Chapter 1
Introduction to Information Quality

1.1 Introduction

The Search query "data quality" entered into Google returns about three million pages, and searching similarly for the term "information quality" (IQ) returns about one and a half million pages, both frequencies showing the increasing importance of data and information quality. The goal of this chapter is to show and discuss the perspectives that make data and information (D&I) quality an issue worth being investigated and understood. We first (Sect. 1.2) highlight the relevance of information quality in everyday life and some of the main related initiatives in the public and private domains. Then, in Sect. 1.3, we show the multidimensional nature of information quality by means of several examples. In Sect. 1.4, we discuss information quality and its relationship with several classifications proposed in the literature for information. Section 1.5 analyzes the different types of information systems for which IQ can be investigated. In Sect. 1.6, we address the main research issues in IQ, application domains in which it is investigated, and related research areas. The research issues (Sect. 1.6.1) concern dimensions, models, techniques, and methodologies; together, they provide the agenda for the rest of the book. The specific application domains for which D&I is relevant can be several, since data and information are fundamental ingredients of all the activities of people and organizations. We focus (Sect. 1.6.2) on three of them that we identified as particularly relevant—e-Government, life sciences, and the World Wide Web—highlighting the role that IQ plays in each of them. Research areas related to IQ will be examined in Sect. 1.6.3.

C. Batini, M. Scannapieco, *Data and Information Quality*, Data-Centric Systems and Applications, DOI 10.1007/978-3-319-24106-7_1

1.2 Why Information Quality Is Relevant

The consequences of poor quality of information can be experienced in everyday life but, often, without making explicit connections to their causes. Some examples are: the late or mistaken delivery of a letter is often blamed on a dysfunctional postal service, although a closer look often reveals data-related causes, typically an error in the address, which can be traced back to the originating database. Similarly, the duplicate delivery of automatically generated mails is often indicative of a database record duplication error.

Information quality seriously impacts on the efficiency and effectiveness of organizations and businesses. The report on information quality of the Data Warehousing Institute (see [162]) estimates that IQ problems cost US 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 IQ 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!) customer registries, updated by different organizational procedures, resulting in inconsistent, duplicate information. Some examples are: it is very difficult for banks to provide clients with a unique list of all their accounts and funds.
- *Corporate householding*. 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 out more effective marketing strategies. This problem is even more complex than the previous one, since in that case, the information to match concerned the same person, while in this case, it concerns groups of persons corresponding to the same household. For a detailed discussion on the relationship between corporate householding information and various business application areas, see [650].
- *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 information is increasing in many contexts. In the following, we summarize some of the major initiatives in both the private and public domains.

1.2.1 Private Initiatives

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

With regard to application providers and systems integrators, IBM's (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 [665] on data integration provides a survey that indicates information 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 information quality is more than just a technological issue. It requires senior management to treat information as a corporate asset and to realize that the value of this asset depends on its quality.

In a research by the Economist Intelligence Unit [617] in 2012 on managers' perception of the most problematic issues in the management of Big Data, "access the right data" (a kind of data quality dimension related to relevance of data) ranks first, while accuracy, heterogeneity reconciliation, and timeliness of data rank, respectively, second, third, and fourth.

The awareness of the relevance of information quality issues has led Oracle (see [481]) to enhance its suite of products and services to support an architecture that optimizes information quality, providing a framework for the systematic analysis of information, with the goals of increasing the value of information, easing the burden of data migration, and decreasing the risks inherent in data integration.

With regard to users, Basel2 and Basel3 are international initiatives in the financial domain that require financial services companies to have a risk-sensitive framework for the assessment of regulatory capital. Initially published in June 2004, Basel2 introduced regulatory requirements leading to demanding improvements in information quality. For example, the Draft Supervisory Guidance on Internal Ratings-Based Systems for Corporate Credit states (see [239]): "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."

Basel3, which was agreed upon by the members of the Basel Committee on Banking Supervision in 2010, proposes further policies for financial services companies (see [295]), by means of fine-tuning their risk-weighted asset models, improving their information quality in terms of improved rating coverage and transparency on underlying assets, and optimizing their asset segmentations. Based on the observation that Basel3 would require even more transparency into the risk positions taken by financial institutions, a novel approach to the evaluation and improvement of information quality in the financial sector is proposed in [172], and the Business Process Modeling Notation is used to represent bank business

processes, to identify where information elements enter the process, and to trace the various information outputs of processes.

1.2.2 Public Initiatives

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

In 2001, the President of the United States 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 information quality issues (see [478]). Obligations mentioned in the guidelines concern agencies, which are to report periodically to the OMB regarding the number and nature of information quality complaints received and how such complaints were handled. The 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, information 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. The 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 information (see [206] and its revision published in 2013 [207]) highlights the importance of reusing the vast information 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 information quality addressed by such a directive are the accessibility to public information 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 information reuse is to guarantee its quality in terms of accuracy and currency, through information cleaning campaigns. This makes it attractive to new potential users and customers.

1.3 Introduction to the Concept of Information Quality

Quality, in general, has been defined as the "totality of characteristics of a product that bear on its ability to satisfy stated or implied needs" [331], also called "fitness for (intended) use" [352], "conformance to requirements" [156], or "user satisfaction" [657].

From a research perspective, *information 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 information quality, by proposing a mathematical theory for considering duplicates in statistical datasets in the late 1960s. They were followed by researchers in management, who at the beginning of the 1980s focused on how to control information manufacturing systems in order to detect and eliminate information quality problems. Only at the beginning of the 1990s computer scientists began considering the problem of defining, measuring, and improving the quality of electronic information stored in databases, data warehouses, and legacy systems.

When people think about information quality, they often reduce 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," and "Basini," all inaccurate versions of the original. Indeed, information is 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, information quality is more than simply accuracy. Other significant dimensions such as completeness, consistency, and currency are necessary in order to fully characterize the quality of information. In Fig. 1.1, we provide some examples of these dimensions for structured data, which are described in more detail in Chap. 3. The relational table in the figure describes movies, with title, director, year of production, number of remakes, and year of the last remake.

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.

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 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 values* 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. Other data quality and schema quality dimensions will be discussed in Chap. 3.

The above examples and considerations show that:

- Data quality is a multifaceted concept, to the definition of which different dimensions concur
- Quality dimensions, such as 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 like happens 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 of the `LastRemakeYear` attribute is wrong)

The above example concerned a relational table of a database. Problems change significantly when other *types of information* different from structured relational data are involved. Let us focus on Fig. 1.2, showing an image representing a flower. Instinctively, the image on the right-hand side is considered of better quality in comparison to the image on the left-hand side. At the same time, it is not immediate to identify the dimension(s) that we are considering to come to such a conclusion.

In the preface, we provided a list of different information types. In the next section, we introduce several classifications of information relevant to quality issues, while issues related to information quality in several *types of information systems* will be considered in Sect. 1.5.

Fig. 1.2 Two images of the same flower with intuitively different image quality levels

1.4 Information Quality and Information Classifications

Structured data represent real-world objects, with a format and a model that can be stored, retrieved, and elaborated by database management systems (DBMSs). The process of representing the real world by means of structured data can be applied to a large number of phenomena, such as measurements, events, characteristics of people, environment, sounds, and smells. Structured data are extremely versatile in such representations but are limited by their intrinsic characteristics.

The types of information introduced in the preface strongly enhance the property of structured data to represent phenomena of the real world. Since researchers in the area of information quality must deal with a wide spectrum of possible information types, they have proposed several classifications for information. The first classification discussed in the preface refers to the perceptual vs. linguistic character of the information, and, among linguistic types of information, we can distinguish among structured, semistructured, and unstructured information.

A second point of view sees information as a product. This point of view is adopted, for example, in the IP-MAP model (see [563]), an extension of the Information Manufacturing Product model [648], which will be discussed in detail in Chap. 6; the IP-MAP model identifies a parallelism between the quality of information and the quality of products as managed by manufacturing companies. In this model, three different types of information 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 information as a product, methodologies and procedures used over a long period for quality assurance in manufacturing processes can be applied to information, with suitable changes having been made to them. This issue will be discussed in Chaps. 6 and 12.

The third classification, proposed in [446], 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, and 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).

Dasu and Johnson in [161] investigate new types of data that emerge from the diffusion of networks and the Web. They distinguish several new types of data; the following 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*, which are "scraped" from the Web and, although characterized by unconventional formats and low control on information, more often constitute the primary source of information for several activities

Previous classifications did not take into account in the time dimension of information investigated in [85]. According to its change frequency, we can classify source information into three categories:

- *Stable* information is information 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 information* is information that has a 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 information* is information 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 information, i.e., stable, long-term-changing, and frequently changing information, are increasingly more complex.

1.5 Information Quality and Types of Information Systems

Information is 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 an organization or the set of organizations that make use of the information system. In order to clarify the impact of information quality on the different *types of information systems*, we adopt the classification criteria proposed in [486] 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. *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, and markup languages. *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 (1) a fully hierarchical system, where only one subject decides for all and no autonomy at all exists, and (2) a total anarchy, where no rule exists and each component organization is totally free in its design and management decisions. For the three criteria, we adopt for simplicity a <yes, no> classification, warning the reader that there is a continuum among extreme solutions.

The three classifications are represented together in the classification space in Fig. 1.3. Six main types of information systems are highlighted in the figure: Monolithic, Distributed, Data Warehouses, Cooperative, Cloud, 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 several tasks, including business analytics and management decision making. The most critical problem in DW design concerns the cleaning and integration of the different data sources that are loaded into

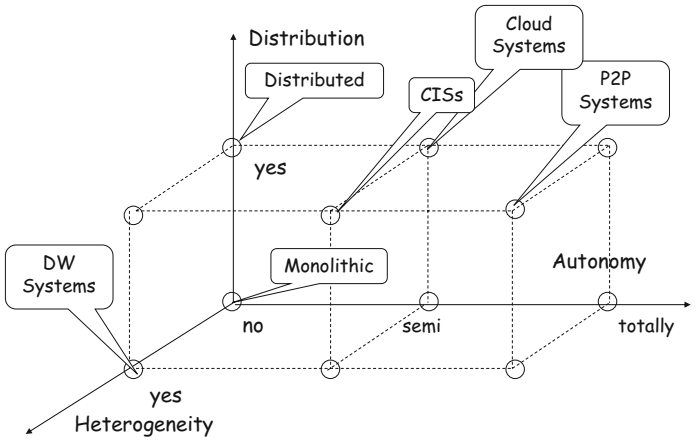

Fig. 1.3 Types of information systems

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 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. 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 [170], the manifesto of CISs, "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 CISs and information quality 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 information. On the other hand, information flows are less controlled than in monolithic systems, and the quality of information, 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 information quality problems, because inconsistencies in information stored at different sites make it difficult to provide integrated information.

- *Cloud information systems* consist of groups of remote servers that are networked to allow centralized data storage and online access to computer services or resources.[1] For these systems, autonomy and heterogeneity are partial due to the presence of a logically centralized data storage.
- *Peer-to-Peer information systems* are based on equal roles with respect to network communications (differently from client-server communication). These systems do not need a central coordination and hence exhibit the maximum level of autonomy as well as of heterogeneity.

The above illustrated typologies do not take into account the fact that there are several "domain/specific" information systems. As a significant example, it is worth mentioning another type of information system largely used in hospitals and clinics: healthcare information systems. These systems are characterized by usage of a vast amount of different types of information, from structured data to semistructured blood test outcomes, handwritten documents, and images as the result of radiographies or ultrasound scans, and exhibit challenging objectives in the integrated analysis of such heterogeneous types of information. Healthcare information systems will be discussed in Chap. 13.

1.6 Main Research Issues and Application Domains

Due to the relevance of information quality, its nature, and the variety of information types and information systems, achieving information quality is a complex, multi-disciplinary area of investigation. It involves several research topics and real-life application areas. Figure 1.4 shows the main ones.

Research issues concern first of all techniques , to some extent models, and two "vertical" areas that cross the first two, i.e., dimensions and methodologies. We will discuss them in Sect. 1.6.1. Three of the application domains mentioned in Fig. 1.4, namely, e-Government, life sciences, and the World Wide Web, in which IQ is particularly relevant, are discussed in Sect. 1.6.2.

Research issues in IQ originate from research paradigms initially developed in other areas of research. The relationship between information quality and these related research areas will be discussed in Sect. 1.6.3.

[1] Wikipedia definition available at http://www.en.wikipedia.org/wiki/Cloud_computing#Architecture.

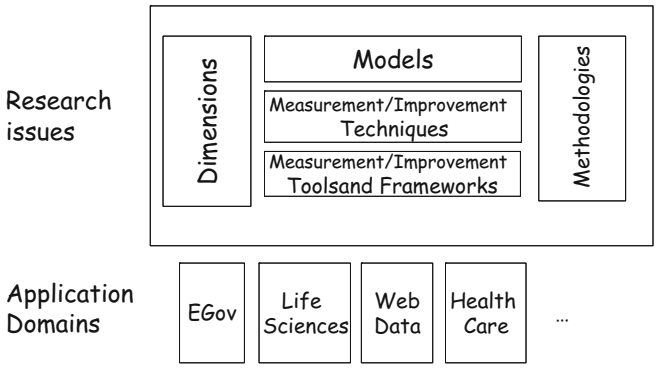

Fig. 1.4 Main issues in information quality

1.6.1 Research Issues in Information Quality

Choosing *dimensions* to measure the level of quality of information is the starting point of any IQ-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 [327]) in order to establish mature concepts in the areas of quality characteristics, measurable indicators, and reliable measurement procedures. Dimensions are discussed in Chaps. 2–5, 13, and 14. Dimensions are applied with different roles in techniques and models.

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 information quality. Models are investigated in Chap. 6.

Techniques correspond to algorithms, heuristics, knowledge-based procedures, and learning processes that provide a solution to a specific IQ problem or to an *information quality activity*, as defined in Chap. 7. Examples of IQ activities are identifying if two records of different databases represent the same object of the real world or not and finding the most reliable source for some specific information. IQ activities are defined in Chap. 7, and techniques are discussed in Chaps. 7–10, and 14.

Methodologies provide guidelines to choose, starting from available techniques, the most effective IQ measurement and improvement process (and hopefully, most economical for comparable results) within a specific information system. Methodologies are investigated in Chap. 12.

1.6.2 Application Domains in Information Quality

In this section, we analyze three distinct application domains of IQ, shown in Fig. 1.4: e-Government, life sciences, and the World Wide Web. Their importance has been growing over the last few years, because of their relevance in the daily lives of people and organizations. A fourth domain, healthcare, will be discussed in detail in a dedicated chapter, namely, Chap. 13.

1.6.2.1 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 information 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
3. The creation of portals that simplify access to services by authorized users

e-Government projects may 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 have 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' information (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 information. Also, errors may occur when personal information on individuals is stored. Some of these errors are not corrected, and a potentially large fraction of them is not detected. Furthermore, information 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 information 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.

1.6.2.2 Life Sciences

Life sciences information and specifically biological information are characterized by a diversity of information 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 information, for instance, in order to perform reliable in silico experiments. However, the effort to actually assess the quality level of such information is entirely in the hands of the biologists; they have to manually analyze disparate sources, trying to integrate and reconcile heterogeneous and contradictory information in order to identify the best information. Let us consider, as an example, a gene analysis scenario. Figure 1.5 shows an example of a simple information analysis pipeline. As the result of a microarray experiment, a biologist wants to analyze a set of genes, with the objective of understanding their functions.

In Step 1, the biologist performs a Web search on a site that is known to contain gene information for the particular organism under consideration. Once the information 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

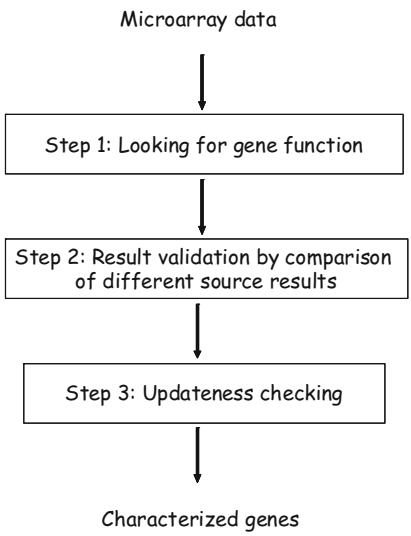

Fig. 1.5 Example of biological information analysis process

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 [619]). 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 information quality techniques.

1.6.2.3 World Wide Web

Web information systems are characterized by the presentation of a large amount of information 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 information structures and, in the case of Web sites, of good navigational paths between pages and certification of information to verify its correctness are costly and lengthy activities. However, the publication of information on Web sites is subject to time constraints.

Web information systems present two further aspects in connection to information 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 *information owner*, responsible for a certain information category. In fact, information 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 information.

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. Information quality for Web and Big Data is discussed in Chap. 14.

1.6.3 Research Areas Related to Information Quality

Information quality is fairly a new research area. Several other areas (see Fig. 1.6) 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 information 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 information. 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 [410], statistics and statistical methodology as the basis of information analysis are concerned with two basic types of problems: (a) summarizing, describing, and exploring information and (b) using sampled data to infer the nature of the process that produced the

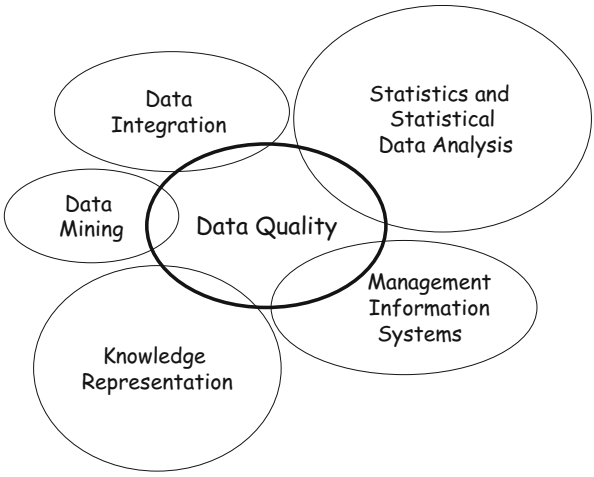

Fig. 1.6 Research areas related to information quality

information. Since low-quality information are an inaccurate representation of the reality, a variety of statistical methods have been developed for measuring and improving the quality of information. We will discuss some statistical methods in Chaps. 7–9.

2. *Knowledge representation* (see [166, 467] 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 (which is called *knowledge reasoning*). Knowledge about an application domain may be represented procedurally in the 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 information quality; we will see some of these techniques in Chaps. 8 and 9.
3. *Data and information mining* (see [294]) 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 [161] as the preliminary process of discovering structure in a set of data using statistical summaries, visualization, and other means. In this context, good-quality information is an intrinsic objective of any data mining activity (see [63]), since otherwise the process of discovering patterns, relationships, and structures is seriously deteriorated. From another perspective, data and information mining techniques may be used in a wide spectrum of activities for improving the quality of data; we will examine some of them in Chaps. 7–9.
4. *Management information systems* (see [164]) are defined as systems that provide the information necessary to manage an organization effectively. Since information and knowledge are becoming relevant resources both in operational and decision business processes, and poor-quality information 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 information resource.
5. *Data integration* (see [397]) has the goal of building and presenting a unified view of data owned by heterogeneous data sources in distributed and CISs. Data integration will be considered in Chap. 7 as one of the basic activities whose purpose is improving information quality and will be discussed in detail in Chap. 10. 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.

1.7 Standardization Efforts in Information Quality

ISO has enacted in 2008 the standard ISO/IEC 25012:2008 (see [330]), for what in the standard is defined as data quality, that is, "the degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions," and provides "a general data quality model for data retained in a structured format within a computer system." The document presents:

- A set of terms and definitions for concepts involved
- Two perspectives that can be adopted when considering data quality characteristics (or dimensions in the following of this book): the inherent perspective and the system-dependent one.

In Fig. 1.7, we see all dimensions defined in the ISO standard.

When we look at the definitions of data and information proposed in the document, we discover that:

1. Data is defined as "reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing"
2. Information is defined as "information-processing knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context have a particular meaning"

DQ characteristic	Definition (all definitions except for completeness and accessibility begin with: the degree to which data has attributes that…")
Correctness	correctly represent the true value of the intended attribute of a concept or event in a specific context of use
Completeness	subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use
Consistency	are free from contradiction and are coherent with other data in a specific context of use
Credibility	are regarded as true and believable by users in specific context of use.
Currentness	are of the right age in a specific context of use
Accessibility	data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability
Compliance	adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use
Confidentiality	ensure that it is only accessible and interpretable by authorized users in a specific context of use
Efficiency	can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use
Precision	are exact or that provide discrimination in a specific context of use
Traceability	provide an audit trail of access to the data and of any changes made to the data in a specific context of use
Understandability	enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use
Availability	enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use
Portability	enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use
Recoverability	enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use

Fig. 1.7 ISO standard data quality dimensions

This choice is different from our choice in the Preface, and it is also specular to the usual one in textbooks and scientific papers, where information is defined in terms of data (see, e.g., [241]).

The ISO effort shows several limitations, such as follows:

- The flat classification adopted among characteristics that is not coherent with, e.g., the classification provided in the document "ISO/IEC 9126 Software engineering—Product quality, an international standard for the evaluation of software quality," where quality characteristics are expressed in terms of sub-characteristics.
- Several characteristics (e.g., completeness) depend on the model adopted for data representation, even though this dependence is not explicitly discussed.
- Data organized in models that neatly distinguish between instances and schemas are considered, e.g., the relational model, while schemaless types of information, such as textual documents, are ignored.
- There is no attempt to distinguish between different types of data and information, from structured data to texts and images.

1.8 Summary

In this chapter, we have perceived that information quality is a multidisciplinary area. This is not surprising, since information, in a variety of formats and with a variety of media, is used in every activity and deeply influences the quality of processes that use information. Many private and public organizations have perceived the impact of information quality on their assets and missions and have consequently launched initiatives of large impact. At the same time, while in monolithic systems information is processed within controlled activities, with the advent of networks and the Web, information is created and exchanged with much more "turbulent" processes and needs more sophisticated management.

The issues discussed in this chapter introduce the structure of the rest of the book; dimensions, models, techniques, and methodologies will be the main topics addressed. While information 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 information quality; with statistical data analysis, also knowledge representation, data and information mining, management information systems, and data integration share some of the problems and issues characteristic of information quality and, at the same time, provide paradigms and techniques that can be effectively used in information quality measurement and improvement activities.