REPRODUCIBLE MEASUREMENT OF DATA FIELD QUALITY

(Practice-Oriented)

Marcus Gebauer, Peter Caspers
WestLB AG, Duesseldorf, Germany
marcus_gebauer@westlb.de, peter_caspers@westlb.de[1]

Niels Weigel
FUZZY! Informatik AG, Ludwigsburg, Germany
niels.weigel@fazi.com

Abstract: Define, quantify, assess and improve are the four cornerstones of a process for continuous improvement of information and data quality. We show how data quality can be measured in a reproducible and understandable way on the basis of defined business rules through the use of a data quality tool. As business rules directly reflect the requirements of the responsible specialists, they are in a position to verify whether the subsequent information meets their requirements.

Key Words: data analysis, data quality metrics, data migration, business rule, profiling, key rule, content rule, matching rule, compound keys, rule induction, patterns

INTRODUCTION

Information quality comes more and more into the focus of many companies - also in banks. The pressure increases due to the often complicated heterogeneous process and system landscape and due to external auditor's regulations. There is a long list of reasons of which only some shall be mentioned:

- Money Laundering
- Basel II and Operational Risk Management
- Sarbanes-Oxley-Act and Corporate Governance
- Multiple data sources and market data management

Quality is defined as the degree to which a product meets the requirements of clients [1]. This "fitness for use" definition is a very intuitive term for quality and is often reflected in statements of data users and suppliers such as:

- "Our data is 100% correct"
- "The data which we receive from system X are absolute junk"

However, data quality cannot be quantitative assessed on the basis of such statements. In most cases, neither the purpose for which data are used nor the underlying requirements to be met by the data are clearly defined. Clear definition is a prerequisite for the assessment of data quality. Requirements should be formulated so as to permit an objective and automatic verification of the ability of data to comply with such requirements. Also in the quantitative assessment, individual opinions play an important role. Whereas the supplier of data wishes to deliver a product which "complies with specifications", the users of the data wish their "expectations to be met" or even "exceeded" [2]. The gap between specifications and expectations often causes dissatisfaction on the client's side.

[1] Mr. Peter Casper can be reached at peter.caspers@ikb.de

Purpose

If data quality is supposed to be a genuine element of control, it must be defined in a reliable, understandable and reproducible manner. Data quality is an on going process. It is known as the Plan-Do-Check-Act (PDCA) cycle and has meanwhile become a reference [4], [5] for continuous improvement of quality. Below the adapted PDCA cycle supported by the data quality measurement tool (DQ measurement tool) is shown (figure 1).

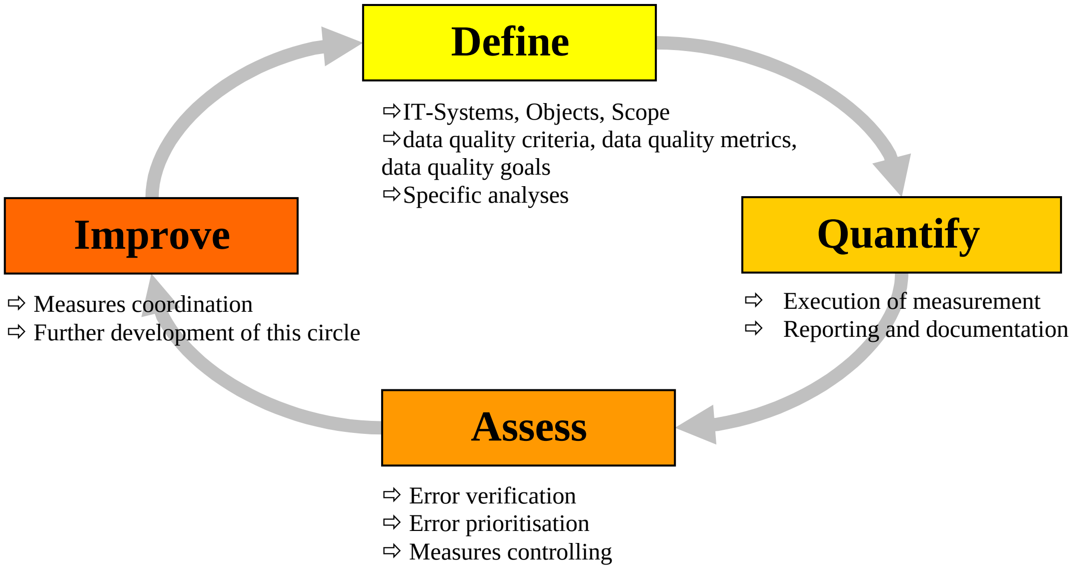

Define
⇨IT-Systems, Objects, Scope
⇨data quality criteria, data quality metrics, data quality goals
⇨Specific analyses
Quantify
⇨ Execution of measurement
⇨ Reporting and documentation
Assess
⇨ Error verification
⇨ Error prioritisation
⇨ Measures controlling
Improve
⇨ Measures coordination
⇨ Further development of this circle

Figure 1: Illustration of the adapted PDCA cycle

The implementation of this cycle is necessary in a rapidly changing environment in order to maintain or improve an achieved quality standard. Modifications caused by IT technology or organisational changes are to be considered equivalent.

In practice, the purpose of data processing and hence underlying requirements, resulting for instance from new legal framework conditions, new products or in-house requirements, is subject to change. A change of systems involving data migrations might impose requirements which are determined by the features of the new system. New requirements can also arise from a change in data-receiving systems.

Requirements

This section deals with the DQ measurement tool developed by the Bank, called The-DQ-Tool. The goal of The-DQ-Tool is to measure and to ensure data quality in the data systems on the level of data fields. As data quality has many categories and dimensions [3] the tool mainly acts on those correlated directly measurable on data fields. Currently, the focus is placed on master data systems which essentially contain static data, and on systems fed by master data systems or external references. The existing heterogeneous systematics and its changes are major challenges for the data storage systems and their quality. In order to meet these challenges, the following fundamental project type checks must be available:

Type a. Consistency check within a data storage system (intrinsic DQ)
Type b. Check of data storage systems against reference systems
Type c. Check against external reference systems
Type d. Check among data storage systems

The distinction between the four types is of a purely functional nature. Type a ensures that the data are consistent within a system. This requirement must be met by all data storage systems, even if it is not a leading system. Types b and c serve to assess data-receiving systems regarding the reliability of data suppliers, both for internal leading systems or external reference systems. Type d verifies the consistency of data kept redundant in the different systems. The differentiation and separation between reference and leading data storage systems must be made, as the reference data storage system might not be electronically or permanently available in the Bank.

The measurements carried out are part of the PDCA cycle (fig. 1). The-DQ-Tool accompanies the entire cycle and supports the specialist units in the individual process steps. The requirements formulated by the individual units for the same data storage system may lead to completely different judgements of the quality of the data. The resulting problem of data interpretation is caused by these heterogeneous contexts of data supplier and data receiver [5]. As a result of this strong heterogeneity, data quality goals, data quality features and data quality metrics are to be determined in accordance with the specific requirements. The standards to be met by the system are derived from the goals of data quality management and the requirements of the specialist departments of the Bank using such data. In order to guarantee permanent improvement and clear definition of data quality in the company, data quality management determines the framework conditions for technical implementation and implemented systematics for The-DQ-Tool:

- Connection of the heterogeneous system environment to The-DQ-Tool must be ensured
- Data quality measurement must be understandable and reproducible
- Rules fixed for measuring data quality must be suitable for repeated application
- Production of standardized reports must be guaranteed
- Data quality analysts must be guided by data field requirements

The tool meets the specific requirements of the different business units, which assess the quality of data storage systems on the basis of their business rules. The results of data quality measurement are also to be used to benchmark the business unit’s performance, to support migrations project in reaching their goals or to get an objective judgement for contract fulfilments.

- Business rules serve as basis for data quality measurement
- Standardized reports provide a basis for comparison of data quality levels
- Effective controlling of measures
- Awareness of the storage of redundant data
- Support for migration projects
- Definition of quality in Service Level Agreements

METHODS

Architecture: The-DQ-Tool was developed as a client-server application in order to comply with the requirements of a central rule repository which can be used by teams on a decentralized basis. The server component also includes a central pool linked to connected data bases which permits efficient use of the data base links by the clients.

Data access: Data access is exclusively made via the generic JDBC interface. This approach offers the advantage of making a large number of data accessible for The-DQ-Tool without requiring a differentiation of these sources within The-DQ-Tool. On the other hand, only the functionalities of data bases provided for in the JDBC interface can be used. Furthermore, some data bases do not completely implement the JDBC interface so that specific The-DQ-Tool functions might not be available. Moreover, The-DQ-Tool requires a specific partial quantity of the SQL 92 standard. This possibly leads to difficulties in the cooperation with individual data bases. The disadvantages are in part offset by import and copy functionalities, which are described in detail later.

Processing principles: The-DQ-Tool carries out all measurement and analysis functions by generating suitable SQL requests, which are processed via the connected data base. Then the results are returned to The-DQ-Tool for further processing, so that The-DQ-Tool mainly has a managing function with the workload being performed by the data base. This outsourcing of the effective processing is advantageous in that the performance of The-DQ-Tool essentially corresponds to the performance of the data base used, so that the performance can be scaled on the data base server.

Systematics: Figure 2 illustrates how The-DQ-Tool supports users in the individual steps of the PDCA cycle by analysing data. Within this process, the three major blocks “Preparation”, “Knowledge Management” and “Measures” can be identified. During Preparation, the systems are connected to The-DQ-Tool, the data being filtered and subjected to preprocessing. In the middle block, Knowledge Management, the most useful work for the specialist units is carried out. Data are analysed, measured on the basis of fixed business rules and the errors are validated. In the third step, measures are defined and controlled, if appropriate, in a further run of the PDCA cycle as to their efficiency.

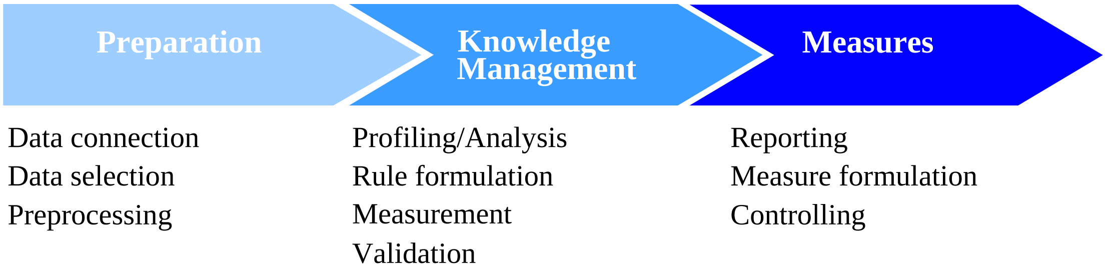


Figure 2: The-DQ-Tool supports the bank’s departments in their data quality analysis

From practical project experiences, we formulate detailed requirements for The-DQ-Tool for further clarifying of the framework requirements indicated. Where appropriate, we will show the implementation of requirements by screen shots.

Preparation

Data Connection

The-DQ-Tool permits the connection of data according to three principles:

- Direct access to data bases, with re-routing of copy access (e.g. for generating error lists), if appropriate, to a separate data base
- Copying data into a separate analysis data base
- Import of data from flat files

A special feature of the import from flat files is that The-DQ-Tool does not carry out any type inferences, i.e. all data are imported as text fields the breadth of which is determined by scanning across the entire flat file. Thus it is guaranteed that the data supplied in the flat file are imported to the analysis data base completely and in their original form. Converting of types, if necessary (e.g. date fields), can be carried out in a next step of data selection/pre-processing. This type of import, albeit simple, has proved to be extremely useful and advantageous in practice.

Data Selection/Preprocessing

Data selection and pre-processing is shown in The-DQ-Tool via own views which substantially correspond to SQL views except that abstracts are made from actual tables. Hence, it is not necessary to adjust the defined views to the actual tables for repeated measurements under different configurations. These views have proved to be an adequate instrument for pre-processing raw data for later analysis.

As a special feature The-DQ-Tool offers the possibility to convert views prior to analysis on an automated basis into physical tables (“flattening”), also into a separate analysis data base, if required. In the case of complex views, this results in considerable increases in performance for measurements and analyses.

Knowledge Management

Profiling / Analysis

The knowledge about data and data models and their business usage vanishes over time. As data quality can be only measured with respect to the business where the data is used, business itself determines the rules to judge the quality of data. So, one major step before determining data quality is to review data and data models. In [6] this is described as KDD – Knowledge Discovery in Databases. The-DQ-Tool contains a selected quantity of profiling and analysis functions which support the retrieval of business rules. In several projects we found this the most important step to the business experts. Often they find long forgotten business rules during this profiling step. On the other side once the business rules are fixed and stored in The-DQ-Tool every new employee can get familiar with the rules. Additionally the rules can now be challenged and validated when either the IT system or the business changes.

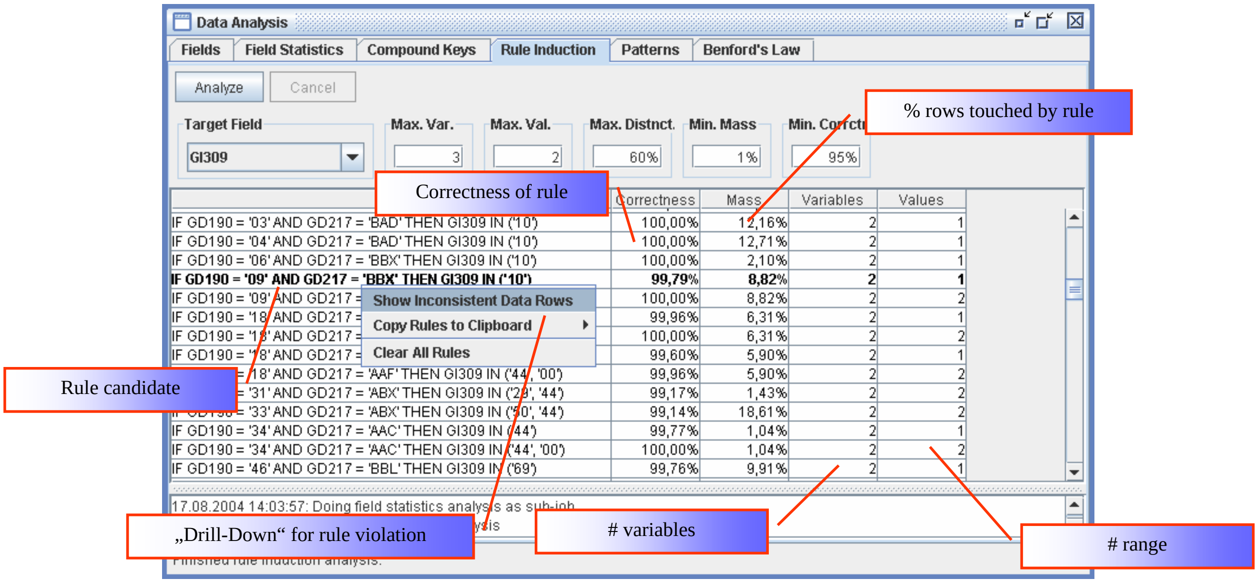
The included analysis functions are:

- Field statistics (descriptive statistics): degree of filling, number of different pattern, minimum, maximum, average, one- and more-dimensional distributions.
- Compound keys, i.e. retrieval of (compound) keys in tables. For the validation of key candidates it is possible to show data lines which violate the key feature.
- Rule induction, i.e. the retrieval of rules of the “IF Field*1*=x1 AND … and Field*n*=xn THEN Target_Field IN (Value*1*, … , Value *m*)“ type, e.g.

  IF X=25 AND Y=’KFO’ THEN Z IN (0, 1, 2, 8)

  For validation of these rules, it is possible to show data lines which are inconsistent with the rule.
- Pattern analysis, i.e. the retrieval of patterns in field contents of text fields, e.g. n2pn2pn4 for a date format. For validation of the patterns, the quantity of patterns which are not among the most k-frequent patterns can be viewed.
- Benford’s law analysis, i.e. the statistical verification of the Benford distribution of initial digits to numeric columns of a table.
- Matching analysis, i.e. the determination of multiplicity and quantity ratios and sub-tables which describe how two given tables correspond to each other via a pre-fixed condition.

Where possible, interfaces to the rule repository are available permitting direct transfer of automatically retrieved rules to the repository. There, the rules can be improved manually and refined to business rules. Figure 3 shows the rule induction within the profiling step in the analysis of the master data system for security paper information used in the Bank. After selection of the data fields, the context of which is to be analysed (here, the effect of using fields GD190 and GD217 on field GI309) the left column shows the logical context of the rules proposed. The “Correctness” column shows the degree of correctness of the rules and the “Mass” column shows the percentage share of data lines to which the rule is applicable. “Variables” and “Values” show the number of variables used and the volume of the result field.



Target Field: GI309
Max. Var.: 3
Max. Val.: 2
Max. Distnct.: 60%
Min. Mass: 1%
Min. Corrctn: 95%

	Correctness	Mass	Variables	Values
IF GD190 = '03' AND GD217 = 'BAD' THEN GI309 IN ('10')	100,00%	12,16%	2	1
IF GD190 = '04' AND GD217 = 'BAD' THEN GI309 IN ('10')	100,00%	12,71%	2	1
IF GD190 = '06' AND GD217 = 'BBX' THEN GI309 IN ('10')	100,00%	2,10%	2	1
IF GD190 = '09' AND GD217 = 'BBX' THEN GI309 IN ('10')	**99,79%**	**8,82%**	**2**	**1**
IF GD190 = '09' AND GD217 =	100,00%	8,82%	2	2
IF GD190 = '18' AND GD217 =	99,96%	6,31%	2	1
IF GD190 = '18' AND GD217 =	100,00%	6,31%	2	2
IF GD190 = '18' AND GD217 =	99,60%	5,90%	2	1
= '18' AND GD217 = 'AAF' THEN GI309 IN ('44', '00')	99,96%	5,90%	2	2
= '31' AND GD217 = 'ABX' THEN GI309 IN ('29', '44')	99,17%	1,43%	2	2
= '33' AND GD217 = 'ABX' THEN GI309 IN ('50', '44')	99,14%	18,61%	2	2
IF GD190 = '34' AND GD217 = 'AAC' THEN GI309 IN ('44')	99,77%	1,04%	2	1
IF GD190 = '34' AND GD217 = 'AAC' THEN GI309 IN ('44', '00')	100,00%	1,04%	2	2
IF GD190 = '46' AND GD217 = 'BBL' THEN GI309 IN ('69')	99,76%	9,91%	2	1

```
17.08.2004 14:03:57: Doing field statistics analysis as sub-job
ysis
Finished rule induction analysis.
```

Figure 3: Illustration of a rule induction result for some static data for securities

Measurement

In The-DQ-Tool, three types of rules are made available for verification of different aspects of data quality. The rules can be used individually or jointly in measurement projects. Typically, different measurement project types involve the use of different rule types:

Type a verification of consistency within a data table: in many cases, use of content rules only

Type b/c/d: Comparison between several data tables: this involves frequently the use of

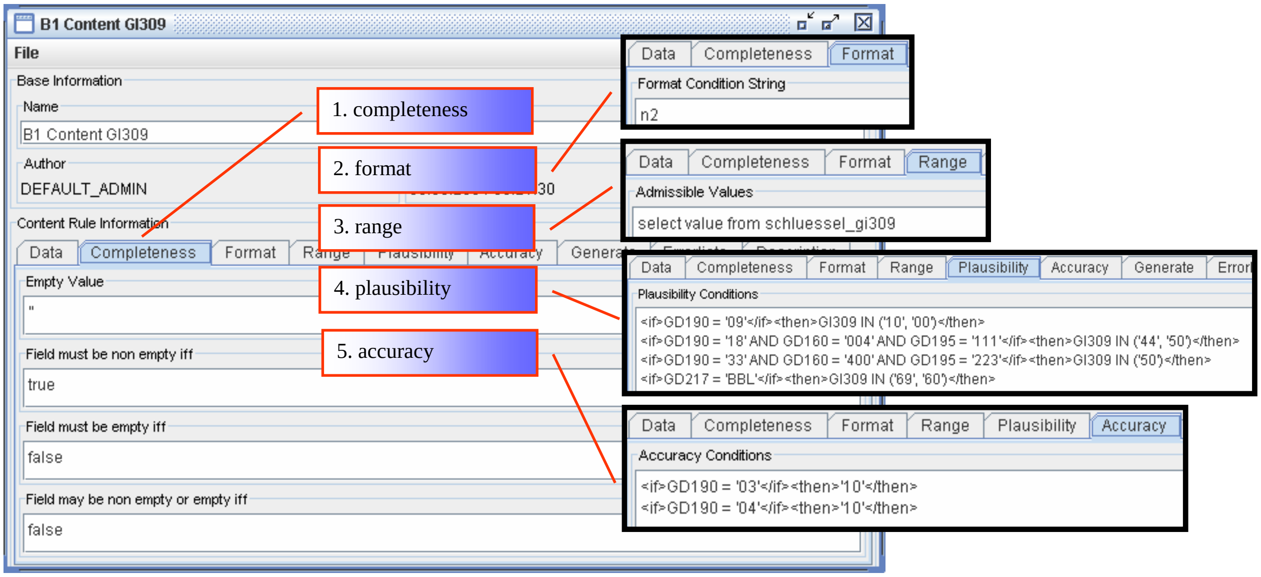
- key rules for verification of the fields connecting the tables, followed by
- matching rules for verifying the extent to which the tables are compatible with each other, followed by
- content rules for verification of the consistency of contents of data records of two compatible tables.

The formulation of business rules is always made by business experts who assess the data used by them against their own business background. Rules can be measured against data under different configurations[2]. A server based central rule repository is implemented which can be updated by the user. All rules generate numeric measurements results which can be combined to quality metrics via self-defined arithmetic formulas. Furthermore, all rules generate error lists, if desired, which contain exactly that part of the analysed data table which violates a given rule. Both outputs – metrics and error lists – are essential for the validation of rules and the evaluation of measurements results.

The rules can be deposited hierarchically in a tree structure, with one rule able to appear in more than one branch of the tree. Thus, it is possible to define different views of the rules, e.g. according to priority, organisational units concerned, data-using systems concerned or functional aspects. The tree structure is reflected in the reporting so that the measurement results can be evaluated according to the views defined.

[2] Configuration means the entire information which characterizes the tables used for measurement.

- **Key rules** are used for verification of key and general multiplicity features of individual or compound fields. The field combination to be verified and the multiplicities permitted for identical combinations are to be specified.
- **Matching rules** are used for verifying the linkage of two tables. The condition according to which the tables are related to each other and the multiplicities permitted are to be specified.
- **Content rules** are used for verifying data contents (figure 4). Data contents are verified step by step, with restrictions applying to contents becoming sharper with each step. These measurement steps are only a part of the 16 dimensions of data quality, as described in [3] for instance. As soon as a data record is recognized as an error, it does not run in the next steps. This precludes multiple counting of an error. The individual steps are:
    - **Completeness**: Verification of correct filling, correct non-filling and definition of 'Empty-Value' of the data field
    - **Format**: Verification of the content of a field as to consistency with a predefined pattern
    - **Range**: Verification of the content of a field as to affiliation to a predefined value area. It is also possible to store the predefined values in a data base table.
    - **Plausibility**: Verification of one- or more-dimensional relations among fields of a data record
    - **Accuracy:** Verification of one- or more-dimensional functional dependencies among the fields of a data record



Name: B1 Content GI309

Author: DEFAULT_ADMIN

Empty Value: ''

Field must be non empty iff: true

Field must be empty iff: false

Field may be non empty or empty iff: false

Format Condition String: n2

Admissible Values: select value from schluessel_gi309

Plausibility Conditions

```
<if>GD190 = '09'</if><then>GI309 IN ('10', '00')</then>
<if>GD190 = '18' AND GD160 = '004' AND GD195 = '111'</if><then>GI309 IN ('44', '50')</then>
<if>GD190 = '33' AND GD160 = '400' AND GD195 = '223'</if><then>GI309 IN ('50')</then>
<if>GD217 = 'BBL'</if><then>GI309 IN ('69', '60')</then>
```

Accuracy Conditions

```
<if>GD190 = '03'</if><then>'10'</then>
<if>GD190 = '04'</if><then>'10'</then>
```

Figure 4: Illustration of the steps of content rules in the analysis of security paper data

Validation / Evaluation

Validation of rules and evaluation of the related measurement results is always carried out in practice on the basis of error lists which are produced by the business rule repository. As these error lists are sometimes very large, it is necessary to provide support for structured evaluation of the lists in the evaluation of potential errors. The-DQ-Tool offers the possibility

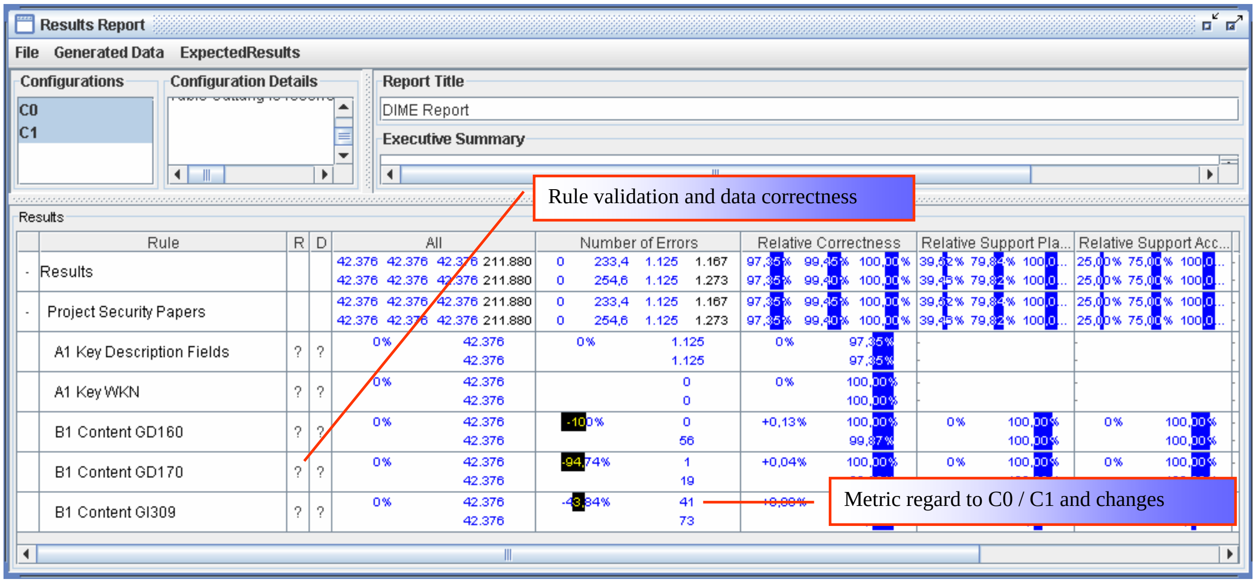
- to sort the lists upwards and downwards regarding any of the fields
- to carry out n-dimensional frequency counts including the illustration of marginal distributions
- to determine pivot fields which define a filter on the list
- to define filters on the basis of the frequency counts

These functions proved to be very useful in practice for an efficient evaluation of large error lists as they offer a practical possibility to sub-divide error lists according to functional aspects in clusters of “similar errors”. The results of functional validation can be allocated to the rules in the form of structured notes and are thus available in The-DQ-Tool to all users directly from the tool.

Measures

Reporting

Result reporting contains the numeric measurement results regarding the defined metrics for the measurement project and the possibility to branch off to the related error lists directly from the measurement results (figure 5). The results are arranged according to the same hierarchy as defined in the rule repository. Measurement results are aggregated (minimum, maximum, average, total) at each hierarchical level, thus permitting an evaluation of the results in line with the views defined in the repository and at different granularity levels. Reporting can be made for more than one configuration, so that measurement results and error lists of several measurements (figure 5: C0 and C1) can be directly compared. This ensures that all information required for the reproduction of the results is stored.



DIME Report

Rule validation and data correctness

	Rule	R	D	All	Number of Errors	Relative Correctness	Relative Support Pla...	Relative Support Acc...
-	Results			42.376 42.376 42.376 211.880 42.376 42.376 42.376 211.880	0 233,4 1.125 1.167 0 254,6 1.125 1.273	97,35% 99,45% 100,00% 97,35% 99,40% 100,00%	39,52% 79,84% 100,0... 39,45% 79,82% 100,0...	25,00% 75,00% 100,0... 25,00% 75,00% 100,0...
-	Project Security Papers			42.376 42.376 42.376 211.880 42.376 42.376 42.376 211.880	0 233,4 1.125 1.167 0 254,6 1.125 1.273	97,35% 99,45% 100,00% 97,35% 99,40% 100,00%	39,52% 79,84% 100,0... 39,45% 79,82% 100,0...	25,00% 75,00% 100,0... 25,00% 75,00% 100,0...
	A1 Key Description Fields	?	?	0% 42.376 42.376	0% 1.125 1.125	0% 97,35% 97,35%	-	-
	A1 Key WKN	?	?	0% 42.376 42.376	0 0	0% 100,00% 100,00%	-	-
	B1 Content GD160	?	?	0% 42.376 42.376	-100% 0 56	+0,13% 100,00% 99,87%	0% 100,00% 100,00%	0% 100,00% 100,00%
	B1 Content GD170	?	?	0% 42.376 42.376	-94,74% 1 19	+0,04% 100,00%	0% 100,00%	0% 100,00%
	B1 Content GI309	?	?	0% 42.376 42.376	-43,84% 41 73	+0,00%		

Metric regard to C0 / C1 and changes

Figure 5: Illustration of the measurement results of the rules applied on two different data records C0 and C1

The measurement report can be made available online within the tool (figure 5), in XML format for further use in any other tools, in CSV format for import to Excel, Access etc., in PDF format and in HTML format.

Measure Formulating / Measure Controlling

Within the framework of the already described note function in the validation and evaluation of rules and measurement results, it is also possible to define measures for eliminating errors found during measurement (content of the measure, addressee, date of implementation). Simple time management permits the identification of overdue measures. Repeated measurements are a simple means of carrying out quality control of the implemented measures, and the result of the quality control can be documented.

EXPERIENCES AND DISCUSSION

The data quality measurement tool has been developed directly from practical experiences gained in data quality projects. The applicability shown also benefits other data quality projects. During the last years, The-DQ-Tool has been used within various data storage systems and for different projects. Hence, this tool is already used for regularly determining the quality of master data systems for customer and security information – both currently host-based systems. Furthermore, the Bank also uses The-DQ-Tool for SAP applications, e.g. in the SAP HR, SAP BP and SAP FI modules. The-DQ-Tool was also successfully used within a migration project, in which deposit master static data had to be adapted and migrated.

During the examination of the quality of data of the security paper systems used in the Bank, it was possible to reduce expenses for data updating considerably. The analysis focused on special data fields, regularly updated by the Bank, which were reportedly used for the processing of turnovers in securities. From a total of 128 so-called "internal security data fields" 73 were no longer used. From the remaining fields, 49 could be automatically derived from externally supplied security paper fields, which were stored in the same system. Hence, at the outcome of the analysis, only 7 fields remained which must be updated and maintained manually. In addition to the time and cost-savings achieved, the responsible specialist units now benefit from being able to concentrate on the updating of fields required. This should have a positive impact on the data quality of the system in the long run.

The support of Knowledge Management by The-DQ-Tool proved to be particularly useful. In old established systems, knowledge of the data structure often exists to a limited extent only. The analysis functions offered help in obtaining a complete overview of data dependencies and structures. The knowledge of the specialist units can be verified against the background of contexts found and used for defining rules for measuring data quality.

LIMITATIONS

The-DQ-Tool is a highly flexible analysis tool the use of which is not confined to financial data. Possibilities of analysises are not limited by static or quasi-static data, except that specialist information is required for the definition of business rules.

Realization of data transfers exclusively via the generic JDBC interface and limitations resulting from the SQL92 standard used can have a restricting effect. As a result of the outsourcing of actual processing, the performance of The-DQ-Tool corresponds mainly to the performance of the data base management system used, so that the scaling of performance depends on the (known) data base side only. So, if a data base is sufficiently performing for large data sets the performance of The-DQ-Tool is of the same level.

For reasons of restricting administration of access rights, the necessary transfer of data into a separate The-DQ-Tool data base can in practice turn out to be a restrictive element for performance.

As the bank implemented its needs in measuring data quality in The-DQ-Tool we nevertheless analysed 22 other commercial tools available. Most of them are used as so called ETL tools from which The-DQ-Tool is different from its usage. The-DQ-Tool was built to ensure data quality in all day's work. In comparison to that ETL tools often come into the game when data sets are migrated. To our opinion this is often too late and the data quality rules are implemented in a way that they are only used once in the migration step. The combination of data profiling, data analysis, rule induction and possible data correction is quite unique among all benchmarked tools. These advantages are a result of many projects run in the bank in the years 1999 to 2001 and The-DQ-Tool therefore meets the bank's requirements of course. Other possible functionalities as address and name checks are not implemented, because they were not needed. This will be subject to further developments as discussed in the following chapter.

CONCLUSION AND OUTLOOK

By means of The-DQ-Tool, the user is directly led through essential parts of the PDCA-cycle. From the profiling of unknown data storages to the measurement of data quality at a field level and the measure controlling involved, the tool supports the user in his or her daily work and in migration projects. The success of the application is partly due to the fact that the functionalities of The-DQ-Tool have been developed from work in practical projects which were then consolidated and implemented in a uniform tool. The high variability in terms of functional systematics and technical embeddedness in the systems environment provides clear benefits for the Bank.

The limiting factors described are compensated by the very flexible use both with regard to the technical environment and the specific contents. This is shown in the daily practical use, in which the applicability of the data quality measurement tool shows its practical use in the heterogeneous system environment is often the major criterion in the data quality analysis rather than speed.

The-DQ-Tool is a module for a holistic approach to data quality, permitting the quality of the data storage systems to become a genuine management tool. It provides the possibility to secure Service Level Agreements, to verify project results in terms of quality, to provide transparency of the degree of data quality and offers many other possibilities.

Currently, the definition and formulation of business rules require a certain technical understanding. In future development stages, emphasis will be placed on a simplified definition of rules supported by assistants, already available in some functions. Furthermore, an intuitive interface will be made available to the unskilled user. As an additional analysis function an enhancement of the original data with so called token types provide the possibility to classify the content of a data field using special reference tables like ‘Mobile area code’. With this classification every token of the data field is checked against the reference tables and tagged with the corresponding token type (e.g {MAC} for the mobile area code). With this information you can easily detect additional or wrong information in a data field. Another extension will be the integration of a complete action management tool. With this extension every necessary step inside The-DQ-Tool can be defined, assigned to team members, tracked and monitored. The defined actions are stored in the centralized database of The-DQ-Tool. Every data quality project member can review the actual status of the data quality project with different reports.

REFERENCES

[1] Juran, J.M.; Gryna, F.M.J.; Bingham, R.S.: *Quality Control Handbook* (5th. edition), McGraw-Hill Book Co, New York, 1974.

[2] Reeves, C.A.; Bednar, D.E.: *Defining Quality: alternatives and implications*, AMR 19, 3 (1994), 419-445.

[3] Pipino, L.L.; Lee, Y.W.; Wang, R.Y.: *Data Quality Assessment*, Communications of the ACM, April 2002/Vol. 45.

[4] English, L.P.: *Improving Data Warehouse and Business Information Quality*, John Wiley & Sons, Inc. New York, 1999.

[5] Wang, R.Y.; Ziad, M.; Lee, Y.W.: *Data Quality* – Chapter 6, Kluwer Academic Publishers, Norwell, 2001.

[6] Roiger, R.J.; Geatz M.W.: *Data Mining*, Addison Wesley, 2003.