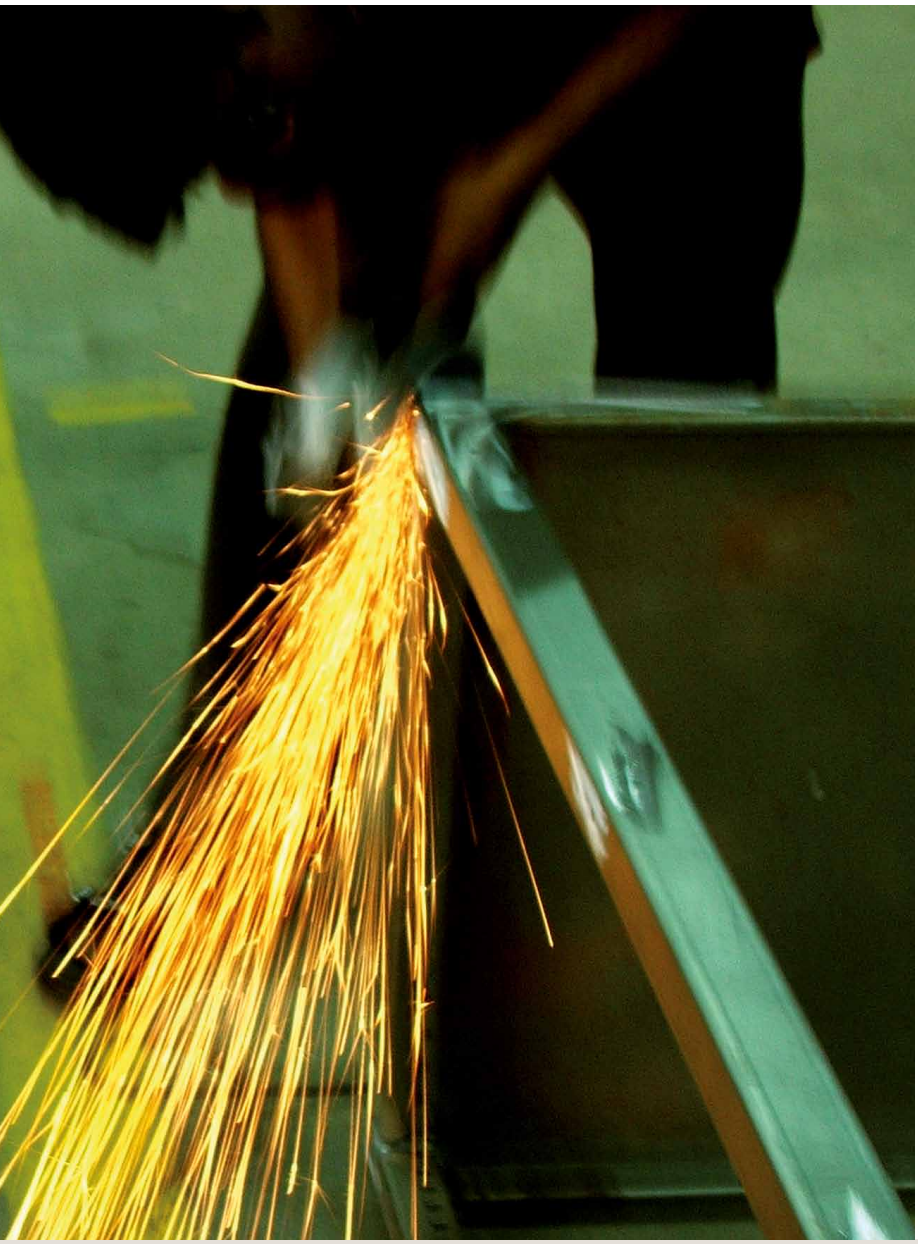
Data quality assessment

Ronald Jonker

This article deals with data quality. Many companies and organizations see data quality as a given fact. To them, it is something that simply exists. Other organizations have come to realize, however, that this is a wrong assumption and that data quality is not self-evident. They have initiated data quality programmes because they are aware of the importance of good data for the efficient and effective management of operational processes. To achieve good data quality, it is necessary for each organization to determine, in a structured way, exactly what 'good data' means to them, as well as finding a way to ensure that the quality of the data remains 'good'.
The description of these themes forms the core of this article. It concludes by providing a number of recommendations for the set-up of data quality programmes.

R.A. Jonker
is a partner at KPMG Management Consulting, IT Advisory.
jonker.ronald@kpmg.nl



Introduction

In the past few years, data quality has been a focus of attention among companies and organizations. In the last decade, the volume of data and databases in which this data has been stored has grown exponentially. Due to the low storage cost and the current possibility of accommodating data in the 'cloud', this trend will continue unabated. To access these databases, great investments have been made in expensive business software. Solutions mostly have been sought in technology. The issue of whether or not this data is of sufficient quality has been completely ignored. It has now become generally known – and accepted, unfortunately – that the quality of data in relational databases is mediocre to poor. In one of the organizations where I once worked as an advisor, management accepted that the data in a database with 7 million records contained four to five per cent of errors. If we presume that each record contains around 20 fields and that this database therefore contains around 140 million data, then we can assume that this database contains approximately 7 million data whose quality is uncertain.

Erroneous data is like a contagious virus. It spreads through the organization, is adopted into other databases, serves as reference data, or forms the foundation of management information reports. With the increasing size of databases, the volume of inaccurate data is also increasing. The result is that inaccurate data is responsible for poor business decisions, which lead to inefficient and ineffective operations, so that companies and organizations incur millions of euros' worth of loss or damage. Now that the economy is stagnating, many companies are seeking to save on costs and to enhance their returns. The general tenor of this article is: stop seeking and start up a data quality programme! The management and control of data quality will be one of the major themes in IT in the coming years.[1]

The article covers the subject of what 'data quality' actually is, and the possible reasons behind poor data quality. It goes on to deal with data quality assessment as a basis for an organization-wide data quality programme.

Data quality

Data quality can best be described as 'data that meets the criteria set by a company or organization'. What these criteria may be depends on that company or organization. Or, to be more specific: on the nature and size of the business operation. With this, it is clear that there is no universal framework of criteria that can be applied to every company. For one company it may be of the utmost importance to have data available timely, to respond to changing market trends for example, another organization may see no problem in receiving the data a little later as long as it is one hundred per cent accurate, such as for the specifications for raw materials and semi-manufactures in a production process. The various company processes may place other demands on the quality of the data. Where the purchase and logistic processes are geared to an optimum process, with specific demands being placed on supplier, contract and article data, a sales and marketing process will require more flexibility from the customer, order and delivery data.

Table 1 gives an overview of quality aspects.

In real-life practice, companies tend to take a more pragmatic approach and declare only several of these quality aspects to be applicable to the data in their environment. Sometimes, the data quality criteria will be related to rules and regulations. Within the financial sector, for example, the following quality criteria have been stated specifically for new sectorial guidelines for insurers (Solvency II): 'appropriate', 'complete' and 'accurate'. When it comes to the observance of criteria for information security, the applicable quality criteria are often related to 'confidentiality', 'integrity' and 'availability'. This article deals, later, with the issue of how companies can generate a framework of quality criteria by means of which they can assure and monitor their data quality.

Causes of poor data quality

Data is not static. Data is continually in motion. It is created, adjusted, exchanged, stored, destroyed – sometimes all of these within a few seconds. Every action contains the threat of a possible impact on the quality of the data. Everyone who works with data will have been surprised, perhaps on several occasions in his or her life, by the fact that data that was once regarded as being correct has suddenly obtained a value or form that does not correspond to reality, without he or she having had any influence on this alteration. There are many reasons for such metamorphosis. A first distinction can be made between active and passive data use. Active data use involves processes that are specifically oriented toward the data. In this context, we can mention data conversion, system consolidation, manual data entry, batch entry, real-time entry, data processing, data cleansing and data integration. Passive data use involves, among other things: changes in the use of data that have not led to any changes in the data itself, system upgrades, new forms of data use, loss of knowledge, processes that are automated. Both the active and the passive use of data by organizations lead to loss of data quality. See the example in the frame, for example. Therefore, we can conclude that data quality is not self-evident, and that data quality can be eroded by various influences. For this reason, organizations should actively guarantee data quality in their processes. Data quality assessments help to determine the actual data quality and its integrity in organizations.

> Due to an error in the Tax Department system, two life partners, one of whom had an income while the other did not, and who had received a general tax discount from the Tax Department for years, received a demand for repayment because they had received this sum illegally. The automatic link between the tax paid by the one partner and the tax discount of the other had mysteriously disappeared.[2]

Data quality assessment

The aim of a data quality assessment is to identify incorrect data, to estimate the impact on the business processes, and to implement corrective action.

Ideally, organizations have no errors in their databases. In the perfect world, we could imagine that we would be capable of researching all data elements and of determining whether they are 'right' or 'wrong'. In real-life practice, there are at least two factors that impede this. First, for each data element in the database you would have to determine what the real value of this element should be. You still need a reliable standard against which the perceived value can be tested. This source is not always available. And even if it is available, then a second challenge arises, namely, that the execution of the test requires so much (manual) effort that it scarcely pays to implement it. If we wish to carry out an assessment of the material master data, in order to discover whether or not the information on the suppliers of all our materials is correct, then we would have to write to all individual suppliers of every individual material (in the absence of

1. See also a recent study by the Hackett Group entitled 'The CFO's Agenda: Finance Top Issues in 2011', which states that 'data governance' is high on the agenda of the CFO.

2. See: http://www.elsevierfiscaal.nl/fiscaal-actueel/themas/navordering/blog/10/kan-de-belastingdienst-zich-zoveel-fouten-permitteren.

Category	Description
Accuracy	The degree to which the data tested is in agreement with a standard source, which is assumed to be correct. It is actually a count of the number of times an attribute was incorrect.
Integrity	The data is not inadvertently deleted, and changes made are applied. You can measure integrity by tracking the number of issues logged or records affected from a call center.
Secure	Users and systems that should not have access rights are properly controlled. You may measure this by logging incidents, tracking the number of individuals with write access, or measuring conformance to security standards by role.
Completeness	The degree to which all of the required data values appear in the data record or business object. You can measure this by tracking the percentage of records with one or more missing required attributes.
Validity	The degree to which the tested data conforms to the data validation requirements. This is a count of the number of rules in conformance.
Timeliness	The degree to which the data is provided at the time required or specified. You can track this by comparing planned versus actual availability.
Coverage	The degree to which the data sample accurately represents the whole population that is to be measured. This can be tested through statistics to determine if the sample population is a true representative sample.
No redundancy	The degree to which there are multiple copies of the exact same data across multiple databases. To measure this, you can track the number of copies of that record throughout your landscape. A more important measure, however, is to track the number of applications that create this data, Systems of Record, and Systems of Destination.
No duplicates	The degree to which there are duplicate records for the same item. You can track this by running a test to count the number of duplicate items with different item identifiers.
Relevance	The data should be tied to a business process, metric, or objective. This can be difficult to track, but you can track attributes in a master data record to determine what process or business document requires that attribute to determine relevance.
Accessibility	The ease of use for data consumers to access and consume the data. There are several methods to measure this, including determining access speed, points of access, or even search quality.
Availability	The degree to which data is available for access to the users. This can be measured by the times at which data is available to be used, planned versus actual.
Consistency	The ability to use or interpret the data accurately across multiple domains, for example, multiple product descriptions for the same product across applications. This can be measured by checking the same item ID across multiple systems or domains for inconsistencies or the same item with different identifiers across multiple domains or applications.

Table 1. **Data quality aspects** (source: [Huan99]).

an independent source) to check whether or not the material description, as recorded in our system, corresponds to that specified by the supplier. If we remember that some companies have thousands of materials in their systems, it will be clear that this task cannot be implemented in an efficient way. If we wanted to examine whether or not the addresses of our suppliers are correct and complete, we could contact all suppliers and also consult the data in the Chamber of Commerce files, if required. This would take a great deal of time.

Partial solutions are available for both challenges. The most important solution is the use of computers in the execution of data quality analyses. This enables us not only to investigate large quantities of data quickly, on the basis of quality criteria, but also to consult other (reliable) sources in an automated way, allowing us to determine the quality of our data elements. Examples of such sources include the previously mentioned Chamber of Commerce files, Dun & Bradstreet address files, GS1 DAS.[3] The application of automation and the consultation of independent sources help to enhance the quality of data. In this framework, sources that are known to be authentic sources provide the most added value. The Municipal Register ('Gemeentelijke Basis Registratie) is a good example here.

In the main, the data quality assessment takes place according to the five phases described in Table 2.

3. See: http://www.gs1.nl/support/gs1-das/.

Phase	Points of concern
Planning phase	Determining the scope, objective, establishing the engagement team, planning schedule, deliverables, method of data download, access to the system environment
Preparatory phase	Planning the actual data download, determining the data, classifying meta-data
Implementation phase	Business rules, data quality criteria and filters are determined and applied to the data sets
Refinement phase	Consultations with process owners about initial outcomes; refining and extending data quality criteria
Reporting phase	Issuing of final report, presentation, discussion of subsequent steps

Table 2. **The five phases of data quality research.**

Planning phase

Determining the scope of the data quality assessment is essential. The scope partly depends on the objective of the assessment. A rough distinction can be made between the goals presented in Figure 1.

Awareness

Many organizations do not know whether or not they have a data quality issue. In such organizations, there is no mention of an embedded data quality function and structured data analyses. Any analyses that have taken place have certainly occurred on an *ad hoc* basis. There is no structure for periodic and structured analysis of data quality, for reports on this quality, and for correction if necessary. The initiative to carry out a data quality assessment can be regarded as the beginning of an 'awareness phase'. The IT department or business operations may indicate that there are issues with the quality of the data. A data analysis is then seen as the instrument to obtain confirmation of the data quality issues, by means of which the management can subsequently be convinced that they ought to initiate a 'data management programme'.

An example of this process is the recently implemented data quality assessment for an international organization. The CIO of this organization acknowledged that the current data quality was indeed causing disruptions to commercial operations, but he did not know how serious 'the problem' was. He required facts to underpin the nature, size and necessity to initiate a data quality improvement programme. In the subsequent assessment, data on materials, customers and suppliers were studied with reference to certain quality aspects. The assessment led to a concrete step-by-step plan to enhance the quality of data, and initiatives have been started up for a data management programme.

(Master) data management programme

Organizations have set up a (master) data management programme. As a component of this programme, 'measurements' of the quality of master data have been taken throughout the entire duration of the programme. After a baseline measurement ('zero measurement') the programme continuously looked for progress and, if observed, measurement was made to assess the extent to which the quality of the master data had improved. For more information on the set-up and execution of the (master) data management programme. See also [Jonk11].

Migration to a new system

There is mention of a large-scale IT renovation project in which data from 'legacy systems' is transferred to a new integrated system environment. In that case, the data to be transferred must be harmonized in order to enable it to be used in the new environment. Data quality analyses are carried out to determine whether or not the legacy data meets the data quality criteria of the new system environment. By implementing quality assessments in an iterative way (i.e. assessment, cleansing of legacy data, assessment etc) an organization works toward a situation in which the data set to be migrated meets the quality criteria of the new environment (as much as possible).

The data quality analysis, whose origin lies in the wish to make the organization aware of the presence of data quality issues, appears at first sight to be an attractive form with regard to the scope, lead-time and therefore also the budget for the project. A meaningful data object can be chosen as a starting point for a quality analysis. Often the emphasis is placed upon an object that the sponsor suspects, albeit speculatively, of causing such quality issues. In close consultation with all parties involved, the scope of the analysis is then focused upon this object. Nowadays we see that organizations are primarily struggling with the task of keeping master data on materials up to date and consistent. It will therefore be no surprise that this data object will be selected in the awareness phase. The next step is that, if the budget and lead-time are limited, and outside-in view test can be executed. In this, the data

Figure 1. **Objectives of data quality assessment.**

The core of the data quality analysis lies in the identification, analysis and definition of good business rules

quality team will execute an automated analysis of the data set, without further examination of business operations and the applicable business rules, while making use of generic quality criteria that are based upon experiences in other organizations ('better practice' queries). This is indeed an attractive alternative from the standpoint of scope and budget. The effectiveness of this form of data quality analysis is limited, however, in view of the fact that the 'better practice' filters applied are largely very generic and can only produce relevant results if the organization is in the initial phase of data management. The quality aspect scrutinized in this way may include:

- attributes not filled in
- double names, material types, addresses
- incorrect names, addresses (with the use of a linked address file).

In addition, profiling techniques can provide insight into the structure of tables and data sets and into value patterns in attributes (minimum and maximum value, average, distribution of the occurrences in numbers and percentage). The effectiveness of such scans is also limited. In the absence of a concrete norm for data quality, it is impossible to establish whether the perceived values are 'right' or 'wrong'.

To enlarge the effectiveness of a scan with a limited but carefully selected scope, it is essential to have interaction between the data analysts and the organization. This interaction is directed toward revealing more and better business rules, which can subsequently be applied to the selected data set. With regard to costs, interaction with the organization has a strongly magnifying effect. Assessing and defining business rules are labor-intensive exploits. But if the awareness-heightening stage is restricted to one or a few data objects, the budget will be lower than would be the case if all relevant objects were to be included in the scope of the study. The argument for a restriction of the assessment budget tends to fade, however, if the choice is made in favor of orienting the assessment to a section of the object data, as when only 100 articles from a data set containing 50,000 articles are examined. At first sight, it would seem cheaper to do this than to investigate all materials. However, it should be kept in mind that defining the business rules is an activity independent of the number of records that are assessed within an object. It is very likely that the same business rules will be used for the data set of 50,000 records as for a section of this data set.

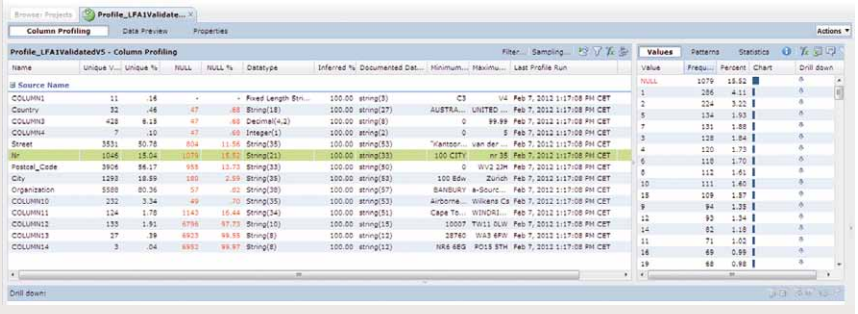
If the cause of a data quality assessment lies in an organization-wide master data management programme (MDM programme), the scope of the assessment will be directly related to the scope of this programme. For example, if the scope of the MDM programme is on the management of supplier, customer and material master data, then the scope of the data quality assessment will also include these objects. The goal of the data quality assessment is, as mentioned previously, to supply information about the actual data quality during the entire programme, in order to obtain insight into the extent to which the programme is effective. From a cost perspective the data quality assessment activities will form a substantial part of the MDM programme budget. In addition to the data quality assessment, during the programme, a data quality management organization will be set up.

In the case of IT-system integration, in which the aim of the data quality analysis is to determine the degree to which the data sets meet the quality criteria of the receiving systems, the scope is self-evidently related to the objects that are migrated.

Engagement team

The composition of the engagement team is also a component of the planning phase. Data quality analyses require collaboration between technical specialists and process specialists. The technical specialists possess knowledge and experience of table structures, of building and analyzing queries and data analysis techniques such as interpreting outcomes, and the recognition of patterns. The process specialists come from the business process and have a steering role there. Their share in the data analysis is directed, in close co-operation with the technical special-

In this example the LFA1 (supplier) table of a SAP system has been analyzed by means of profiling techniques. A profile has been drawn up from the 'number' field. What is conspicuous here is that, in more than fifteen per cent of all records, no value has been entered. This may be a trigger to discuss such occurrences with the organization.



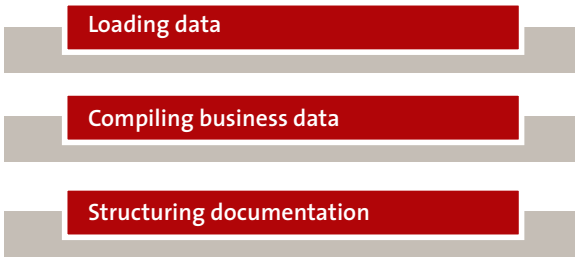


Figure 2. **Activities in the preparatory phase.**

ists, toward the identification, processing and testing of business rules that are used as quality criteria in the data quality analyses. The IT department is often approached to provide information on table structures, the working of the system and to realize the data download, complementing the technical specialists.

Preparation phase

In the preparation phase, various activities can be distinguished, as shown in Figure 2.

Loading data

The cyclical execution of data analyses, in which a data set is examined against a multitude of business rules, may have an influence on the performance of the system. If, in addition to this, there is mention of a migration programme in which source data is adjusted in accordance with quality criteria and is consistently implemented until the required quality has been achieved, it will be evident that the execution of data quality analyses requires an isolated environment. This is often referred to as a 'staging area'.[4] This can best be described as an isolated environment in which a copy of the production data is stored in order to test data quality and transformation activities. In the staging area, data can be adjusted, relations established, and analyses performed to one's heart's content. Specific software is needed to lay an audit trail in which all changes to the source data are recorded.

The download itself is often performed by the IT department, which possesses the required authority to access tables and to make them available for the data analysts.

Compiling business rules

The core of the data quality analysis lies in the identification, analysis and definition of good business rules. The better the business rules, the better the quality analyses and the better the insight into the real quality of the operational data that is used to make commercial decisions.

But the real question is: How do we acquire good business rules? It has already been stated that intensive interaction between data analysts and knowledge workers of the company forms the primary means to gain good insight into, and an intensification of, the business rules. Jointly examining the outcomes of 'data profiling' and analyzing patterns and trends leads to the identification of business rules. For example, analysis of the occurrences of the 'payment term' attribute (for which the general business rule is 'within 30 days') can lead to the insight that other payment terms are applied in certain other cases within the company. Seeking the causes ('root cause analysis') may lead to the identification of scenarios within which it is allowed to make use of a different payment term, such as: 'payment term within 60 days if the client yielded a turnover of x or higher last year', for example.

Intensive co-operation with knowledge workers is not always sufficient to reveal all business rules. There is not always sufficient knowledge of and experience with the business process available. The turnover of staff may cause a loss of much expertise. There are also other potential sources that the data analyst can make use of. In this context, one may consider technical documentation. Data rulebooks ('data dictionaries') and data models may also offer a helping hand. Good rulebooks contain a source of information about table structure, the attributes used, ownership, business rules and suchlike. Data management projects are also often aimed at updating or, as is frequently the case in our practice, realizing such rulebooks.

Data models describe the structure of data and particularly the relationship between the various objects and attributes. Just as is the case with rulebooks, data models are not always available; alternatively, they may be out of date. There are various kinds of data models. It is beyond the scope of this article to deal with them here. The list of references provides more information on this theme.

Structuring documentation

A substantial data quality assessment makes use of existing rulebooks and data models. But the assessment also produces much documentation that must be stored in a structured manner. In this context, we can think of analysis reports, mapping documentation, versions of updated rulebooks. It is important to create, at the start of the data quality assessment, a repository in which such documentation can be stored in an orderly and transparent way. This facilitates the retrieval of data and the underpinning of proposals to improve data quality.

Implementation phase

The implementation phase focuses on the identification, development and implementation of business rules. The concept of 'data profiling' has already been mentioned several times. Profiling can best be described as a collection of methods and techniques directed toward the analysis of data sets, with the aim of obtaining insight into the struc-

4. In a sense, the 'staging area' can be compared to the term 'DTAP', which is more common in IT circles. This DTAP represents the isolated environment that a newly developed system passes through before going live (Development, Test, Acceptance and Production).

Attribute profiling	Relationship profiling	State transition model profiling	Dependency profiling
investigates the value of individual attributes and provides statistical information, the frequency of certain values.	concerns the identification of keys in the data set to be analyzed; also counts the number of occasions that a certain relationship occurs in the data model.	techniques that are used to determine the number of statuses that an object/attribute can assume, and can also count the number of records with a certain status.	techniques that seek hidden relationships between attribute values.

Figure 3. **Sorts of data profiling.**

tures of that data set, and the dependences within and between data sets.

In relation to documentation, profiling enables us to see how data is truly structured. Documentation may be obsolete and the insight that architects or knowledge workers share with us may also be subject to erosion. A profile presents the current status and, as such, may be an important source to give direction to the data quality analysis and to the identification of business rules. The literature ([Mayd07]) recognizes the sorts of data profiling mentioned in Figure 3.

Business rules

In [Mayd07] the following kinds of business rules are distinguished:

- restriction in the value of attributes
- rules relating to relational integrity
- rules for historical data, including restrictions in time values
- rules for status values of objects and attributes
- general dependence rules.

Rules for the value of attributes

Depending on the significance of an attribute, a limit to the value that can be assigned to a field can be specified. For instance, it is impossible for the 'age' field to have a value higher than a certain number or lower than zero. Business rules that are directed toward the value of attributes aim at applying a restriction in the permitted values. This helps prevent incorrect values being entered, which enhances the quality of the data set in question. Defining a field as 'obligatory', for example, can prevent the occurrence of an empty field.

The article earlier provided an example of the outcome of a 'data profile' of supplier data. It was observed that the 'street and number' field contained no value in a number of cases. However, the 'supplier' record also contains a field called 'postbox number'. It may be that no 'street and number' has been entered because a 'postbox number' has been filled in. A business rule could then be that either 'street and number' is filled in, or 'postbox number', but it is not possible to leave both fields empty.

Rules relating to relational integrity

These business rules are based upon the starting point that entities (such as 'supplier', 'customer', 'product') in a relational database have a certain mutual relationship with the attributes of these entities. The most well-known example of this kind of business rule is the rule that attempts to guarantee the uniqueness of a record. The question is, for example how can we guarantee that a certain supplier only occurs one single time in a database. In the logical database, the so-called 'primary key' for the 'supplier' entity is the 'supplier's number'. In database design terms, this is the 'identifier' by means of which the uniqueness of the entity is guaranteed. However, the business rule that each supplier's number may only occur once in the data set does not guarantee the uniqueness of the supplier by definition. For this, there must be inspection of a combination of attributes. In this case, that could be the combination of 'supplier's number', 'name', 'address', and 'VAT number'. Various supplier's numbers, but the same name, address and VAT number indicate double entry. Searching through a data set for double supplier entries in which the combination of these attributes must be examined demands the presence of very flexible data analysis software in which so-called 'fuzzy logic' is applied. Here, on the basis of previously defined logic founded on perceived data, it is established whether or not certain data contains possible errors. (Example: J. Johnson, J Johnson and J. Jonson may possibly be the same person, certainly if they have one or more common attributes.)

Rules for historical data

Many data sets contain historical data: order history, overview of former suppliers' addresses, functions that members of staff have fulfilled, with the corresponding salaries. These are generally only of value if they are accompanied by a date to which the registered piece of data relates. With regard to certain personal data, this should actually be deleted if it has lost its value to the organization. The combination of the date and the nature of the item of data registered produces business rules. For instance, in a database with active materials, the date on which the price was determined will have to be recent.

Rules for status values

Objects in a company can undergo a transition and thus acquire a certain status. For example, an object 'insured'

The knowledge worker is ideally equipped to interpret and validate the outcomes of the analyses

with a life insurance company may have the following statuses: 'application', 'offer', 'offer assessment', 'examination', 'acceptation', 'active', 'terminated'. It will be clear that the nature of these statuses and the conditions under which a status may change can provide information for business rules (for example: an insured person can have only one status, for acceptation it is necessary that an examination take place, etc.) The entire configuration of the logical sequence of statuses, the number of statuses an object can have in a certain time, and the relationships that these statuses have with one another, may be very complex. It is often necessary to express these relationships in a model to determine the business rules.

Dependence rules

Dependences between attributes occur in business rules for historical data and business rules for status values, as touched upon previously. But these two forms of business rules do not completely cover all possible dependences between attributes. Supplementing these, there are also other dependences between attributes, which are also relevant to the quality of a data set we are investigating. One form of this type of business rule occurs when the same attribute appears in various databases. For instance, the address of a client may occur in both the CRM system and the invoicing system. The values of the 'address' attribute ought to be identical. This dependence can then be converted into a business rule.

Refinement phase

It is not easy to design good and adequate business rules. The application of better-practice rules is not without risk. They have not been specifically designed to comply with the business environment within which an assessment is being made. Although such better-practice rules may be applicable to a great many environments, they could lead to an erroneous outcome in certain situations. In this context, 'erroneous' means that they mistakenly lead to the identification of 'wrong' data, or they may not identify data that is indeed wrong. For example, payments are recorded as being double when two records show the same receiving-party bank account, order reference, and sum stated. However, regular payments for a subscription may then also be erroneously classified as double. Therefore the specific business rules must be taken into careful consideration.

The refinement phase is geared to sifting through initial results, in conjunction with knowledge workers, with the aim of refining the applied rules. Data analysts often lack the specific knowledge to be able to fully interpret profiling and other outcomes of analysis. Knowledge workers are involved in the everyday implementation of the process. They are involved in the realization of business transactions established in the system, as well as in the generation and alteration of master data. They are often also concerned with process optimization, where possibilities to enable the process to run faster and smoother are consistently sought at the interface of process and system. All these activities ensure that the knowledge worker has far-reaching insight into the way in which the process is structured and implemented. He or she is ideally equipped to interpret and validate the outcomes of the analyses, and to further refine the business rules.

Reporting phase

At the end of the refinement phase, an optimum set of business rules has been generated that can be applied to objects within the scope of the assessment. The application of these rules provides an optimum picture of the quality of the data studied, with clear identification of the 'errors' that are contained in the data set. A number of possibilities are now available:

- The management decides to extend the scope of the data quality analysis with relevant and critical objects that initially fell outside the scope of the analysis. The observed flaws in the data set are subsequently repaired.
- The outcomes are part of an organization-wide analysis of the quality of data management. The defects provoke the question as to how these could have arisen. The broad investigation can supply information on and perhaps construe that the governance processes could be improved, that the maintenance processes for master data and the relevant management measures must be improved, and that technical documentation is lacking. In that sense, the data quality assessment and the supplementary findings from the organization-wide assessment may form a trigger to initiate an organization-wide master data management programme.
- If the data quality analysis is a component of a system implementation, the outcomes will be used to cleanse the legacy data before it is migrated to the target environment.

Conclusion

Many organizations are not aware of the importance of data quality for the performance of business processes and for the provision of management information about the results of business operations. This article has provided an introduction to the theme of 'data quality' and the execution of data quality analysis. In view of the fact that business data forms the basis of decision making within an organization, it is important that the quality of the data is adequate and effective in supporting good decision making. The definition and application of the proper data quality rules occupies a central position in data quality assessment. The coming years will bring an increase in data analysts, data analysis software, and companies and organizations that will give substance to data quality management in a structural way.

Literature

[Huan99] Huang, Lee and Yang, *Quality Information and Knowledge*, Prentice Hall, 1999.

[Jonk11] R.A. Jonker et al., *Effective Master Data Management*, Compact, 2011/0.

[LeBl08] Andrew LeBlanc, *Enterprise Data Management with SAP Netweaver MDM*, Galileo Press / SAP press, 2008.

[Mayd07] Arkady Maydanchik, *Data Quality Assessment*, Technics Publications LLC, 2007.

About the author

R.A. Jonker is a partner at KPMG Management Consulting, IT Advisory, and Service Leader for Enterprise Data Management. He advises companies and organizations about setting and implementing organization-wide data management applications.