

Data Quality Management and Evolution of Information Systems

Carlo Batini¹, Barbara Pernici²

¹ Università degli Studi di Milano Bicocca
Via degli Arcimboldi 8, Milano (Italy)
batini@disco.unimib.it

² Politecnico di Milano
Piazza Leonardo da Vinci 32, Milano (Italy)
barbara.pernici@polimi.it

WWW home page: <http://www.elet.polimi.it/people/pernici>

Abstract. Information systems have been rapidly evolving from monolithic/transactional to network/service based systems. The issue of data quality is becoming increasingly important, since information in new information systems is ubiquitous, diverse, uncontrolled. In the paper we examine data quality from the point of view of dimensions and methodologies proposed for data quality measurement and improvement. Dimensions and methodologies are examined in their relationship with the different types of data, from structured to unstructured, the evolution of information systems, and the diverse application areas.

1 Introduction

Managing data quality is a complex task, especially in modern organizations where information is ubiquitous and diverse. Information is processed inside organizations, but it can also be provided to 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 [1–3]. There is a general agreement in the literature on the business value of data quality (IQ) and on the cost implications of errors in information management for both internal and external users.

The evolution of information systems from systems which are accessed by a set of well identified users in a well defined context to services available online to different and unknown types of users has greatly increased the potential impact of poor data quality. Data quality requirements have changed accordingly to this different context of use, and information management has evolved and requires a more organized planning, monitoring, and control of data quality. The nature of data itself has

changed in information systems: different types of information, different types of information systems, wide increase of application domains are relevant issues.

The goal of this paper is to analyze data quality issues in information systems, in particular considering the evolution of information systems characteristics. Three coordinates of analysis are proposed: types of information, types of information systems, and application areas. Providing a complete map of the evolution of dimensions, techniques, methodologies, and tools is a complex issue. In the following sections we focus on a few representative dimensions, highlighting the evolution path of data quality dimensions and methodologies for data quality measurement and improvement among the three coordinates of analysis introduced above.

The paper is organized as follows: in Section 2, we discuss the main dimensions of data quality proposed in the literature. The three coordinates for analyzing data quality in information systems proposed in the paper are discussed in Sections 3-5. Finally, in Section 6 we discuss approaches being developed to improve data quality in information systems.

2 Dimensions of Data Quality

In the data-quality literature, various authors have proposed a rich set of possible data-quality dimensions, considering conformance to specifications, user requirements, context of use, and so on; however, a comprehensive list of commonly agreed quality dimensions is still not available. Dimensions are distinguished typically in two categories: schema quality dimensions, that refer to the quality of the intensional representation of data, the schema, and data quality dimensions, that refer to the quality of the extensional representation of data, i.e. their values. In the following we focus mainly on data quality dimensions, which are more relevant in influencing the quality of business processes and services, and we will interchange the terms data and information depending on the context.

Lists of data dimensions can be found in [4], [5], [1], [6] [7], and [2]. For a comparison of the different definitions and classifications see also [8]. Data quality dimensions reported by most authors in the field are:

- *Accuracy*: “inaccuracy implies that the information system represents a real world state different from the one that should have been represented”. Inaccuracy refers to a garbled mapping into a wrong state of the information system, where it is possible to infer a valid state of the real world though not the correct one.
- *Timeliness*: refers to “the delay between a change of the real-world state and the resulting modification of the information system state”. Lack of timeliness may lead to an information system state that reflects a past real world state. A particular form of timeliness is Synchronization among different time series concerns proper integration of data having different time stamps.

- *Completeness*: is “the ability of an information system to represent every meaningful state of the represented real world system”. Completeness is of course tied to incomplete representations.
- *Consistency*: consistency of data values occurs if there is more than one state of the information system matching a state of the real world system, therefore “inconsistency would mean that the representation mapping is one-to-many”.
- *Interpretability* concerns the documentation and metadata that are available to interpret correctly the meaning and properties of data sources.
- *Accessibility* measures the ability of the user to access the data as from his/her own culture, physical status/functions and technologies available.
- *Usability* measures the effectiveness, efficiency, satisfaction with which specified users perceive and make use of data.
- *Trustworthiness* (of an organization) measures how reliable is the organization in providing data sources.

3 Types of Data

Researchers in the area of data quality must deal with a wide spectrum of possible data representations. A basic classification proposed by several authors is to distinguish, implicitly or explicitly, three types of data:

- 1 *structured*, when the data are distinguished in elementary items, and each of them is represented with a format that can be described by a grammar.
- 2 *semi structured*, when data has a structure which has some degree of flexibility. Semi structured data are also “schemaless” or “self-describing” data (see [9-11]).
- 3 *unstructured*, when data are expressed in natural language and no specific structure or domain types are defined.

Dimensions and techniques for data quality have to be adapted for the three types of data described above, and are progressively more complex to conceive and use from structured data to unstructured ones. 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. 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 timeliness is instead easier and more common. The frequency of updates can be easily measured for a text and related to benchmarks [12].

A second point of view sees data as a particular manufacturing product. This point of view is adopted. In the Information Production Map (IP-MAP) model ([13], an extension of the Information Manufacturing Product model [14]; the IP-MAP model identifies a parallelism between the quality of data, and the quality of products as

managed by manufacturing companies. In this model three different types of data are distinguished:

- *raw data items* are considered as the smaller data units. They are used to construct information and component data items that are semi-processed information;
- *component data items* are stored temporarily until the final product is manufactured. The component items are regenerated each time an information product is needed. The same set of raw data and component data items may be used (sometimes simultaneously) in the manufacturing of several different information products;
- *information products* are the result of a manufacturing activity performed on data.

Looking at data as a product, as we will see later in Section 6, allows exploiting well-established methodologies and procedures developed for quality assurance in manufacturing processes.

The third classification, proposed in [15], addresses a typical distinction made in information systems between elementary data and aggregated data. Elementary data are data managed in organizations by operational processes, and represent atomic phenomena of the real world (e.g., social security number, age, sex). Aggregated data are obtained from a collection of elementary data by applying some aggregation function to them (e.g. the average income of tax payers in a given city). This classification is useful to distinguish different levels of severity in measuring and achieving the quality of data. As an example, the accuracy of a field sex changes dramatically if we input M (male) instead of F (female); if the age of a single person is wrongly imputed as 25 instead of 35, the accuracy of the average age of a population of millions of inhabitants is minimally affected.

4 Types of Information Systems

Different criteria can be adopted in classifying the different types of information systems, and corresponding architectures; they are usually related to the overall organizational model and rules adopted in the organization or the set of the organizations that make use of the information system. In order to clarify the impact of data quality on the different types of information systems, we adapt the classification criteria proposed in [16] for distributed databases. Three different criteria are proposed: distribution, heterogeneity, autonomy.

Distribution deals with the possibility of distributing the data and the applications over a network of computers. *Heterogeneity* considers all types of semantic and technological diversities among systems used in modeling and physically representing data. *Autonomy* has to do with the degree of hierarchy and rules of coordination, establishing rights and duties, defined among the organization using the information system. The two extremes are: (i) a fully hierarchical system, where

only one subject decides for all, and no autonomy at all exists; and (ii) a total anarchy, where no rule exists, and each component organization is totally free in its design and management decisions.

The three classifications are represented all together in the classification space of Figure 1. Among all possible combinations, five main types of information systems are highlighted in the figure: Monolithic, Distributed, Data Warehouses, Cooperative, Peer to Peer, Service-oriented.

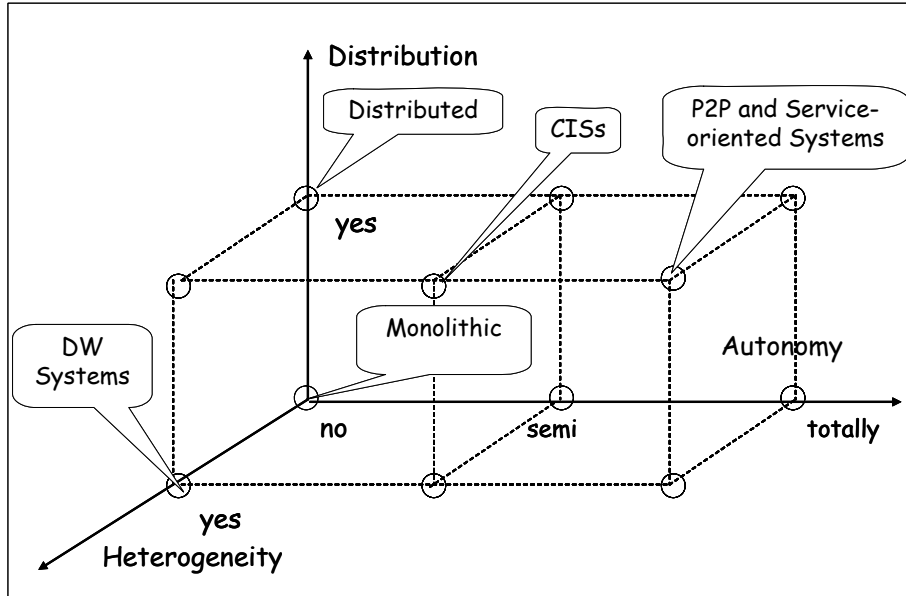


Fig. 1. Types of information systems

In a *Monolithic information system*, presentation, application logic, and data management are merged into a single computational node. Many monolithic information systems are still in use. While being extremely rigid, they provide advantages to organizations, such as reduced costs due to homogeneity of solutions and centralization of management. In monolithic systems data flows have common format, decided at design time, and data quality control is facilitated by the homogeneity and centralization of procedures and management rules.

A *Data Warehouse (DW)* is a centralized collection of data collected from different sources designed to support management decision making. Data warehouses contain a wide variety of data that present an integrated and coherent picture of all data used in operational processes of an organization, eventually enriched with external sources, at a single point in time. Development of a data warehouse includes development of systems to extract data from operating systems plus installation of a warehouse database system that provides managers flexible access to the data. The most critical problem in DW design concerns the cleaning and integration of the

disparate sources that are loaded in the DW, in that much of the implementation budget is spent on data cleaning activities.

A *Distributed information system* relaxes the rigid centralization of monolithic systems, in that it allows distribution of resources and applications among a network of systems, geographically distributed. Data design is usually performed centrally, but to a certain extent some degree of heterogeneity can occur, due to the impossibility to establish unified procedures. When the information system of an organization moves from a monolithic architecture to a distributed one, problems of data management become more complex, due to the reduced level of centralization. Heterogeneities and autonomy usually increase with the number of nodes. Data sources are typically characterized by various kinds of heterogeneities, that can be roughly classified into technological heterogeneities, schema heterogeneities and instance-level heterogeneities. Schema heterogeneity is principally due to the usage of (i) different data models that originate model conflicts, while Instance level heterogeneity deals with conflicts occurring on data values provided by different sources. This type of heterogeneity can be caused by accuracy, completeness, currency, and consistency errors, related to the processes that feed the different data sources.

A *Cooperative Information System (CIS)* can be defined as a large scale information system that interconnects various systems of different and autonomous organizations, geographically distributed while sharing common objectives [17]. Cooperation with other information systems presupposes the ability to exchange information and to make a system's own functionality available to other systems. These features are often referred to as inter-operability in the literature and should be treated as prerequisites to cooperation". The relationship between cooperative information systems and data quality is double face: from one side one can take profit of the cooperation among agents in order to choose best quality sources, and improve in such a way the quality of circulating data; from the other side, data flows are less controlled then in monolithic systems, and the quality of data, when not controlled, may rapidly decrease in time. In cooperative information systems it is difficult, although crucial for achieving data quality, to individuate a subject, usually called data owner, that is responsible for a certain data category. In fact, data are typically replicated among the different participating organizations, and one does not know how to state that an organization or subject has the primary responsibility for some specific data. These aspects make difficult for a new user of the information system to become confident and assess the reputation of other users. Trustworthiness issues are also related to identification of ownership of data. Another issue crucial for data quality becoming relevant in cooperative information systems is to keep track of the provenance of data. Data provenance is defined in [18] as the "description of the origins of a piece of data and the process by which it arrived in the database". The typical mechanism used to trace the provenance is using annotations. The annotation concept can be used to represent a wide spectrum of information about data, such as comments or other types of metadata, and, in particular, quality data.

In a *Peer to Peer information system* (usually abbreviated into P2P), the traditional distinction, typical of distributed systems, between clients and servers is disappearing. Every node of the system plays the role of a client and a server. The node pays its participation in the global exchange community by providing access to its computing resources, without no obligation on the quality of its services, and data. A P2P system can be characterized by a number of properties: no central coordination, no central database, no peer has a global view of the system, global behavior emerges from local interactions, peers are autonomous, and peers and connections are unreliable. It is clear that P2P systems are extremely critical from the point of view of data quality, since no obligation exists for agents participating in the system, and it is costly and risky for a single agent to evaluate the reputation of other partners.

In a *Service-oriented information system* the functionalities of information systems are provided as services to external and potentially unknown users. Services are published in registries and a service-oriented architecture provides the infrastructure to select and invoke services. Services may be annotated with information about their quality-of-service characteristics, including data quality dimensions (see e.g. [19]). In this way the service provider can declare which are the quality characteristics of the service it provides, and the service user can use these annotations, both to select the services to use and as an element to verify the trustworthiness of the service provider.

5 Application Areas

Data quality has a leading role in a wide number of application domains, under different perspectives. In the following, we discuss the role of data quality in some of the most relevant application areas, selected among a wide spectrum of domains, and, specifically, e-Government, web applications, and Life Sciences.

5.1 e-Government

The main goal of all e-Government projects is the improvement of the relationship between government agencies and citizens, as well as between agencies and businesses, through the use of information and communication technologies. E-Government projects must face the problem that similar information about one citizen or business is likely to appear in multiple databases, each autonomously managed by different agencies that, historically, have never been able to share their data about the citizens and businesses. The problem of data quality in e-Government is worsened by the many errors usually contained in the legacy databases of agencies, due to many causes. First, due to the nature of the administrative flows, several citizens' (and businesses') data (e.g., addresses) tend to become stale and are not updated for long periods of time. Also, errors may occur when personal data are stored. Furthermore, data provided by distinct sources differ in format, following local conventions, that can even change in time and result in multiple versions. One

major consequence of having multiple disconnected views for the same information, is that citizens and businesses experience consistent service degradation during their interaction with the agencies.

5.2 Web Applications

Web applications are characterized by the presentation to a wide audience of a large amount of data, the quality of which can be very heterogeneous. There are several reasons for this variety, but a significant reason is the conflict between two needs. On the one hand information systems on the web need to publish information in the shortest possible time after it is available from information sources. On the other hand, the most relevant dimensions are, from one side, accuracy, currency, and completeness, relevant also in the monolithic setting, from the other side a new dimension arises, namely trustworthiness of the sources. With the advent of internet-based systems, web information systems, and peer to peer information systems, sources of data increase dramatically, and provenance on available data is difficult to evaluate in the majority of cases. This is a radical change with respect to old centralized systems (still widespread in some organizations, such as banks), where data sources and data flows are accurately controlled and monitored. So, evaluating trustworthiness becomes crucial in web information systems. Several papers deal with this issue, see e.g. [20] and [21]. These two requirements are in many aspects contradictory: accurate design of data structures, and in the case of web sites, of good navigational paths between pages, and certification of data to verify its correctness are costly and lengthy activities, while publication of data on web sites requires stringent times. Web information systems present three peculiar aspects with respect to traditional information sources: first, a web site is a continuously evolving source of information, and it is not linked to a fixed release time of information; second, the process of producing information changes, additional information can be produced in different phases, and corrections to previously published information are possible. Such features lead to a different type of information with respect to traditional media.

Interpretability is a dimension whose importance has increased dramatically while moving from monolithic systems to cooperative and internet based systems. More and more data and information is exchanged within flows in which the meaning of data is not explicitly represented.

Accessibility is also a major quality issue in web systems to guarantee understandability and feasibility when navigating and querying the web.

5.3 Life Sciences

Life sciences data and, specifically, biological data are characterized by a diversity of data types, very large volumes, and highly variable quality: data are available through vastly disparate sources and disconnected repositories, and their quality is difficult to assess and often unacceptable for the required usage. Biologists typically search several sources looking for good quality data, for instance in order to perform reliable experiments. However, the effort to actually assess the quality level of such

data is entirely in the biologists' hands; they have to manually analyze disparate sources, trying to integrate and reconcile heterogeneous and contradictory data in order to identify the best ones.

As an example, a typical activity of a biologist is to analyze a set of genes, with the objective of understanding their functions. The biologist may perform a web search on a site that is known to contain gene data for the particular organism under consideration. Once the data is obtained, the biologist must assess its reliability; in order to do so, he may perform a new web search in order to check if other sites provide the same gene information. It may happen that different sites provide conflicting results. The biologist has also to check that the provided results are up to date, i.e., if a gene is "unknown" in the queried sites, or no recent publication on that gene is available, e.g. through Pubmed (see [22]). In order to overcome such problems, Life Sciences and Biology need for a robust use of data quality techniques.

6 Methodologies for Data Quality

Two types of recovery actions are proposed in case of insufficient data quality: reactive recovery actions and proactive recovery actions. Reactive recovery actions are performed contextually with the execution services of the information system and allow the recovery of running services. Proactive recovery actions are mostly based on data mining techniques and can only be executed in an off-line mode; proactive recovery actions are complex and require the support of an environment that is able to execute services, detect runtime faults, and perform recovery actions without compromising the running instances of the monitored Web services. A long term approach to improvement is adopted, where recovery actions have the goal to improve services and applications in order to avoid future failures. Process-oriented methods allow the identification of the causes of data errors and their permanent elimination through an observation of the whole process in which data are involved. In the following, we discuss first a proactive approach, while in the subsequent section reactive approaches are illustrated.

6.1 Quality as Manufacturing of Information Products: IP-UML

As mentioned in Section 2, an information product approach to data quality can be the basis for proactive process-based methods for data quality improvement. The IP-UML approach [26] proposes to define a UML profile for data quality within a methodology that is based on the IP-MAP framework, with some original additional proposals: (i) the artifacts to produce during the improvement process are defined using UML elements defined in the profile for data quality; (ii) it uses IP-MAPs not only in order to assess quality and to detect the need for improvement actions, but also as a diagrammatic way to design improvement actions. The starting concepts are the ones defined in the IP-MAP framework. The Information Quality profile consists of three different models, namely: (i) the Data Analysis Model, identifying which data are of interest, (ii) the Quality Analysis Model, identifying the quality requirements for each data type, and

(iii) the Quality Design Model, modeling data and processes together in order to model process-based improvement actions. As for techniques to produce artifacts, design patterns for quality improvement are proposed. Quality improvement is a complex activity that typically requires investments in terms of money and of people skills. The reuse of solutions and experiences can be very useful in supporting quality improvement, and can reduce time and costs considerably. For instance, a variety of techniques for data improvement are proposed in the literature and can be adopted as the basis for design patterns. The most straightforward solution suggests the adoption of data-oriented inspection and rework techniques, such as data bashing or data cleaning [10]. These techniques focus on data values and can solve problems related to data accuracy and data consistency quality dimensions [10]. A fundamental limitation of these techniques is that they do not prevent future errors. They are considered appropriate only when data are not modified frequently [10]. On the other hand, a more frequent use of data bashing and data cleaning algorithms involves high costs that can be difficult to justify.

A Process to Improve data quality is proposed, consisting of three distinct phases: Data Analysis, Quality Analysis and Quality Improvement Design. The Data and Quality Analysis phases are inspired by the IP-MAP framework, and are simply explained by the specific UML artifacts that should be produced in each phase in conformance with the Information Quality profile. The Quality Improvement Design phase consists of two distinct sub-phases, namely: the Quality Verification phase and the Quality Improvement phase. The former is inspired by the IP-MAP framework; while, the latter has been introduced with the specific aim of using IP-MAPs to explicitly model improvement processes. The task of data quality improvement needs to be adequately supported by software tools. Such tools can be classified as follows: Modeling tools, supporting non-automatic activities, such as modeling and design. Measurement tools, implementing quality measurement techniques. Ad-hoc tools.

6.2 Self Healing Systems

Self healing systems are proposed in [27] to provide repair actions when faults in the information system are identified. Among the possible faults we consider in the present paper data level faults as a type of application-level faults. Internal data level faults include data quality faults that are related to the data manipulated during the execution of a specific service. It is important to evaluate the quality of information flow along a specific service since failures can be caused by incorrect or missing information.

The healing approaches proposed are based on repair actions defined for the different types of faults, leveraging on adaptivity mechanisms at the web-service and application levels. In particular, we focus on healing mechanisms based both on service selection and substitution and on the addition of new services in composed processes to support self-healing functionalities. For each diagnosed fault, one or more recovery actions are executed and, with respect to the way in which these actions are performed.

Run time recovery actions in data quality require the identification of the causes of data errors and their permanent elimination through an observation of the whole process in which data are involved. Data tracking methods are required in order to determine the exact stage or steps in information process where the causes of data quality decreasing occur[6].

Tools are being proposed for data quality monitoring, analysis, and improvement. For instance, the Poliqua Quality Factory [28] allows performing both (i) *on line evaluation*: the property quality values and aggregate quality values are computed using the corresponding algorithms, as illustrated above, and (ii) *off-line evaluation*: aggregate quality values are precomputed and stored in a quality repository. Moreover, decision rules for applying the above mentioned evaluation techniques can be designed, focusing in particular on evaluation of quality metadata and improvement actions.

7 Concluding Remarks

Evolving characteristics of information systems pose new requirements with respect to data quality. Poor data quality has implications in terms of additional costs, both for data quality improvement actions and for indirect effects of poor data quality such as, for instance, reduced trustworthiness of providers of information system services. Additional research is needed to evaluate which are the most relevant quality dimensions according to the different characteristics of a specific information system. Moreover, cost implications need further investigation and techniques for planning, monitoring, and controlling data quality, both with reactive and proactive approaches are still being investigated.

References

1. T. Redman. Data Quality for the Information Age. Artech House (1996)
2. T. Redman. Data Quality The field guide. The Digital Press (2001)
3. 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)
4. Y. Wand and R. Wang, R. Anchoring data quality dimensions in ontological foundations. Communications of the ACM 39(11) (1996)
5. R. Wang and D. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4) (1996)
6. M. Bovee, R. Srivastava, and Mak, B. A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. In: Proceedings of the 6th International Conference on Information Quality. (Boston, MA, 2001)
7. M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, P., eds. Fundamentals of Data Warehouses. Springer Verlag (1995)
8. M. Scannapieco, P. Missier, and C. Batini, C. Data Quality at a Glance. Datenbank Spectrum 14 (2005) 6–14

9. S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers (2000)
10. P. Buneman. Semistructured data. In: *Proceedings of PODS '97*, Tucson, Arizona (1997)
11. D. Calvanese, D.D. Giacomo, and M. Lenzerini. Modeling and querying semi-structured data. *Networking and Information Systems Journal* 2(2) (1999) 253–273
12. B. Pernici and M. Scannapieco. Data quality in web information systems. *J. Data Semantics* 1 (2003) 48–68
13. G. Shankaranarayan, R. Wang, and M. Ziad. Modeling the Manufacture of an Information Product with IP-MAP. In: *Proceedings of the 6th International Conference on Information Quality (ICIQ 2000)*, (Boston, MA, 2000)
14. R. Wang, Y. Lee, L. Pipino, and D. Strong. Manage your information as a product. *Sloan Management Review* 39(4) (1998) 95–105
15. P. Missier, G. Lack, V. Verykios, F. Grillo, T. Lorusso, P. Angeletti. Improving data quality in practice: a case study in the italian public administration. *Parallel and distributed Databases* 13(2) (2003) 135–160
16. M.T. Ozsu, P. Valduriez. *Principles of Distributed Database Systems*, Second edition. Prentice Hall (2000)
17. G. De Michelis, E. Dubois, M. Jarke, F. Matthes, J. Mylopoulos, M. Papazoglou, K. Pohl, J. Schmidt, C. Woo, E. Yu. *Cooperative Information Systems: A Manifesto*. In Papazoglou, M., Schlageter, G., eds.: *Cooperative Information Systems: Trends & Directions*. Accademic-Press (1997)
18. P. Buneman, S. Khanna, W. Tan. Why and where: A characterization of data provenance. In: *Proceedings of the International Conference on Database Theory (ICDT)*. (London, United Kingdom, 2001) 316–330
19. B. Pernici. ed. *Mobile Information Systems*. Springer (2006)
20. L. De Santis, M. Scannapieco, T. Catarci. Trusting Data Quality in Cooperative Information Systems. In: *Proceedings of 11th International Conference on Cooperative Information Systems (CoopIS 2003)*. (Catania, Italy, 2003)
21. A.F. Cardenas, R. Pon. Data quality inference. In: *IQIS '05: Proceedings of the 2nd International Workshop on Information Quality in Information Systems*, New York, NY, USA, ACM Press (2005)
22. U.S. National Institutes of Health (NIH). (<http://www.pubmedcentral.nih.gov/>)
23. H. Krawczyk, B. Wiszniewski. Digital document life cycle development. In: *ISICT 2003: Proceedings of the 1st International Symposium on Information and Communication Technologies*, Trinity College Dublin (2003) 255–260
24. H. Krawczyk and B. Wiszniewski. Visual GQM approach to quality-driven development of electronic documents. In: *Proceedings of the Second International Workshop on Web Document Analysis (WDA2003)* (2003)
25. International monetary fund. (<http://dsbb.imf.org/>)
26. M. Scannapieco, B. Pernici, E.M. Pierce. IP-UML: Towards a methodology for quality improvement based on the IP-MAP framework. In: *IQ*. (2002) 279–291
27. D. Ardagna, C. Cappiello, M. Fugini, P. Plebani, B. Pernici. Faults and recovery actions for self-healing web services (2006)
28. C. Cappiello, C. Francalanci, and B. Pernici. A self-monitoring system to satisfy data quality requirements. In: *OTM Conferences (2)*. (2005) 1535–1552