute of Technology, USA, 2000.

[27] R. Y. Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2), 1998.

[28] R. Y. Wang and H. B. Kon. *Information technology in action: trends and perspectives*, chapter Toward total data quality management (TDQM), pages 179–197. Prentice-Hall, Inc., 1993.

[29] W. E. Winkler. Methods for evaluating and creating data quality. *Inf. Syst.*, 29(7):531–550, 2004.