

Introduction to Data Quality

A Web search of the terms “data quality” through the search engine Google, returns about three millions of pages, an indicator that data quality issues are real and increasingly important (often, in the following, the term data quality will be shortened to the acronym DQ). The goal of this chapter is to introduce the relevant perspectives that make data quality an issue worth being investigated and understood. We first introduce the notion of data quality (Section 1.1), highlighting its relevance in real life and some of the main related initiatives in the public and private domains. Then, in Section 1.2, we show by means of several examples the multidimensional nature of data quality. Sections 1.3 and 1.4 analyze the different types of data, and the different types of information systems for which DQ can be investigated. In Section 1.5, we address the main research issues in DQ, application domains and related research areas. The research issues (Section 1.5.1) concern dimensions, models, techniques, methodologies, and tools; together, they provide the agenda for the rest of the book. Application domains are large sets, since data and information are fundamental ingredients of all the activities of people and organizations. We focus (Section 1.5.2) on three of the most relevant application domains, e-Government, Life Sciences, and the World Wide Web, highlighting the role that DQ plays in each of them. Research areas related to DQ will be examined in Section 1.5.3.

1.1 Why Data Quality is Relevant

The consequences of poor quality of data are often experienced in everyday life, but, often, without making the necessary connections to their causes. For example, the late or mistaken delivery of a letter is often blamed on a malfunctional postal service, although a closer look often reveals data-related causes, typically an error in the address, originating in the address database. Similarly, the duplicate delivery of automatically generated mail is often indicative of a database record duplication error.

Data quality has serious consequences, of far-reaching significance, for the efficiency and effectiveness of organizations and businesses. As already mentioned in the preface, the report on data quality of the Data Warehousing Institute (see [52]) estimates that data quality problems cost U.S. businesses more than 600 billion dollars a year. The findings of the report were based on interviews with industry experts, leading edge customers, and survey data from 647 respondents. In the following, we list further examples of the importance of data quality in organizational processes.

- *Customer matching.* Information systems of public and private organizations can be seen as the result of a set of scarcely controlled and independent activities producing several databases very often characterized by overlapping information. In private organizations, such as marketing firms or banks, it is not surprising to have several (sometimes dozens!) of customers registries, updated with different organizational procedures, resulting in inconsistent, duplicate information. As an example, it is very complex for banks to provide clients with a unique list of all their accounts and funds.
- *Corporate house-holding.* Many organizations establish separate relationships with single members of households, or, more generally, related groups of people; either way, they like, for marketing purposes, to reconstruct the household relationships in order to carry on more effective marketing strategies. This problem is even more complex than the previous one, since in that case the data to match concerned the same person, in this case it concerns groups of persons corresponding to the same household. For a detailed discussion on the relationship between corporate house holding information and various business application areas, see [200].
- *Organization fusion.* When different organizations or different units of an organization merge, it is necessary to integrate their legacy information systems. Such integration requires compatibility and interoperability at any layer of the information system, with the database level required to ensure both physical and semantic interoperability.

The examples above are indicative of the growing need to integrate information across completely different data sources, an activity in which poor quality hampers integration efforts. Awareness of the importance of improving the quality of data is increasing in many contexts. In the following, we summarize some of the major initiatives in both the private and public domains.

Private Initiatives

In the private sector, on the one hand, application providers and system integrators, and, on the other hand, direct users are experiencing the role of DQ in their own business processes.

With regard to application providers and systems integrators, IBM's recent (2005) acquisition of Ascential Software, a leading provider of data integration

tools, highlights the critical role data and information stewardship plays in the enterprise. The 2005 Ascential report [208] on data integration provides a survey that indicates data quality and security issues as the leading inhibitors (55 % of respondents in a multi-response survey) to successful data integration projects. The respondents also emphasize that data quality is more than just a technological issue. It requires senior management to treat data as a corporate asset and to realize that the value of this asset depends on its quality.

In the last few years, SAP [84] has set up a project for testing in the area of DQ and to build an internal methodology, with important savings (documented in [84]) in several internal business processes.

The awareness of the relevance of data quality issues has led Oracle (see [151]) to recently enhance its suite of products and services to support an architecture that optimizes data quality, providing a framework for the systematic analysis of data, with the goals of increasing the value of data, easing the burden of data migration, and decreasing the risks inherent in data integration.

With regard to users, Basel2 is an international initiative in the financial domain that requires financial services companies to have a risk sensitive framework for the assessment of regulatory capital. The planned implementation date for Basel2 is December 2006, with parallel operation from January 2006. The regulatory requirements of Basel2 are demanding improvements in data quality. For example, the Draft Supervisory Guidance on Internal Ratings-Based Systems for Corporate Credit states (see [19]): “institutions using the Internal Ratings-Based approach for regulatory capital purposes will need advanced data management practices to produce credible and reliable risk estimates”; and “data retained by the bank will be essential for regulatory risk-based capital calculations and public reporting. These uses underscore the need for a well defined data maintenance framework and strong controls over data integrity.”

Public Initiatives

In the public sector a number of initiatives address data quality issues at international, European, and national levels. We focus in the rest of the section on two of the main initiatives, the Data Quality Act in the US and the European directive on reuse of public data.

In 2001 the President of the US signed into law important new Data Quality legislation, concerning “Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies,” in short the Data Quality Act. The Office of Management and Budget (OMB) issued guidelines referred for policies and procedures on data quality issues (see [149]). Obligations mentioned in the guidelines concern agencies, which are to report periodically to the OMB regarding the number and nature of data quality complaints received, and how such complaints were handled. OMB must also include a mechanism through which

the public can petition agencies to correct information that does not meet the OMB standard. In the OMB guidelines data quality is defined as an encompassing term comprising utility, objectivity, and integrity. Objectivity is a measure to determine whether the disseminated information is accurate, reliable, and unbiased, and whether that information is presented in an accurate, clear, complete, and unbiased manner. Utility refers to the usefulness of the information for its anticipated purpose, by its intended audience. OMB is committed to disseminating reliable and useful information. Integrity refers to the security of information, namely protection of the information from unauthorized, unanticipated, or unintentional modification, to prevent it from being compromised by corruption or falsification. Specific risk-based, cost-effective policies are defined for assuring integrity.

The European directive 2003/98/CE on the reuse of public data (see [71]) highlights the importance of reusing the vast data assets owned by public agencies. The public sector collects, produces, and disseminates a wide range of information in many areas of activity, such as social, economic, geographical, meteorological, business, and educational information. Making public all generally available documents held by the public sector, concerning not only the political process but also the legal and administrative processes, is considered a fundamental instrument for extending the right to information, which is a basic principle of democracy. Aspects of data quality addressed by such a directive are the accessibility of public data and availability in a format which is not dependent on the use of specific software. At the same time, a related and necessary step for public data reuse is to guarantee its quality in terms of accuracy and currency, through data cleaning campaigns. This makes it attractive to new potential users and customers.

1.2 Introduction to the Concept of Data Quality

From a research perspective, data quality has been addressed in different areas, including statistics, management, and computer science. Statisticians were the first to investigate some of the problems related to data quality, by proposing a mathematical theory for considering duplicates in statistical data sets, in the late 1960's. They were followed by researchers in management, who at the beginning of the 1980's focused on how to control data manufacturing systems in order to detect and eliminate data quality problems. Only at the beginning of the 1990's computer scientists begin considering the problem of defining, measuring, and improving the quality of electronic data stored in databases, data warehouses, and legacy systems.

When people think about data quality, they often reduce data quality just to accuracy. For example, let us consider the surname "Batini"; when this is spelled during a telephone call, several misspellings are reported by the other side, such as "Vatini," "Battini," "Barini," "Basini," all inaccurate versions of the original last name. Indeed, data are normally considered to be of poor

quality if typos are present or wrong values are associated with a concept instance, such as an erroneous birth date or age associated with a person. However, data quality is more than simply data accuracy. Other significant dimensions such as completeness, consistency, and currency are necessary in order to fully characterize the quality of data. In Figure 1.1 we provide some examples of these dimensions, which are described in more detail among others in Chapter 2. The relation in the figure describes movies, with title, director, year of production, number of remakes, and year of the last remake.

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

Fig. 1.1. A relation `Movies` with data quality problems

In the figure, the cells with data quality problems are shaded. At first, only the cell corresponding to the title of movie 3 seems to be affected by a data quality problem. In fact, there is a misspelling in the title, where `Rman` stands for `Roman`, thus causing an accuracy problem. Nevertheless, another accuracy problem is related to the exchange of the director between movies 1 and 2; Weir is actually the director of movie 2 and Curtiz the director of movie 1. Other data quality problems are a missing value for the director of movie 4, causing a completeness problem, and a 0 value for the number of remakes of movie 4, causing a currency problem because a remake of the movie has actually been proposed. Finally, there are two consistency problems: first, for movie 1, the value of `LastRemakeYear` cannot be lower than `Year`; second, for movie 4 the value of `LastRemakeYear` cannot be different from null, because the value of `#Remakes` is 0.

The above examples of dimensions concern the *quality of data* represented in the relation. Besides data, a large part of the design methodologies for the relational model addresses properties that concern the *quality of the schema*; for example, several normal forms have been proposed with the aim of capturing the concept of good relational schema, free of anomalies and redundancies. For instance, the relational schema of Figure 1.1 is in the Boyce Codd normal form, since all attributes that do not belong to a superkey are functionally dependent on the superkeys (`Id` and `Title`). Other data quality and schema quality dimensions will be discussed in Chapter 2. The above examples and considerations show that:

- Data quality is a multifaceted concept, as in whose definition different dimensions concur.
- The quality dimensions, e.g., accuracy, can be easily detected in some cases (e.g., misspellings) but are more difficult to detect in other cases (e.g., where admissible but not correct values are provided).
- A simple example of a completeness error has been shown, but as with accuracy, completeness can also be very difficult to evaluate (e.g., if a tuple representing a movie is entirely missing from the relation `Movie`).
- Consistency detection does not always localize the errors (e.g., for movie 1, the value or the `LastRemakeYear` attribute is wrong).

The above example concerned a relational table of a single database. Problems change significantly when other *types of data* are involved, and more complex *types of information systems* than a single database are considered. We now address these two aspects.

1.3 Data Quality and Types of Data

Data represent real world objects, in a format that can be stored, retrieved, and elaborated by a software procedure, and communicated through a network. The process of representing the real world by means of data can be applied to a large number of phenomena, such as measurements, events, characteristics of people, the environment, sounds, and smells. Data are extremely versatile in such representation. Besides data, other types of *information* are used in real-life and business processes, such as paper-based information, and information conveyed by the voice. We will not deal with all these types of information, and we concentrate on data.

Since researchers in the area of data quality must deal with a wide spectrum of possible data representations, they have proposed several classifications for data. First, several authors distinguish, implicitly or explicitly, three types of data:

1. *Structured*, when each data element has an associated fixed structure. Relational tables are the most popular type of structured data.
2. *Semistructured*, when data has a structure which has some degree of flexibility. Semistructured data are also “schemaless” or “self-describing” (see [1], [35], and [40]). XML is the markup language commonly used to represent semistructured data. Some common characteristics are (i) data can contain fields not known at design time; for instance, an XML file does not have an associated XML schema file; (ii) the same kind of data may be represented in multiple ways; for example, a date might be represented by one field or by multiple fields, even within a single set of data; and (iii) among fields known at design time, many fields will not have values.
3. *Unstructured*, when data are expressed in natural language and no specific structure or domain types are defined.

It is intuitive that 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 to unstructured data.

A second point of view sees data as a product. This point of view is adopted, for example, in the IP-MAP model (see [177]), an extension of the Information Manufacturing Product model [201], which will be discussed in detail in Section 3.4; 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 smaller data units. They are used to construct information and component data items that are semi-processed information;
- while the raw data items may be stored for long periods of time, the *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 products;
- *information products*, which are the result of a manufacturing activity performed on data.

Looking at data as a product, as discussed in Chapters 3 and 7, methodologies and procedures used over a long period, with suitable changes having been made to them, can be applied to data for quality assurance in manufacturing processes.

The third classification, proposed in [133], addresses a typical distinction made in information systems between elementary data and aggregated data. *Elementary data* are 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 an attribute **Sex** changes dramatically if we input **M** (male) instead of **F** (female); if the age of a single person is wrongly recorded as 25 instead of 35, the accuracy of the average age of a population of millions of inhabitants is minimally affected.

Dasu and Johnson in [50] investigate new types of data that emerge from the diffusion of networks and Internet, and observe that the definition of data itself has changed dramatically to include “any kind of information that is analyzed systematically.” They distinguish several new types of data, among them are relevant in this book:

- *federated data*, which come from different heterogeneous sources, and, consequently, require disparate data sources to be combined with approximate matches;

- *web data*, that are “scraped” from the Web and, although characterized by unconventional formats and low control on data, more often constitute the primary source of information for several activities.

Previous classifications were not interested in the time dimension of data, investigated in [30]. According to its change frequency, we can classify source data into three categories:

- *stable* data is data that is unlikely to change. Examples are scientific publications; although new publications can be added to the source, older publications remain unchanged;
- *long-term-changing data* is data that has very low change frequency. Examples are addresses, currencies, and hotel price lists. The concept of low frequency is domain dependent; in an e-trade application, if the value of a stock quote is tracked once an hour, it is considered to be a low frequency change, while a shop that changes its goods weekly has a high-frequency change for clients;
- *frequently-changing data* is data that has intensive change, such as real-time traffic information, temperature sensor measures, and sales quantities. The changes can occur with a defined frequency or they can be random.

For this classification, the procedures for establishing the time dimension qualities of the three types of data, i.e., stable, long-term-changing, and frequently-changing data, are increasingly more complex.

Among the different types of data resulting from the above classification, we are mainly interested in focusing our attention on *structured* and *semistructured elementary data*, and on *information products*. Such types of data have been deeply investigated in the literature, and, to a certain extent, consolidated techniques and methodologies have been conceived. This does not mean that we will exclude other types of data from our analysis: dimensions for time-dependent data will be introduced and discussed in Chapter 2, and web data will be considered in Chapter 9, dedicated to open problems.

As a terminological note, when we give generic examples of structured data, we use the term *tuple* to indicate a set of *fields* or *cell values*, corresponding usually to different *definition domains* or *domains*, describing properties or *attributes* of a specific real world object; we use interchangeably the terms *relational table* or *table* or *relation* to indicate a set of tuples. As a consequence, *tuple* can be used in place of *record* and *table/relation* can be used in place of *structured file*. When we refer to generic data, we use the term *record* to indicate a set of fields, and we use interchangeably the terms *file* or *data set* to indicate a set of tuples.

1.4 Data Quality and Types of Information Systems

Data are collected, stored, elaborated, retrieved, and exchanged in *information systems* used in organizations to provide services to business processes. Different criteria can be adopted for classifying the different types of information systems, and their corresponding architectures; they are usually related to the overall organizational model adopted by 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 [153] for distributed databases. Three different criteria are proposed: distribution, heterogeneity, and autonomy.

Distribution deals with the possibility of distributing the data and the applications over a network of computers. For simplicity, we adopt a $\langle \text{yes}, \text{no} \rangle$ domain for distribution. *Heterogeneity* considers all types of semantic and technological diversities among systems used in modeling and physically representing data, such as database management systems, programming languages, operating systems, middleware, markup languages. For heterogeneity we also adopt a simple $\langle \text{yes}, \text{no} \rangle$ domain. *Autonomy* has to do with the degree of hierarchy and rules of coordination, establishing rights and duties, defined in 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. In this case we adopt a three-value $\langle \text{no}, \text{semi}, \text{totally} \rangle$ domain.

The three classifications are represented together in the classification space of Figure 1.2. Among all possible combinations, five main types of information systems are highlighted in the figure: Monolithic, Distributed, Data Warehouses, Cooperative, and Peer-to-Peer.

- 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 the homogeneity of solutions and centralization of management. In monolithic systems, data flows have a common format, and data quality control is facilitated by the homogeneity and centralization of procedures and management rules.
- A *data warehouse* (DW) is a centralized set of data collected from different sources, designed to support management decision making. The most critical problem in DW design concerns the cleaning and integration of the different data sources that are loaded into 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 the distribution of resources and applications across a network of geographically distributed systems. The network

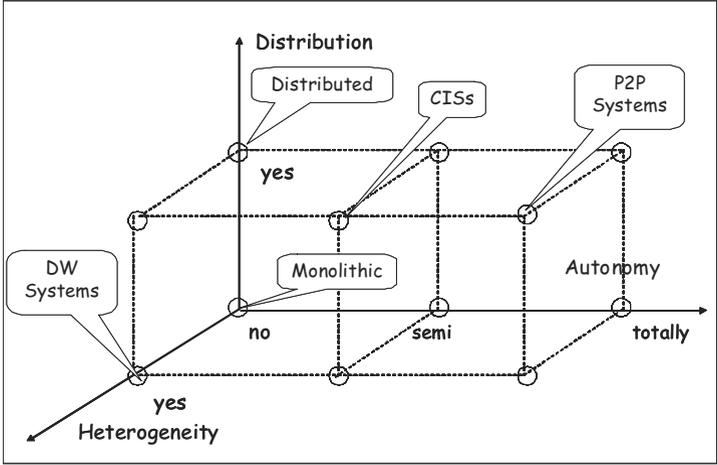


Fig. 1.2. Types of information systems

can be organized in terms of several tiers, each made of one or more computational nodes. Presentation, application logic, and data management are distributed across tiers. Usually, the different tiers and nodes have a limited degree of autonomy, data design is usually performed centrally, but to a certain extent some degree of heterogeneity can occur, due to the impossibility of establishing unified procedures. Problems of data management are more complex than in monolithic systems, due to the reduced level of centralization. Heterogeneities and autonomy usually increase with the number of tiers and nodes.

- A *cooperative information system* (CIS) can be defined as a large-scale information system that interconnects various systems of different and autonomous organizations, while sharing common objectives. According to [58], the manifesto of cooperative information systems, “an information system is cooperative if it shares goals with other agents in its environment, such as other information systems, human agents, and the organization itself, and contributes positively toward the fulfillment of these common goals.” The relationship between cooperative information systems and DQ is double-faced: on the one hand it is possible to profit the cooperation between agents in order to choose the best quality sources, and thus improve the quality of circulating data. On the other hand, data flows are less controlled than in monolithic systems, and the quality of data, when not controlled, may rapidly decrease in time. Integration of data sources is also a relevant issue in CISs, especially when partners decide to substitute a group of databases, that have been independently developed, with an integrated in-house database. In *virtual data integration* a unique virtual integrated schema is built to provide unified access. This case is affected by

data quality problems, because inconsistencies in data stored at different sites make it difficult to provide integrated information.

- In a *peer-to-peer information system* (usually abbreviated P2P), the traditional distinction between clients and servers typical of distributed systems is disappearing. A P2P system can be characterized by a number of properties: peers are highly autonomous and highly heterogeneous, they have no obligation for the quality of their services and data, no central coordination and no central database exist, no peer has a global view of the system, global behavior emerges from local interactions. 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. It is also costly for a single agent to evaluate the reputation of other partners.

In the rest of the book, we will examine DQ issues mainly conceived for monolithic, distributed, data warehouses, and cooperative information systems, while issues for P2P systems will be discussed in Chapter 9 on open problems.

1.5 Main Research Issues and Application Domains in Data Quality

Due to the relevance of data quality, its nature, and the variety of data types and information systems, achieving data quality is a complex, multidisciplinary area of investigation. It involves several research topics and real-life application areas. Figure 1.3 shows the main ones.

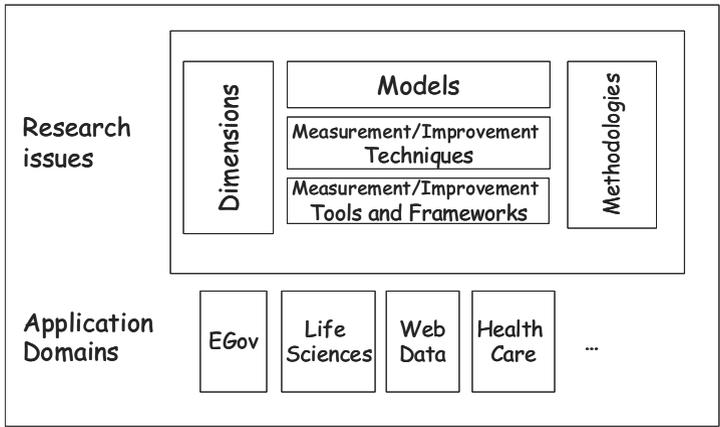


Fig. 1.3. Main issues in data quality

Research issues concern models, techniques, and tools, and two “vertical” areas, that cross the first three, i.e. dimensions and methodologies. We will

discuss them in Section 1.5.1. Three of the application domains mentioned in Figure 1.3, namely e-Government, Life Sciences, and the World Wide Web, in which DQ is particularly relevant, are discussed in Section 1.5.2.

Research issues in DQ originate from research paradigms initially developed in other areas of research. The relationship between data quality and these related research areas will be discussed in Section 1.5.3.

1.5.1 Research Issues in Data Quality

Choosing *dimensions* to measure the level of quality of data is the starting point of any DQ-related activity. Though measuring the quality of ICT technologies, artifacts, processes, and services is not a new issue in research, for many years several standardization institutions have been operating (e.g. ISO, see [97]) in order to establish mature concepts in the areas of quality characteristics, measurable indicators, and reliable measurement procedures. Dimensions are discussed in Chapter 2. Dimensions are applied with different roles in models, techniques, tools, and frameworks.

Models are used in databases to represent data and data schemas. They are also used in information systems to represent business processes of the organization; these models have to be enriched in order to represent dimensions and other issues related to DQ. Models are investigated in Chapter 3.

Techniques correspond to algorithms, heuristics, knowledge-based procedures, and learning processes that provide a solution to a specific DQ problem or, as we say, to a *data quality activity*, as defined in Chapter 4. Examples of DQ activities are identifying if two records of different databases represent the same object of the real world or not; or finding the most reliable source for some specific data. DQ activities are defined in Chapter 4 and techniques are discussed in Chapters 4, 5, and 6.

Methodologies provide guidelines to choose, starting from available techniques and tools, the most effective DQ measurement and improvement process (and hopefully, most economical for comparable results) within a specific information system. Methodologies are investigated in Chapter 7.

Methodologies and techniques, in order to be effective, need the support of *tools*, i.e., automatized procedures, provided with an interface, that relieve the user of the manual execution of some techniques. When a set of coordinated tools is integrated to provide a set of DQ services, we will use the term *framework*. Tools and frameworks are discussed in Chapter 8.

1.5.2 Application Domains in Data Quality

In this section, we analyze three distinct application domains of DQ. Their importance has been growing over the last few years, because of their relevance in daily lives of people and organizations: e-Government, Life Sciences, the World Wide Web.

e-Government

The main goal of all e-Government projects is the improvement of the relationship between the government, agencies, and citizens, as well as between agencies and businesses, through the use of information and communication technologies. This ambitious goal is articulated in different objectives:

1. the complete automation of those government administrative processes that deliver services to citizens and businesses, and that involve the exchange of data between government agencies;
2. the creation of an architecture that, by connecting the different agencies, enables them to fulfill their administrative processes without any additional burden to the users that benefit from them; and
3. the creation of portals that simplify access to services by authorized users.

e-Government projects must face the problem that similar information about one citizen or business is likely to appear in multiple databases. Each database is autonomously managed by the different agencies that historically has never been able to share data about citizens and businesses.

The problem is worsened by the many errors usually present in the databases, for many reasons. First, due to the nature of the administrative flows, several citizens' data (e.g. addresses) are not updated for long periods of time. This happens because it is often impractical to obtain updates from subjects that maintain the official residence data. Also, errors may occur when personal data on individuals is stored. Some of these errors are not corrected and a potentially large fraction of them is not detected. Furthermore, data provided by distinct sources differ in format, following local conventions, that can change in time and result in multiple versions. Finally, many of the records currently in the database were entered over years using legacy processes that included one or more manual data entry steps.

A direct consequence of this combination of redundancy and errors in data is frequent mismatches between different records that refer to the same citizen or business. One major outcome of having multiple disconnected views for the same information is that citizens and businesses experience consistent service degradation during their interaction with the agencies. Furthermore, misalignment brings about additional costs. First, agencies must make an investment to reconcile records using clerical review, e.g., to manually trace citizens and businesses that cannot be correctly and unequivocally identified. Secondly, because most investigation techniques, e.g., tax fraud prevention techniques, rely on cross-referencing records of different agencies, misalignment results in undetected tax fraud and reduced revenues.

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. Their quality is difficult to assess and often unacceptable for the required usage. Biologists typically search several sources, for good quality data, for instance, in order to perform reliable in-silico experiments. However, the effort to actually assess the quality level of such data is entirely in the hands of the biologists; they have to manually analyze disparate sources, trying to integrate and reconcile heterogeneous and contradictory data in order to identify the best information. Let us consider, as an example, a gene analysis scenario. Figure 1.4 shows an example of a simple data analysis pipeline. As the result of a micro-array experiment, a biologist wants to analyze a set of genes, with the objective of understanding their functions.

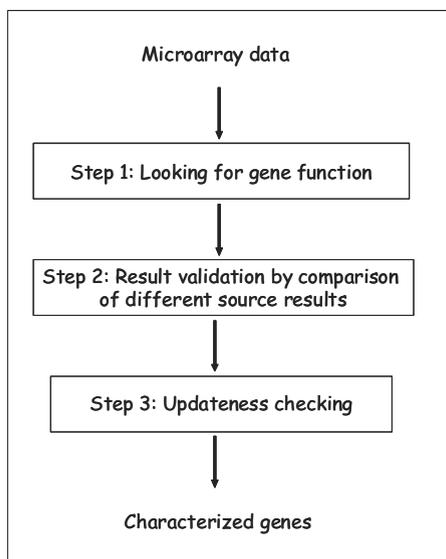


Fig. 1.4. Example of biological data analysis process

In Step 1, the biologist performs 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. Therefore, in Step 2 the biologist performs 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. Then (Step 3) the biologist also has 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 [192]). The described scenario has many weaknesses:

1. the biologist must perform a time-consuming manual search for all the sources that may provide the function of the interested gene. This process

is also dependent on the user having personal knowledge about which sites must be queried;

2. the biologist has no way of assessing the trustworthiness of a result;
3. in Step 2, the biologist has no way of evaluating the quality of the results provided by different sites.
4. in Step 3, a new web search must be performed which again can be very time consuming.

In order to overcome such weaknesses, life sciences and biology need robust data quality techniques.

World Wide Web

Web information systems are characterized by the presentation of a large amount of data to a wide audience, the quality of which can be very heterogeneous. There are several reasons for this variety. First, every organization and individual can create a Web site and load every kind of information without any control on its quality, and sometimes with a malicious intent. A second reason lies in 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, information has to be checked with regard to its accuracy, currency, and trustworthiness of its sources. 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. However, the publication of data on Web sites is subject to time constraints.

Web information systems present two further aspects in connection to data quality that differentiate them from 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, in the process of changing information, additional information can be produced in different phases, and corrections to previously published information are possible, creating, in such a way, further needs for quality checks. Such features lead to a different type of information than with traditional media.

As a final argument, in Web information systems it is practically impossible to individuate a subject, usually called *data owner*, 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.

All previously discussed aspects make it difficult to certify the quality of data sources, and, for a user, to assess the reputation of other users and sources.

1.5.3 Research Areas Related to Data Quality

Data quality is fairly a new research area. Several other areas (see Figure 1.5) in computer science and other sciences have in the past treated related and overlapping problems; at the same time, such areas have developed in the last decades (in the case of statistics, in the last centuries) paradigms, models, and methodologies that have proved to be of major importance in grounding the data quality research area. We now discuss such research areas.

1. *Statistics* includes a set of methods that are used to collect, analyze, present, and interpret data. Statistics has developed in the last two centuries a wide spectrum of methods and models that allow one to express predictions and formulate decisions in all contexts where uncertain and imprecise information is available for the domain of interest. As discussed in [121], statistics and statistical methodology as the basis of data analysis are concerned with two basic types of problems: (i) summarizing, describing, and exploring data, (ii) using sampled data to infer the nature of the process that produced the data. Since low quality data are an inaccurate representation of the reality, a variety of statistical methods have been developed for measuring and improving the quality of data. We will discuss some statistical methods in Chapters 4 and 5.
2. *Knowledge representation* (see [144] and [54] for insightful introductions to the area) is the study of how knowledge about an application domain can be represented, and what kinds of reasoning can be done with that knowledge (this is called *knowledge reasoning*). Knowledge about an application domain may be represented procedurally in form of program code, or implicitly as patterns of activation in a neural network. Alternatively, the area of knowledge representation assumes an explicit and declarative representation, in terms of a *knowledge base*, consisting of logical formulas or rules expressed in a representation language. Providing a rich representation of the application domain, and being able to reason about it, is becoming an important leverage in many techniques for improving data quality; we will see some of these techniques in Chapters 5 and 8.
3. *Data mining* (see [92]) is an analytic process designed to explore usually large sets of data in search of consistent patterns and/or systematic relationships between attributes/variables. *Exploratory data mining* is defined in [50] as the preliminary process of discovering structure in a set of data using statistical summaries, visualization, and other means. In this context, achieving good data quality is an intrinsic objective of any data mining activity (see [46]), since otherwise the process of discovering patterns, relationships and structures is seriously deteriorated. From another perspective, data mining techniques may be used in a wide spectrum of activities for improving the quality of data; we will examine some of them in Chapter 4.

4. *Management information systems* (see [53]) are defined as systems that provide the information necessary to manage an organization effectively. Since data and knowledge are becoming relevant resources both in operational and decision business processes, and poor quality data result in poor quality processes, it is becoming increasingly important to supply management information systems with functionalities and services that allow one to control and improve the quality of the data resource.
5. *Data integration* (see [116]) has the goal of building and presenting a unified view of data owned by heterogeneous data sources in distributed, cooperative, and peer-to-peer information systems. Data integration will be considered in Chapter 4 as one of basic activities whose purpose is improving data quality, and will be discussed in detail in Chapter 6. While being an autonomous and well-grounded research area, data integration will be considered in this book as strictly related to data quality, regarding two main issues, providing query results on the basis of a quality characterization of data at sources, and identifying and solving conflicts on values referring to the same real-world objects.

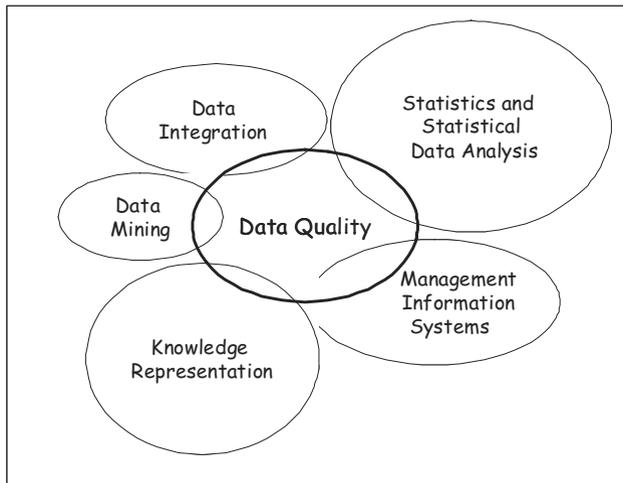


Fig. 1.5. Research areas related to data quality

1.6 Summary

In this chapter we have perceived that data quality is a multidisciplinary area. This is not surprising, since data, in a variety of formats and with a variety of media, are used in every real-life or business activity, and deeply influence

the quality of processes that use data. Many private and public organizations have perceived the impact of data quality on their assets and missions, and have consequently launched initiatives of large impact. At the same time, while in monolithic information systems data are processed within controlled activities, with the advent of networks and the Internet, data are created and exchanged with much more “turbulent” processes, and need more sophisticated management.

The issues discussed in this chapter introduce to the structure of the rest of the book: dimensions, models, techniques, methodologies, tools, and frameworks will be the main topics addressed. While data quality is a relatively new research area, other areas, such as statistical data analysis, have addressed in the past some aspects of the problems related to data quality; with statistical data analysis, also knowledge representation, data mining, management information systems, and data integration share some of the problems and issues characteristic of data quality, and, at the same time, provide paradigms and techniques that can be effectively used in data quality measurement and improvement activities.



<http://www.springer.com/978-3-540-33172-8>

Data Quality

Concepts, Methodologies and Techniques

Batini, C.; Scannapieco, M.

2006, XX, 262 p. 134 illus., Hardcover

ISBN: 978-3-540-33172-8