

Chapter 7

Strategic Issues Relating to Data Quality for E-Government: Learning from an Approach Adopted in Belgium

Isabelle Boydens

Abstract Data quality is a strategic matter in the context of e-government as the integration of services requires authentic, coherent, and reliable data. However, establishing databases that are devoid of duplication, redundancy, or ambiguity isn't simple either in theory or in practice. In the context of e-government, this problem has been neglected for too long, particularly because administrative databases have often been wrongly regarded as "simple." We demonstrate in this chapter that this is not the case at all, in particular because of the questions of interpretation that they raise. This chapter is based on case studies stemming from the Belgian federal administration (social security, business directories, federal authentic sources, etc.). Contrary to the assertions of common theories postulating a permanent bijective relationship between data and the corresponding reality, we argue that an empirical information system evolves over time along with the interpretation of the values that it allows one to determine. To address data quality, we propose a temporal framework that provides new operational strategies to improve administrative data quality (mainly, new ways to define quality indicators for continuous monitoring and re-engineering strategies). We finally demonstrate how our approach is generally applicable in the context of empirical information systems.

7.1 Introduction

The dematerialization of information and the placing online, via the Internet, of transverse services for citizens, based on electronic government, make the question of data quality more crucial than ever. We present firstly a general outline of the "data quality" concept and its practical challenges (Sect. 7.1.1) and, secondly, offer an introduction to the strategic data quality issues for e-government (Sect. 7.1.2).

I. Boydens (✉)

Département Sciences de l'Information et de la Communication – CP 123,
Université Libre de Bruxelles, Faculté de Philosophie et Lettres,
CP 123 Avenue F.D. Roosevelt, 50, B-1050, Bruxelles, Belgium
e-mail: iboydens@ulb.ac.be

7.1.1 *The Quality of Data*

The quality of data is today considered as a strategic matter [BAT06]. The question is of significant importance when the information is used as a tool to assist with decision making, or even with real-world action. For example, in May 1999, during the war in Kosovo, NATO mistakenly bombed the Chinese embassy in Belgrade: the mapping databases then used to guide the missiles contained a plan of the city that was obsolete and therefore inadequate, hence, the untimely attack and the diplomatic incident which followed [BOY07].

The quality of a data element denotes its adequacy with respect to the objectives assigned to it. “Total quality” does not exist, because the concept is relative: on the basis of a cost–benefit type analysis, the most pertinent quality criteria (freshness of information, rapidity of data transmission, relevancy, etc.) must be adopted for a given context (*fitness for use*) [BOY11].

These questions are of increasing concern in the private sector [MAD09]. Several surveys carried out in the United States indicate that factors such as the multiplication of partially redundant heterogeneous sources and of incomplete or poorly documented data could entail a cost of up to 15% of businesses’ revenue [RED01]. Added to this are the costs incurred for the implementation of new technologies [FRI07] as well as the consequences in terms of credibility in the eyes of clients or users.

7.1.2 *Strategic Quality Issues for E-Government*

Although the problem is just as significant in the context of electronic government, it has been neglected for too long, particularly because administrative databases have often been wrongly regarded as “simple” information systems. We show in this chapter that this is not the case at all. The management of such systems is complex, not only because of the questions of interpretation that they raise, but also because administrative information gives rise to rights and duties. The quality of the corresponding online services is therefore of considerable importance in social and financial terms. For example, every quarter, the Belgian social security databases store approximately four million records, with several hundred corresponding attributes. These databases allow the collection and redistribution of some 40 billion euros each year. Every quarter, hundreds of thousands of formal anomalies are detected (one finds similar proportions in other countries and other sectors, such as the banking sector: “Recent works about the quality of large databases have shown that about 10% of XML documents (or data records) contain at least one error. This level of quality is unacceptable for many applications” [VAN03]).

On the basis of our research work in the area of administrative database quality [BOY99, BOY11] and of our practical experience in consultancy [BOY07], we propose to examine these questions in greater depth. In Belgium, we are head of a “Data Quality Competency Center” for evaluating and improving the quality of databases deployed in electronic government, which has been created at Smals.¹

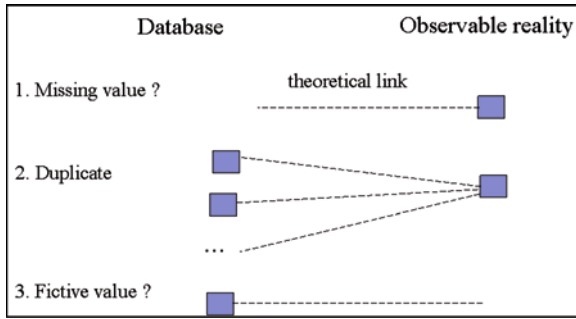


Fig. 7.1 Typology of formal data quality problems

This Data Quality Competency Center covers various domains (social security, business directories, federal authentic sources, etc.) and aims, by improving data quality, to strengthen the responsiveness of the administrations in the context of online services provided to citizens. We then demonstrate the generalisable character of this in the context of electronic government globally.

The integration of services requires that authentic, coherent, and reliable sources be put in place. For instance, in order to allow a company to fulfill its tax and social security obligations, or indeed to notify any events (e.g., change of address), via a set of integrated services on the Internet, there must be a back-office body of homogeneous databases relating to the target population (companies, in our example). These databases must be pertinent, accurate, up to date, and documented, because they are liable to be used transversely by various departments according to their respective needs. As we show, establishing files that are devoid of duplication, redundancy, or ambiguity is not a simple matter in practice. Two types of cases may arise: formal data quality problems and formal data interpretation problems.

Formal data quality problems (Fig. 7.1) may include the following situations.

1. *Missing values*: for instance, due to a lack of reactivity, a new company created in the “observable reality” is not yet registered in a database.
2. *Duplicate*: due to a lack of organization and harmonization, an international company created in the “observable reality” is registered twice because, for instance, two social security institutions in various countries use different meta-data types and formal identifications to categorize a same item.
3. *Fictive value*: a company that stopped its activities (because it is bankrupt, for instance) did not inform the institution in charge of the database management.

Data quality interpretation problems are due to the fact that administrative databases are empirical information systems subject to human experience. Administrative databases, by their nature, raise particularly complex questions of interpretation that we tackle in this chapter. For example, the notion of “principal activity” of a company, which is fundamental information in businesses registers, is a mutable factor whose reliability is difficult to evaluate (for instance, in the woodwork sector, the difference between manufacture and installation, as “principal economic activity,” may be very subtle and evolutive). These problems are also encountered on a daily basis by labor

inspectors doing fieldwork: “In almost 30 years of doing this job [...], I have seen many points become subject to interpretation, and therefore to challenge. It’s a little like a policeman stopping you for going too fast and telling you that you’re driving at a “dangerous” speed. You could easily contest this notion” [BAR06]. Consequently, there is no “absolute reference” to check the correctness of an administrative database (nor, as we show later, of the correctness of any empirical database).

Moreover, in the context of a shared exploitation of information gathered in a single flow, as in the architecture of electronic government, a tradeoff may arise among the needs of the various departments using that single source. We see this phenomenon in the case of the multifunction online declarations developed in Belgium in the field of Social Security. For legal reasons associated with the payment of fixed-date welfare benefits, some user bodies need to have the information very quickly, in spite of any anomalies by which it may be tainted. Conversely, other social security institutions prefer all formal anomalies to be dealt with before the data are distributed. There is therefore a problematic tradeoff between the speed of distribution of administrative information and its relative reliability, with the quality criteria varying according to the respective uses.

However, if the deployment of electronic government raises new challenges regarding the quality of information, it also offers fresh prospects for improving the quality of data. Thus, it is now possible to provide citizens with online simulation environments in order to facilitate the declaration of social security contributions, for example, and to strengthen the partnerships between public and private sectors. Furthermore, it is essential to document the services offered and the corresponding digital resources. Meta-information systems for documenting administrative regulations through their successive versions, but also for generating the corresponding XML schemas (with regard to social security legislation, in particular), can now be made available to all.

The remainder of this chapter is structured as follows: after this introduction, we describe the characteristics of the data exploited in the field of electronic government (Sect. 7.2) on the basis of various operational definitions. It is then possible to specify the most appropriate quality indicators for the objectives pursued (Sect. 7.3). We then consider methodological recommendations for evaluating and improving the quality of the corresponding databases, including the new prospects now offered by electronic government (Sect. 7.4). By way of conclusion (Sect. 7.5), we present future work and widen the question to include other empirical information systems.

7.2 Characteristics of Administrative Data

Electronic government services rely essentially on data directories. For instance, the governmental business databases mentioned in the introduction may be exploited for a variety of purposes by different institutions. In Belgium, the directory of business companies can be used by various authorities, such as social security institutions or the agency responsible for monitoring the food chain. This agency needs the addresses of the production units of food sector companies in

order to carry out its functions (food safety supervision). It is thus important for the various data (relating to the businesses and their production units) to be interpreted in the same way by each institution and handled in a homogeneous and reliable manner. For this reason, we consider in greater depth the characteristics and the nature of administrative data.

Administrative databases exhibit a number of distinctive characteristics: frequency and nature of legislative changes, compliance with probative force, volume of data and of formal anomalies to be handled, and finally, the social and financial stakes.

The structure of administrative databases evolves according to changes in the corresponding legal directives. In the domain of Belgian social security, for example, legislative changes, entailing an equivalent number of schema versions, must be implemented every 3 months. The question becomes more complex when these changes are retroactive: even if contested, retroactive changes occur frequently in administrative life in various European countries. Furthermore, the successive versions must be jointly maintained at least for the period of prescription, which specifies the time for which the administrative files must legally be taken into account. In the case of Belgian social security, this period varies from 5 to 30 years depending on the sector concerned.

Moreover, most of the original administrative information entered into the databases has a probative force status: in other words, it serves as court evidence in the event of any dispute. This is true, for example, of the quarterly declarations submitted by employers to prove payment of their employees' social security contributions. Consequently, the original information, even if tainted by formal anomalies, must be conserved. Similarly, the history of its processing must also be retained for several schema versions (these anomalies may be corrected or validated following inspections on the ground or after the interpretation of legislation). Ultimately, no error tolerance is theoretically permitted within databases. Citizens legitimately expect their administrative affairs to be handled equitably, whether with regard to taxes to be paid or welfare benefits to be received.²

Finally, administrative databases are generally extremely voluminous. Potentially, they may house the files of an entire state's population. Considerable social and financial stakes are involved in their management, as already mentioned in the introduction of this chapter.

7.3 Quality Indicators

Any process aimed at improving the quality of information involves the prior specification of indicators. These will then make it possible to evaluate the progress made with respect to the objectives pursued. In order to identify the most appropriate indicators, it is first necessary to examine the nature of the handled concepts.

To this end, we consider in succession the following three questions (in the light of the administrative information characteristics, as cited in the previous point): what is a data element (Sect. 7.3.1), what is a "correct" data element (Sect. 7.3.2), and how is the information progressively constructed (Sect. 7.3.3).

7.3.1 *What Is a Data Element?*

A data element is a set of three components (i, d, v): an identifier (i), referring to a concept (e.g., a category of activity); a domain of definition (d), comprising a set of formal assertions specifying all the values admissible in the database for this concept (e.g., a controlled list of alphabetical values), and finally, a value (v) at a time t (e.g., the chemicals sector). In addition, there are also interactions among the different components of the database schema, which we do not consider here.

It is important to distinguish *deterministic data* from *empirical data*. The first are characterized by the fact that there is, at any moment, a theory which makes it possible to decide whether a value (v) is correct. This is the case with algebraic data: inasmuch as the rules of algebra do not change over time, we can know at any time whether the result of a sum is correct. But for empirical data, which are subject to human experience, theory changes over time along with the interpretation of the values that it has made possible to determine. This is true, for example, in the medical domain, where theory evolves with the accumulation of experience, as witnessed, for instance, in the current research into influenza A(H1N1)) and the economic domain (e.g., with regard to the calculation of national wealth), but also in the legal and administrative realms, where the interpretation of legal concepts changes with the constant evolution of real-world circumstances and jurisprudence. For example, when “copy centers” first began to appear, (i.e., shops offering photocopying services to their customers), the nomenclature for European business activities (used in administrative databases to categorize companies) was quickly found to be inadequate for their classification: the best it could offer were the categories of “printing,” “book retailing,” or “secretarial services.” To take the category “copy centers” into account, it was first necessary to amend the regulatory texts, and then to adapt the structure of the administrative databases accordingly. Problems related to empirical data quality remain crucial in the context of electronic.

7.3.2 *What Is a “Correct” Data Element?*

For obvious operational reasons, the functioning of a database is predicated on the *closed world assumption*, according to which any value not included in the domain of definition (d) is regarded as false. However, in the case of empirical data, if we step outside this formal framework, it may happen that between the moment at which the structure of the database was formalized and the moment at which the information was entered, new characteristics have appeared within the domain in question (contrary to the assertions of some theories postulating a permanent bijective relationship between data and the corresponding empirical reality [WAN96]). In this case, it is impossible to verify the correction of the database values automatically. Consequently, when an inconsistency appears between a value entered in the database and the reference tables which allow the validity of that value to be tested, it may become essential, depending on what is at stake, to carry out a manual verification, for example, by contacting the citizen or company concerned.

There is therefore no “absolute” formal reference for testing the correction of a huge empirical database. Let us take an example. We know that social legislation differs according to whether it is applied to manual or clerical staff, with these two groups being distinguished in the law according to the preponderant nature of their manual or intellectual activities. In practice, this distinction is not easy to apply, but no fuzziness is tolerated in a database: everything must be clear-cut. In order to arrive at a clear-cut answer, it will often be necessary to be doing fieldwork to interpret de facto situations and examine supporting documentation. As new interpretations are made and jurisprudence evolves, the meaning of the notions of “clerical” or “manual” staff will evolve over time.

We can conclude from this discussion that data are not a given: they are progressively constructed. It is all the more essential for these conclusions to be taken into account when several institutions are involved. Thus, as mentioned in the introduction, in Belgium, electronic government projects have given rise to the adoption of a multifunction declaration (one information flow can be used for multiple administrative institutions) in the social security sector, so that the citizen now has to enter his or her information online only once, this information then being exploited by the various social security sectors. Consequently, the question of conceptual interpretation also takes on a multifunctional dimension. As we saw in the introduction, this question may give rise to tradeoffs according to the needs of the institutions, particularly between the relative quality of the information and the speed of its distribution.

7.3.3 *How Are Data Progressively Constructed?*

Braudel’s temporal framework (“*temporalités étagées*” [BRA76]) can be applied in the database field. Three levels of transformation are interacting within the information system: the evolution of jurisprudence, the changes made within databases, and the categories observable in the field. These three levels of reality are interlinked, interlinked, but asynchronous [BOY04]. They operate, according to their nature, on different timescales. Thus we have the long-term for legal rules, renewed from one quarter or one year to the next, the medium-term for the management of databases, and the short-term for the observable reality, that is, that of the citizens or companies subject to administration, which is continuously evolving. Companies regularly merge, split, or disappear altogether, and new professions or categories of activity not covered in the official nomenclature are constantly being born, as with the diversification of IT jobs, for example. From a dynamic point of view, an ideal database should therefore match the rhythm of its updates to the (unforeseeable) division into “layered timescales” of the changes in the reality that it seeks to grasp. To what looks like a gamble we must add the necessity, always revealed a posteriori, to integrate unforeseen observations, prohibited a priori by the closed world assumption.

Let us take an example in the domain of employment creation initiatives (Fig. 7.2). Following the directives issued by the European Council in Brussels in December 1993, on the basis of Jacques Delors’ white paper on growth,

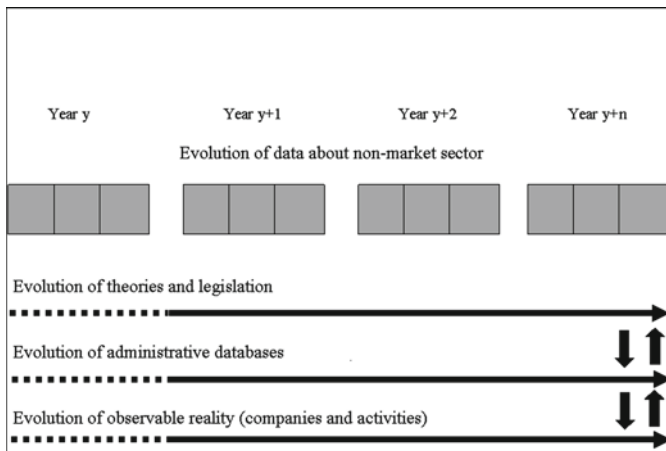


Fig. 7.2 Transformation mechanisms to interpret administrative data

competitiveness and employment, multiple job creation initiatives have been taken in most European countries with the aim to fight unemployment. Among this legislation, reductions in social charges have resulted in lowering the social security contributions payable by employers, in order to enable them to take on additional employees. In many countries of the European Union, these initiatives have produced a flood of legislative directives and adjustments which complicate not only their implementation but also the evaluation of their effectiveness.³ The problem is ever-present: owing to the proliferation of legal texts in France, in March 2006 the Council of State proposed drafting a law to reduce their number: “[...] to the 9000 laws and 120,000 decrees on the books in 2000, an average of 70 laws, 50 orders and 15,000 decrees have been added every year [...]” [DEM06].

In the case of Belgium, during the implementation of a governmental directive aimed at the “nonmarket” sector, the question arose with regard to the reality that is progressively reflected in the database, of whether this “nonmarket” sector should include private nursing homes, which were a priori excluded because they operate for profit. Initially regarded as “erroneous” cases with respect to the domain of definition for the “nonmarket sector,” these businesses were eventually included after legal interpretation. This led to a restructuring of the database schema. This restructuring was the result of a human decision aimed at bringing the model temporarily into line with the new observations. This phenomenon of transformation corresponds to the so-called “strange loop” mechanism defined by Hofstadter [HOF80]. In the absence of such an intervention, the gap between the database and the reality widens. We show later (Sect. 7.4) the operational extensions of these mechanisms when it comes to evaluating and improving the quality of administrative data.

What are the consequences of this analysis with regard to specifying appropriate quality indicators for administrative information? Because administrative data are

empirical in nature, there is no direct frame of reference for testing their correction. Their appropriateness to the needs can be determined only indirectly, via a series of lateral indicators. Firstly, it is necessary to consider the relative relevance of the information with respect to the objectives pursued: relevance is a nonquantifiable indicator whose operational scope is examined in Sect. 7.4.1. (“Master Data Management”). Next, a series of quantifiable indicators relating to the detected anomalies and their handling may be produced with a view to deploying management strategies for the database (Sect. 7.4.2). Finally, in all cases, there must be a tool for the critical interpretation of the data: we present a meta-information system implemented to document the online services of the Belgian social security authorities (Sect. 7.4.3).

7.4 Methods for Improving the Quality of Administrative Databases

The methods presented below for improving the quality of information should ideally be applied during the design of a database, because questions of quality arise at the very outset. They should also be accompanied by continuous monitoring, carried out by a committee designed for this purpose, with a prohibition of hasty ad hoc actions [BLO05].

1. *Master Data Management* is a general methodology to analyze and improve the quality of the concepts and flows judged to be the most fundamental within the information system.
2. *Anomalies and Management Strategies* are an original operational approach that we applied in the scope of our research about interpretation of the Belgian social security database.
3. *Documentation of Application and Services* aims to present an electronic data dictionary (*glossaires de la sécurité sociale*) that was implemented in Belgium to improve interpretation of e-government databases by the Belgian Data Quality Competency Center presented in the introduction.

We do not consider the data cleansing technique here, which involves automatically smoothing the content of a database a posteriori by eliminating (using a set of formal correction algorithms) the values considered to be aberrant, for example. This technique has its place in the statistical domain, where an error tolerance is permissible. It is, however, not valid in the administrative domain, where each individual case must be considered [OLS03]. We must also mention that data cleansing does not act on the causes of the “no-quality.” As suggested by Thomas Redman [RED01], an information system can be compared to a river: the algorithms of data cleansing clean up the river bed in an ad hoc manner, but they do not act on the procedures situated upstream. However, it should be noticed that these types of algorithms are useful operational techniques that facilitate the implementation of the three methods of data interpretation

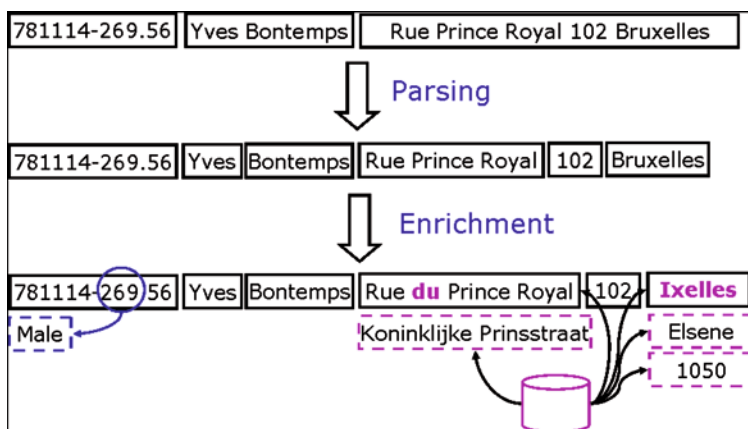


Fig. 7.3 Data standardization: illustration of data parsing and enrichment

improvement presented in the rest of this chapter (for instance, as illustrated in Fig. 7.3, to parse and automatically enrich multilingual data).

7.4.1 Master Data Management

On the basis of the core business and main needs of the application domain concerned, “Master Data Management” [LOS09] aims to analyze and improve the quality of the concepts and flows judged to be the most fundamental within the information system.

For instance, in the intersectorial relationships involved in electronic government, the question of the identification of citizens is crucial for reasons of security and confidentiality, but also for operational reasons. It is essential that a given citizen or a given company is correctly identified by the various services accessible online.

In the first instance, it will be necessary to check whether the unique identifier used corresponds to the intended target of the application domain [VOL06]. Service to the citizen must take precedence over internal organization: the identifier must therefore not include any “content” information. In the case of a bank, for example, a customer must be identified by a randomly generated number and not by his or her account number, inasmuch as this may change and generate double entries if it is used as the primary key.

This question has special implications in the administrative domain. Even if the above rule of Master management has been observed, problems may arise in relation to the unique identifier, because the flows that allow the creation of a company in a directory, for example, have been imperfectly defined. This may entail double entries or a phenomenon of “undercoverage” (absence of pertinent elements). Thus we sometimes see, in the architecture of an information system, the

phenomenon of the “phantom factory” or the “factory within the factory,” meaning the time and energy an organization devotes (unknowingly) to producing and correcting errors. In company directories, “overcoverage” is also sometimes observed (presence of “false actives” in the database) owing to realities in the field: for example, companies which have ceased trading but omitted to inform the administration that they are no longer active.

Finally, we must note that the handling of double entries requires complex procedures: firstly, it is necessary to define what constitutes a double entry and to implement procedures for identifying the corresponding cases. There are algorithms for this purpose which make it possible to take account of a certain imprecision where certain strings of characters (relating to a name or an address, for instance) may have been entered on several separate occasions, giving rise to double entries in the database. So-called “matching” techniques [BAT06] are used to compare the records of a database with those of an authoritative competing source (or “frame of reference”), and these also make it possible to detect any inconsistencies in the identifier. Once the double entries have been detected, homogeneous rules must be defined in order to use a priority number and resolve any discrepancies between the values associated with the fields of the various records constituting cases of double entries.

Finally – and this is specific to the administrative domain – procedures must be defined for carrying the correction made over into legislation. Thus, when errors of identification concerning companies with “legal person” status have been dealt with in the database, those companies’ deeds of incorporation must also be amended. In Belgium, a procedure has been enacted into legislation for handling these cases in the directory of natural persons.⁴ In Germany, the problem is complicated by the need to manage heterogeneous numbering systems relating to different sectors of the administration.

7.4.2 Anomalies and Management Strategies

Quantitative monitoring of anomalies and their handling allows the deployment of original strategies for database interpretation and management, with measurable cost–benefit results. The strategies we present here are based on analysis presented in Sect. 7.3 (“Quality Indicators”) and have been implemented in Belgium by the “Data Quality Competency Center” of Smals.

There are several essential prerequisites for supporting this approach. A system for detecting anomalies at the time of entry, but also *ex post*, must be put in place (owing to the potential interactions between data elements, formal anomalies may arise subsequently after the correction of other anomalies). Clear procedures for their handling must be established, particularly when the database forms part of a federalized environment and several institutions are each responsible for a subset of the information. It must be clearly defined which authority handles which part of the database and what the authorized handling procedures are. This is often a tricky

matter in practice, because it falls under the political responsibility of each institution concerned. The database must be structured in such a manner that traceability is guaranteed (i.e., the history of data handling operations is stored and can be queried). Finally, a clear procedure for the production of indicators at a given frequency must be defined.

On the basis of the analysis proposed in Sect. 7.3, we present an example of the operational exploitation of the presented quality indicators. Statistical monitoring of integrity constraint violations (“formal anomalies”) makes it possible to detect not only “abnormal” increases in anomalies (with respect to a given threshold), but also increases in “validations” of anomalies during the handling phase. A validation operation means that, after examination, an operator has judged that the anomaly, which is a presumption of error, corresponds to a relevant value. The operator can “force” the system to accept the value. If the rate of these anomaly validations is high and recurring, there is a high probability that the structure of the database itself is no longer relevant. An algorithm then issues a “signal” to the database manager so that she can examine whether a structural modification of its schema is required. When there are large numbers of validations, it is worthwhile to examine the phenomenon closely: as we have seen (Sect. 7.3), a new circumstance (e.g., the emergence of a new category of activity or a change in the interpretation of a concept, such as the nonmarket sector as cited above) may have arisen, which requires a modification of the database structure. If the schema is not modified accordingly, the anomalies corresponding to these cases will continue to appear in large numbers, demanding a potentially large-scale manual examination and considerably slowing down the administrative file handling.

For the Belgian social security system, the implementation of this method has made it possible to improve the precision and speed of social security contribution handling, reducing by up to 50% the volume of formal anomalies, which had previously accounted for 100,000 to 300,000 occurrences to be managed manually every quarter [BOY99]. Other types of indicators for identifying, quantifying, and categorizing anomalies and the nature of their handling are essential for the implementation of efficient electronic government services. For example, it is possible to evaluate the speed of anomaly handling in order to determine the timeliest moment for the database exploitation. This type of method is all the more useful when data are collected at a single point and then exploited in a federalized manner by different departments, as with the procedures offered by electronic government [BOY11].

7.4.3 Documentation of Applications and Services

Because the legislation is complex and constantly evolving, an “electronic data dictionary” is regarded as highly necessary in order to facilitate the administrative information interpretation. To this end, in the framework of the services involved in

electronic government, a collaborative multilingual meta-information system was designed within the Belgian social security department. This was deployed in a Web environment in order to document the XML messages exchanged between the citizens and the administration. This meta-information system was put into operation in 2001, and has been enhanced since then.⁵ This Web-based application is aimed both at the IT personnel responsible for the management of the databases and at the authorities responsible for sending electronic messages, the goal being for all parties to work on a common basis.

We would note that meta-information systems are potentially prone to three pitfalls:

- The first is associated with the fact that these systems are infinitely expandable, particularly when the fields to be completed are “free,” because the natural language is its own metalanguage. This involves significant management costs when there are numerous manual updates to be made.
- The second snag is that the metadata may themselves be erroneous or uncertain. When the data are contextual in nature, their validation cannot be made subject to rigorous integrity constraints.
- The third snag concerns the time lag between the updating of a data element and that of the corresponding metadata, because the latter, especially when it takes a textual form, is generally created only after the completion of an analysis phase.

These characteristics may appear in every empirical application domain [VAN06]. On the basis of these observations, a system aimed at preserving the coherence of the information and facilitating its management has been implemented by the Belgian Data Quality Competency Center. The system (“*glossaires de la sécurité sociale*”) includes the following functions.

- Semi-automatic management of multilingualism (via precontrolled tables).
- Reuse of common definitions via an inheritance procedure (see Figs. 7.4 and 7.5); generic definitions such as the codification of locations, addresses, and so on are updated only once and then propagated in all the specific documentary applications where they are used.
- Version management (when technical definitions evolve over time, as shown in Fig. 7.6, the system makes it possible to monitor the various versions and specifies, for each new version, the list of changes made with respect to the immediately preceding version).
- Implementation of the concept of WOPM (“Write Once Publish Many”): the application includes structured lists (postcodes, categories of activity, etc.) which, in practice, must be distributed not only for documentary purposes but also to test the data entered into the databases. In order to accommodate these two functions, the application was designed to automatically generate a single structured table (e.g., list of postcodes) in different formats: ASCII, XML, Word, Excel, and PDF. The same source can thus be used in interdependent applications.
- Navigation system and search engine.
- Validation procedures (Fig. 7.6): owing to the legal, social, and financial stakes associated with the management of databases and their documentation, each

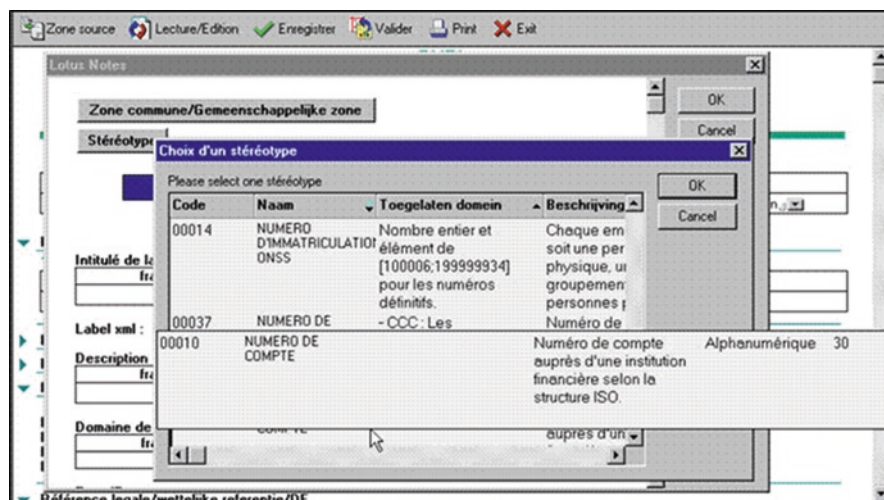


Fig. 7.4 Heritage of common definitions

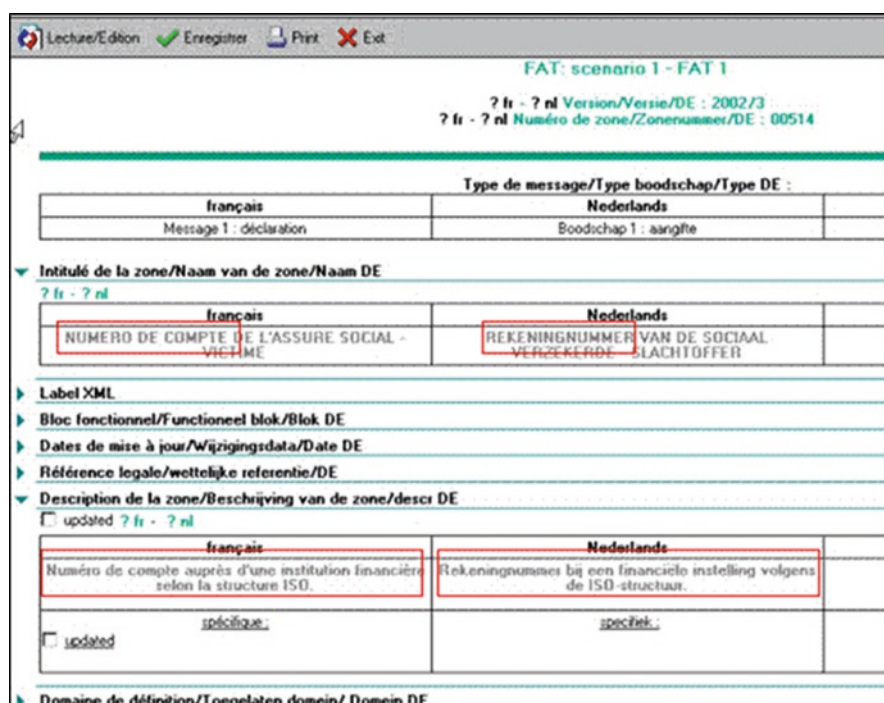


Fig. 7.5 Heritage and reuse (multilingual framework)

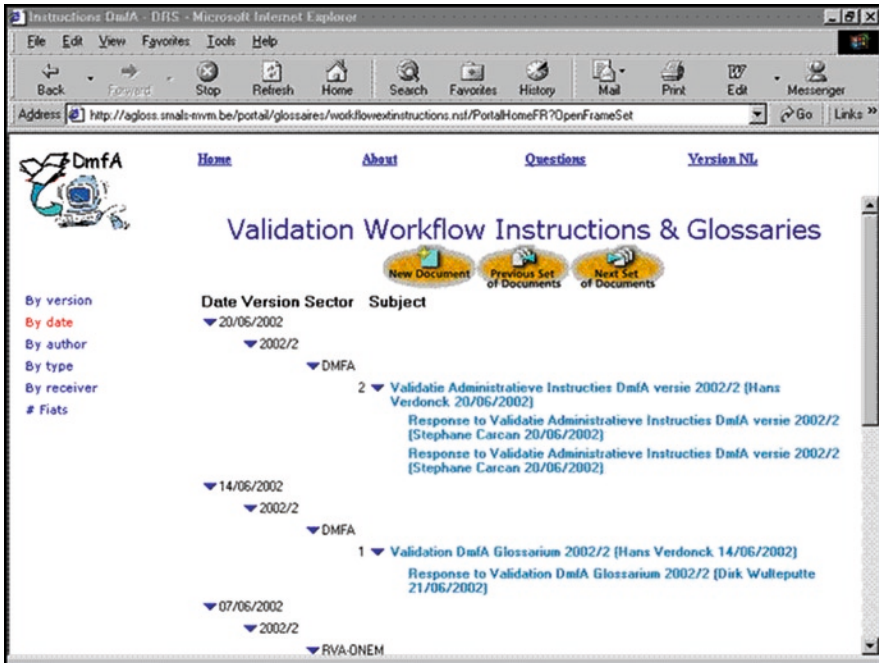


Fig. 7.6 “Glossaires de la sécurité sociale”: validation procedures and version management

new version must be validated by the information manager in both technical and legal terms. In order to structure this validation process, a workflow system guides the deployment of the electronic dictionary. This system is enshrined in a procedure (a schedule rigorously specifies the timing of each update, validation, acceptance, and putting into production). The workflow is “piloted” in a centralized manner by a team dedicated to this task, and deployed in a decentralized manner in a Web environment. At the time of the creation of each new version, the history of exchanges between the various managers is conserved, so that the interpretation process can continue to be monitored.

The implementation of this kind of system facilitates management of the data that feed the online administrative services and helps to ensure their quality.

7.5 Conclusions, Future Work, and Generalization of the Approach

Examination of the quality of a database is a multidisciplinary process. It requires joint intervention both by technicians and by specialists in the concerned application domain. In this regard, the administrative concepts have certain specific

characteristics, including their empirical nature: they are subject to human interpretation and their meaning changes over time.

In the context of electronic government, these characteristics, which we have detailed in the course of this chapter, pose fresh challenges: the pooling of data and dematerialization of procedures demands interoperability between sectors and departments, and this potentially multiplies the interpretation difficulties to be overcome. As shown by a recent report published by the United Nations [UNG05], this makes it essential to implement procedures for the continuous evaluation and improvement of the quality of administrative information. These procedures, as we have seen, involve the deployment of indicators for data quality evaluation, analysis of flows and basic concepts, piloting of management strategies, and documentation of the databases on which the online government services rely.

Alongside these new challenges, the takeoff of electronic government also offers new prospects, such as operational avenues for improving information quality. Thus, we have described the adoption of a multifunction online electronic dictionary, supporting the management of versions and feeding the structured and unstructured applications. Other new solutions are appearing in a spirit of increased partnership between private and public sectors. Among these are the emergence of online simulation environments and the distribution of targeted information via citizens' personal spaces on electronic government portals: a list of the most common anomalies occurring in their official declarations can be sent to them, along with information about the legal consequences of these anomalies and various courses of action for putting them right. On this basis, "process re-engineering" projects can be put in place: errors are frequently committed by certain declaring parties because of an inadequate or heterogeneous interpretation of the procedures or concepts to be handled. In collaboration with the administration, a procedure of "data tracking" allows us to examine the internal processes of the 50 Belgian companies with the highest numbers of anomalies in their social security declarations: the objective is to determine the reason (problem of data interpretation, of interface usability, of formal error, of internal or external organization, etc.) for this and remedy it at the source.

In conclusion, we would note that phenomena similar to those observed in the administrative sector are found in various empirical domains. This is true, for example, in the world of stratospheric databases: before the discovery of falling ozone levels by British researchers in the 1980s, the corresponding low values were, as a matter of course, treated as anomalies in NASA's database for more than a decade [BOY99]. The prevailing theory of the time, which was modeled in the NASA database, did not allow any entertainment of the possibility that such values might be correct. After the British discovery, NASA adapted the structure of its database, integrating the values previously regarded as "anomalies" into the set of permissible values. The management strategy described in this chapter therefore applies to all information systems whose structure evolves according to the interpretation of the realities which they aim to grasp [BOY11]. The body of propositions put forward elsewhere ("Master Data Management," Sect. 7.4.1, documentation of databases, etc.) are all the more essential when the information

system is dynamic in nature. This is particularly true of administrative databases, in which the homogeneity of the formal codifications clashes with the heterogeneity of the empirical categories.

Notes

¹ICT company supplying services to the Belgian federal administration (<http://www.smals.be>).

²Conversely, the statistical exploitation of a database may be predicated on error tolerance. Let us take a simple example: the total of all wages paid to salaried employees is used to evaluate the national wage bill, which in turn allows the calculation of statistical aggregates. If all the records used incorporate wage inversions, with the pay of an individual A incorrectly attributed to an individual B, the overall evaluation of the national wage bill will be unaffected. Such inversions are, however, extremely damaging at the level of individual administrative processings (B being paid the salary of A).

³As in France, where the ‘stratification of measures, the short lifespan of mechanisms, the multiplication of the effects of policy announcements and the growing complexity of the field of these policies, which are still recent in terms of the history of social policy’ complicate their long-term evaluation [DAN 98].

⁴Royal Decree of 8/02/91 on the composition and procedures for allocation of the identification number of natural persons not entered in the National Register of natural persons, *Belgian Official Gazette*, 19 February 1991.

⁵https://www.socialsecurity.be/lambda/portail/glossaires/dmfda.nsf/web/glossary_home_fr https://www.socialsecurity.be/lambda/portail/glossaires/dmfda.nsf/web/glossary_home_nl.

References

- BAR06 Baroux R., “En première ligne sur le front du travail”. *Le Monde*, 2006, p. 3.
- BAT06 Batini C. and Scannapieco M., *Data quality: concepts, methodologies and techniques*, Heidelberg, Springer, 2006.
- BLO05 Bloch L., *Systèmes d’information, obstacles et succès*, Paris, Vuibert, 2005.
- BOY99 Boydens I., *Informatique, normes et temps*, Bruxelles, Bruylant, 1999.
- BOY04 Boydens I., *La conservation numérique des données de gestion (Numéro spécial “Archivage et pérennisation”)*, vol°8, no. 2, Paris, Hermès Sciences, 2004, pp. 13–22.
- BOY07 Boydens I., “Qualité de l’information et e-administration: enjeux et perspectives”. In Assar S. and Boughazala I., (eds.), “Administration électronique: constats et perspectives,” Paris, Lavoisier – Hermès Sciences, 2007, pp. 103–120 (chapter 5).
- BOY11 Boydens I., Van Hooland S., “Hermeneutics applied to the quality of empirical databases”. *Journal of Documentation Emerald*, volume 67, issue 2, 2011.
- BRA76 Braudel F., *“La Méditerranée et le monde méditerranéen à l’époque de Philippe II,”* Paris, Armand Colin, 1976.
- DEM06 De Montvallon J.-B. “Une Loi pour enrayer la prolifération des textes”. *Le Monde*, 2006, p. 10.
- FRI07 Friedman T., *Key Issues for Data Quality, 2007*. Gartner Research Note, 2007, no. G00147383.
- HOF80 Hofstadter R. D., *Gödel, Escher, Bach: an eternal Golden Braid. A metaphorical fugue on minds and machines in the spirit of Lewis Carroll*, New York, Penguin Books, 1980.
- LOS09 Loshin D., *Master Data Management*, Burlington, Elsevier, 2009.

- MAD09 Madnick S. E., Wang R. Y., Lee Y. W., Zhu H., "Overview and Framework for Data and Information Quality Research," *ACM Journal of Data and Information Quality*, 2009, vol. 1, n 1, pp. 2–22.
- OLS03 Olson J., *Data quality: the accuracy dimension*, San Francisco, Elsevier, The Morgan-Kaufmann Series in Database Management, 2003.
- RED01 Redman T. C., *Data quality: the field guide*, Boston, Digital Press, 2001.
- UNG05 *UN Global E-governement Readiness Report 2005. From E-governement to E-inclusion*. New York, United Nations, 2005.
- VAN03 Van Der Vlist E., *Relax NG*, Cambridge, O'Reilly Media, 2003.
- VAN06 Van Hooland S., "Spectator becomes annotator: possibilities offered by user-generated metadata for image databases," *Immaculate Catalogues: Taxonomy, Metadata and Resource Discovery in the 21st Century, Proceedings of CILIP Conference*, University of East Anglia, UK, 13–15 September 2006.
- VOL06 Volle M., *De l'informatique (Savoir vivre avec l'automate)*, Paris, Economica, 2006.
- WAN95 Wand Y., Wang R. Y., "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, 1996, vol. 39, n 11, pp. 86–95.