A Novel Approach to the Evaluation and Improvement of Data Quality in the Financial Sector

Karel Dejaeger[*,a], Bart Hamers[b], Jonas Poelmans[a], Bart Baesens[a,c]

[a]*Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*
[b]*Dexia S.A., Brussels, Belgium*
[c]*School of Management, University of Southampton, Highfield Southampton, SO17 1BJ, United Kingdom*

Abstract

Until recently, a universal approach for analyzing the quality of data generated by business processes has been missing. In this paper, a structured approach to data quality management is presented and applied to the credit rating process of a company in the financial sector. Starting from a comprehensive data quality definition, a structured questionnaire is composed. It is used for guided interviews to distill the key elements in the credit rating process from a data quality perspective. Business Process Modeling Notation (BPMN) is used to visualize the process, to identify where data elements enter the process, and to trace the various data outputs of the process. The visual process representation allows to identify error prone elements in the process and thus enabling a more focussed data quality management. It was found that data quality issues appear manifold in larger data driven organizations in the financial sector. The presented methodology provides a cross-departmental overview of the data flow and manipulations from source to end usage.

Key words: Data quality, Financial institutions, Credit rating, Data quality axes, BPMN

1. Introduction

In modern companies, more and more data is generated on a daily basis. This situation allows for a more quantitative use of data [18]. However, the results of a quantitative analysis are strongly dependent on the data quality and in case of insufficient data quality such an analysis may fail to provide the desired results and may even lead to incorrect conclusions being drawn. This is particularly true for the financial sector, where e.g. crucial decisions are made concerning the estimation of another party's credit rating. This credit rating, often expressed as an alpha–numeric code (e.g. AAA to D), reflects the risk the other party will experience payment problems within the next 12 months. An incorrect assessment of the credit rating will lead to a wrong estimation of the portfolio risk. Under Basel II regulation [28], financial institutions can calculate credit risk by using the ratings published by international rating agencies as Moody's and Standard & Poor's. In addition to this basic approach, the Basel II regulation allows for internal calculation of the credit ratings [26]. This is termed 'Advanced Internal Rating-Based' (AIRB) approach and it enables financial institutions to construct their own credit risk management models which provide more flexibility and the incorporation of context specific details. In the near future, a new regulation will be introduced, Basel III. Although details are not yet officially communicated, it is known that this regulation will require even more transparency into the risk positions taken by financial institutions. On the level of quantitative modeling, more attention is put on reducing the model risk and measurement error for all crucial risk parameters. This should be achieved by using more long term data horizons for modeling and model calibration. Also the attention to the stress testing by both the regulators and the bank management will result in an even more quantitative driven risk management. Essential in all these evolutions is the role of accurate and reliable data. To monitor both the internal and external information and data flow, strict and stringent data quality programs have to be put in place. Special attention

[*]Corresponding author. Tel. +32 16 32 68 87; Fax +32 16 32 66 24

Email addresses: Karel.Dejaeger@econ.kuleuven.be (Karel Dejaeger), Bart.Hamers@dexia.com (Bart Hamers), Jonas.Poelmans@econ.kuleuven.be (Jonas Poelmans), Bart.Baesens@econ.kuleuven.be (Bart Baesens)

to data quality will directly result into good and reliable risk management decisions based on both historical facts, statistical predictions and business knowledge. Data quality can be considered to be of high importance both to credit model construction and to risk reporting.

The remainder of this paper is structured as follows. In Section 2, the issues of data quality for the specific situation of financial institutions are explained. In Section 3, the different axes to which data quality was evaluated are discussed. In Section 4 the methodology applied for this study is explained. Section 5 summarizes the results of the analysis. Section 6 gives an overview of some shortcomings of the applied techniques. Section 8 concludes the paper.

2. Literature overview

Today's economy is characterized by a tendency towards a more quantitative use of data enforced by international regulations. Especially in the financial sector the quantitative use of data is highly regulated. For instance, European insurance companies must comply with the 'Solvency II' directive, requiring the harmonization of the capital requirements across the different member states of the EU [13] and thus forcing a common measurement framework for assets and liabilities. Analogous, methods to deal with anti money laundering are harmonized by the 3^{th} 'Anti-Money Laundering directive' [12]. The European Union also introduced various legislation concerning marketing in the financial sector [11]. Another well known international regulation is Basel II which, among other elements, determines how financial institutions are obliged to calculate exposed credit risk and the associated financial buffer [28]. To meet the requirements prescribed in these and other regulations, data quality in the financial sector is of the utmost importance. The process under consideration in this case study is the estimation of credit ratings which are used in calculating the financial buffers and is subjected to the Basel II-regulation which demands complete transparency and traceability of data. The elaboration and validation of the used credit models are required under the Basel II regulation on a yearly basis [28]. In [10], an overview of the backtesting process and background information is presented.

In the literature, the concept of data quality (sometimes also referred to as 'Information Quality') is envisioned as multi dimensional [5]. Table 1 provides an overview of the data quality aspects considered by different studies. In this table, both theoretic and practical oriented studies are considered.

A generic definition of data quality is 'fitness of use' meaning data should be readily usable by the end user [3]. According to this definition, the precise aspects of data quality under consideration are domain dependent [9]. Therefore, in this paper a comprehensive data quality definition is presented oriented to the financial background of the case study. The costs connected to poor data quality in the financial sector are of considerable importance. For instance, one study estimated these costs at 25 billion dollars in 2008 [22].

2.1. Credit risk modeling context

The analysis described in this paper took place in a financial credit risk modeling context. Within this context, the main goal is to determine the 'regulatory capital'. This is the monetary buffer financial institutions need to maintain to compensate for potential losses at the 99.9% confidence level. This regulatory capital is calculated as a percentage of the Risk Weighted Assets (RWA).

$$\text{regulatory capital} = 0.08 \times \text{RWA}$$

The RWA itself is a function of Loss Given Default (LGD), Exposure At Default (EaD), and Probability of Default (PD).

$$\text{RWA} = 12.5 \times \text{EaD} \times K(\text{LGD}, \text{PD})$$

with $K(.,.)$ called the capital requirement function. These and other calculations are part of Pillar I of the Basel II accord introduced in 2006 [28].

3. Data Quality: the different axes

In the literature, five recurring data quality dimensions can be found: accuracy, comprehensibility, consistency, completeness, and time (cfr. Table 1). A number of further refinements were made to these dimensions in line with information provided by the domain experts. Within the specific financial background of the case study, it was found that the time dimension is of high importance as credit ratings should be calculated using up to date information and are only valid for a certain period. Three possible sources of time-related data quality problems are identified (volatility, timeliness, and currency). Additionally, a security dimension was identified, as stringent privacy and security requirements are imposed on data within this domain. It can be seen from Table 1 that some domain specific data quality dimensions were not considered in this study; e.g. usefulness and accessibility. Fig. 1 depicts the different data quality axes.

Authors	Year	Data Quality aspects	Acc.	Compr.	Cons.	Compl.	Sec.	Time
D. P. Ballou and H. L. Pazer [5]	1985	Accuracy, completeness, consistency, and timeliness	✓		✓	✓		✓
K. C. Laudron [24]	1986	Ambiguity, completeness and inaccuracy	✓	✓		✓		
Y. Wand and R. Y. Wang [30]	1996	Correctness, completeness, unambiguity, and meaningfulness	✓	✓		✓		
P. Cykana, A. Paul and M. Stern [14]	1996	Accuracy, completeness, consistency, timeliness, uniqueness, and validity	✓		✓	✓		✓
R. Y. Wang and D. Strong [31]	1996	4 categories of data aspects: intrinsic, contextual, representational, and accessibility	✓	✓	✓	✓	✓	✓
E. Gardyn [17]	1997	Accessibility, completeness, consistency, correctness, and currency	✓		✓	✓		✓
R. Kovac, Y. W. Lee and L. L. Pipino [23]	1997	Accuracy, reliability, and timeliness	✓					✓
M. Helfert and C. Herrmann [19]	2002	Interpretability, plausibility, timeliness, and usefulness		✓				✓
P. Missier, G. Lalk, V. Verykios	2003	Accuracy, consistency, completeness, and precision	✓		✓	✓		
F. Grillo, T. Korusso and P. Angeletti [27]		Correctness and currency	✓					✓
C. Batini and M. Scannapieco [8]	2006	Accuracy, completeness, currency, consistency, timeliness, and volatility	✓		✓	✓		✓
B. Baesens [2]	2008	Accuracy, distortion, completeness, timeliness, and definition	✓	✓		✓		✓

Table 1: Overview of the concept of data quality

The exact definition of the data quality dimensions is presented and demonstrated within a credit risk modeling context (cfr. Section 2.1).

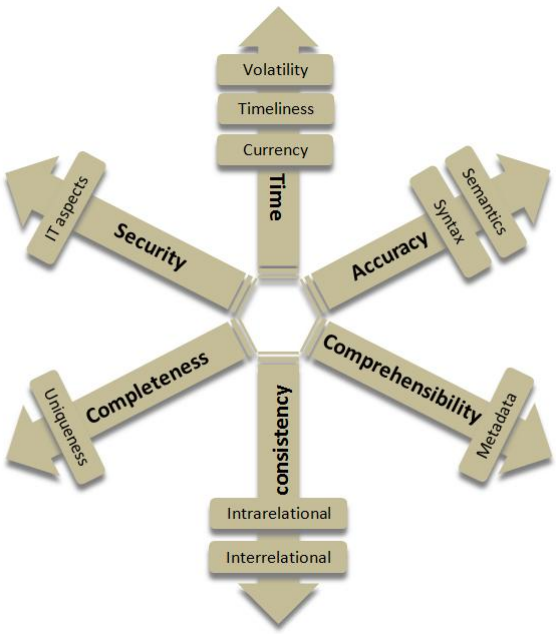

Figure 1: The data quality axes

3.1. Accuracy

The dimension 'accuracy' is defined as the degree to which a representation α' offers a correct value of a real-life counterpart α. A distinction can be made between syntactic and semantic accuracy. The domain D is the collection of all syntactic correct elements needed to describe the real-life counterparts.

Syntactic accuracy represents the approximation of a value α to the elements of the corresponding domain D.

For instance, 'Citibank N.A.' is a correct name while 'Citibank' is incorrect as it is unclear whether this refers to the head office or a local branch of Citibank.

Semantic accuracy describes the approximation of a value α' to the true value α.

E.g. both AA- and AAneg ratings are possible but have a different meaning. AAneg is an AA class company with a negative outlook while AA- is a rating in between A+ and AA. Failing to recognize the correct one of the two will lead to a faulty PD to be used to calculate the RWA and thus an incorrect regulatory capital.

3.2. Comprehensibility

This dimension refers to whether the end-user can fully understand the data. An optimal comprehensibility can be obtained by establishing extensive data definitions in the metadata, by creating conformity with standardized data-exchange formats and finally by ensuring the use of a clear and consistent syntaxis [21].

In order to correctly calculate the regulatory capital, a number of inputs are required (refer to Section 2.1). The credit analyst will need to understand the subtle differences in the various inputs. E.g. $LGD_{Secured}$ versus $LGD_{Unsecured}$; in case of $LGD_{Secured}$ recoveries on the collateral values are taken into account in the recovery computation while for $LGD_{Unsecured}$ this is not the case.

3.3. Consistency

A data set can be judged to be consistent when the constraints within each observation are met. Based on the nature of the constraints, a distinction can be made between interrelational and intrarelational consistency.

Interrelational consistency deals with rules established for all the records within the data set.

For instance, financial institutions typically use interrelational rules related to turnover to determine the type of company involved, e.g.

IF Turnover ≥ 1.000.000 $ THEN Company = 'Corp' ELSE Company = 'MidCorp'

Failing to adhere to this rule will give way to different PD ratings and LGDs used in risk calculations.

Intrarelational consistency verifies whether rules which are applicable within one record are being respected (e.g. whether a particular value is situated between the preset boundaries).

Typically, major banks consist of different branches and companies can be client of different branches at the same time. Suppose a client is in default at one branch, he should also be flagged as in default at other branches otherwise the risk for future draw dawns will increase. This potentially has a large impact on the EaD linked to the defaulted counterpart.

3.4. Completeness

Completeness can be defined as to what extent there are no missing values. A missing value can be causal or not causal. A causal missing value is allowed and it implies the presence of a specific and accepted reason for the missing value. For instance, the field containing the tav-number will remain empty if the client is a private person.

Uniqueness can be viewed as a supplement to completeness by checking the presence of doubles in the data set.

Suppose the loan to a company is booked twice, the bank will calculate the RWA twice and thus will also book the regulatory capital twice. On a full bank portfolio, such data errors could unnecessarily consume a significant part of the capital buffer.

3.5. Time

The dimension 'time' refers to time-related aspects of data quality and embraces three elements.

Currency: this sub dimension concerns the immediate updating when a change occurs in the real-life counterpart α.

Typically, banks will update the RWA of their investments on a monthly basis. Suppose at the 31^{th} of the month the RWA is calculated but a loan granted to a client on the 30^{th} is not included in the calculations. The real-life change is not propagated thus resulting in a currency problem.

Volatility: this item describes how frequent data change in time.

The changes in the exposure on the full bank portfolio can be high mainly caused by the typical high exposure volatility. This is driven by the typical short term interbank lending behavior. More frequent updating of the EaD estimations and corresponding capital buffer could be beneficent from risk perspective.

Timeliness: this represents how recent the data are in relation to their purpose. This item was added because having recent data at your disposal doesn't necessarily guarantee that they are available at the moment one needs them.

This issue occurs in the case of untimely re-rated counterparts. Normally, counterparts are rated on a yearly basis but it can happen that a new rating is unavailable and an older rating must be used which might not reflect the current risks associated with that counterpart.

3.6. Security

This dimension is of paramount importance for financial institutions due to privacy and safety regulations. Distinction can be made between *IT-elements* (e.g. passwords) and *human aspects* such as separation of function.

The privileges of credit analysts are typically not clearly defined. For instance, by which type of credit analyst should a central bank be rated?

4. Methodology

In order to correctly assess the data quality throughout a process, it is important to correctly map the dataflows and communication channels at the beginning of the assessment [7], to facilitate the detection of error prone process steps. A process can be decomposed into several smaller steps by adopting an ETL (Extract Transfer Load) approach in which the process steps are

identified where data is created or extracted from another source, where data is manipulated or transferred and finally where data is stored or loaded into another database. We considered these parts of the process to be the most critical from a data quality point of view as it are these parts of a process where typically most data errors are created. This generic view on the data flow of processes is not limited to processes in a financial context. The subprocesses identified using this approach, together with the communication channels between them, can be efficiently represented using BPMN (Business Process Modeling Notation). BPMN is a formal graphical representation format [15] specifically devised to model processes in which the participating parties are represented by horizontal layers (referred to as 'swim lanes'). As it is the business that mainly possesses the information required to model the process and identify the most critical process steps, in depth contacts with the business side are required (e.g. in this case the credit analysts, database managers and quantitative analysts). A simplified BPMN representation of the credit rating process is given in Fig. 2. More information regarding the BPMN standard can be found in [32].

To better guide these contacts and to readily identify data quality problems, we suggest to use a targeted questionnaire, in line with [1, 25, 29]. This original questionnaire consists of 65 questions and is publicly available [1]. The questionnaire is also designed based on the ETL approach and consists of five parts (or 'dimensions'): a collection, an analysis and a warehousing dimension, supplemented by a 'human' and a 'goal' dimension. In the *collection* dimension, questions concerning the origin of the data are posed. E.g.:

- Collection-Comprehensibility:
  To what extent is the source data obtained in a standardized format ?
- Collection-Consistency:
  To what extent is the procedure to deal with inconsistent data standardized ?

The analysis dimension focusses on data analysis or transformations. Sample questions are:

- Analysis-Comprehensibility
  Is the way of calculating the statistics clear to the analysts ?
- Analysis-Time
  What is the delay between the collection and the analysis of the data ?

The warehousing dimension investigates how to send the transformed or analyzed data to the next step in the chain. Some of the questions concerning the collection dimension can be repeated in this dimension such as questions related to the exchange format of the data. Examples are the following:

- Warehousing-Consistency:
  Are there interface business rules that check the imputed data with predefined data definitions ?
- Warehousing-Comprehensibility:
  Is an inventory maintained of all the data stored and how comprehensive and complete is this inventory ?

Additionally, a human dimension and a goal dimension are added. The first is concerned with aspects of training and communication of the data users while the latter is concerned with the goal of the process (e.g. whether end-users understand the goal of the actions).

Within each of the five dimensions, the questions are structured according to the definition of data quality (i.e. structured alongside the previously explained six axes) where applicable. This allows to highlight particular aspects of data quality within each dimension. Hereto, a score can be attributed to the questions, and the results for each dimension can be visualized using radarplots by applying simple descriptive statistics on these scores (summation and average). The same uniform set of questions were used to assess the data quality across the different process steps.

5. Results

5.1. Process description

Using a graphical BPMN representation in combination with a questionnaire specifically targeted on the assessment of data quality can be applied to different kinds of processes. In this particular case, this approach was used to evaluate the rating process of a large financial institution in Belgium (Europe). A high level overview of the rating process using BPMN is provided in Fig. 2. Following our methodology, a more detailed process description was also created, which provided a clear overview of the role of the different parties involved. Also which specific data is shared using the different information channels was reported. As this process model contained sensitive information, we are unable to include this detailed process model in this paper. The

[1] The questionnaire can be found at http://dataminingapps.com/staff/karel.php

graphical process representation allows to easily define the origin of the data and can be used to trace problems with certain data flows.

As the BPMN representation does not require prerequisite knowledge to interpret, this output can be readily used in communication with management and other stakeholders. Data quality problems can be indicated in the graphical representation using e.g. traffic signs associated to the different parts of the process.

5.2. Semi-quantitative presentation

The structuring of the questions in the questionnaire alongside the five dimensions and arranging the different questions for each dimension according to the six data quality axes allows to identify data quality issues at different granularity levels for each subprocess. This can be done by looking at the scores of the (sub)dimensions thus allowing detection of error prone elements. Fig. 3 shows a fictitious example of a radarplot representing the five dimensions of the questionnaire; the scores for each dimension are an aggregation of the scores of the questions for the dimension. A low score indicates a potential problem. Radarplots can also be generated for each of the dimensions separately, Fig. 3, bottom. These plots reveal more details concerning the different aspects of each dimension. The dimension 'collection' for instance contains seven subaspects. Not receiving answers to the questions connected with a particular item leads to a zero-score as in our example is the case for consistency. In this example, we can also see a low score for 'uniqueness' (collection dimension). This might indicate a problem with the appearance of doubles in this step of the process. The dimension 'analysis' contains three sub-aspects. Such high scores indicate that there are no major problems concerning analysis. The dimension 'warehousing' has five sub-aspects. The aspect safety has a high priority for financial institutions. In this example the questions related to timeliness were not answered.

5.3. Rating process analysis

By applying our methodology to the rating process, we detected a number of data quality related issues that impacted the data quality experienced at the end of the process (the backtesting or validation of the credit models).

This backtesting requires a number of different inputs collected from the rest of the process: a document containing data concerning trade partners who have severe payment problems (the 'default list'), quantitative information from the rating analyst and additional data relating to the trade partner which is extracted from a database. Therefore, data quality issues experienced earlier can have a serious impact on this sub process. A number of key persons in the process chain were interviewed (a total of six persons), and radar plots were generated for each interview. The analysis indicated for instance that the default list was constructed with a higher focus on business continuity with respect to client relations instead of the strict adherence to the risk rules (e.g. there may be a perfectly good business reason to grant a trade partner more time). This resulted in multiple revisions of the default list and thus gave way to a timeliness problem in the data quality. Also was found that the credit analyst and other parts in the process are required to input some data manually into the system and also used manually kept files thus potentially giving way to accuracy related data quality problems. Some of the reports received are also difficult to understand due to the lack of additional explanations. These and other issues resulted in slow downs of the backtesting or even into restarts and thus in the inability of the backtesting team to meet certain deadlines.

Another issue uncovered was the existence of double entries for the same trade partner in the databases. This can have far-reaching consequences (e.g. setting aside risk-capital twice for only one loan thus affecting the profit made on the loan) for the company. This problem can be labeled under the aspect uniqueness.

6. Discussion

The presented methodology can be used as a framework for testing various data quality problems. In large, data driven organizations like those in the financial sector, data quality issues appear to be manifold. Tackling these problems requires a unified methodology that is generic enough to be applicable to different tasks, while being specific enough to capture and identify the real cause of the data quality problem. Our methodology provides a global overview of the data flow and manipulations from source to end usage. The results of our assessment have proven to be useful cross-departmental as the findings are not limited to one database, one (risk) reporting task or department.

The analysis of the credit rating process revealed some discrepancy between the problems experienced at the end of the process (the backtesting) and the results of the data quality assessment. In general, the results of the separate sub-processes appeared to be fairly good as reflected in medium to high scores on the radarplots. The problems experienced during backtesting could not

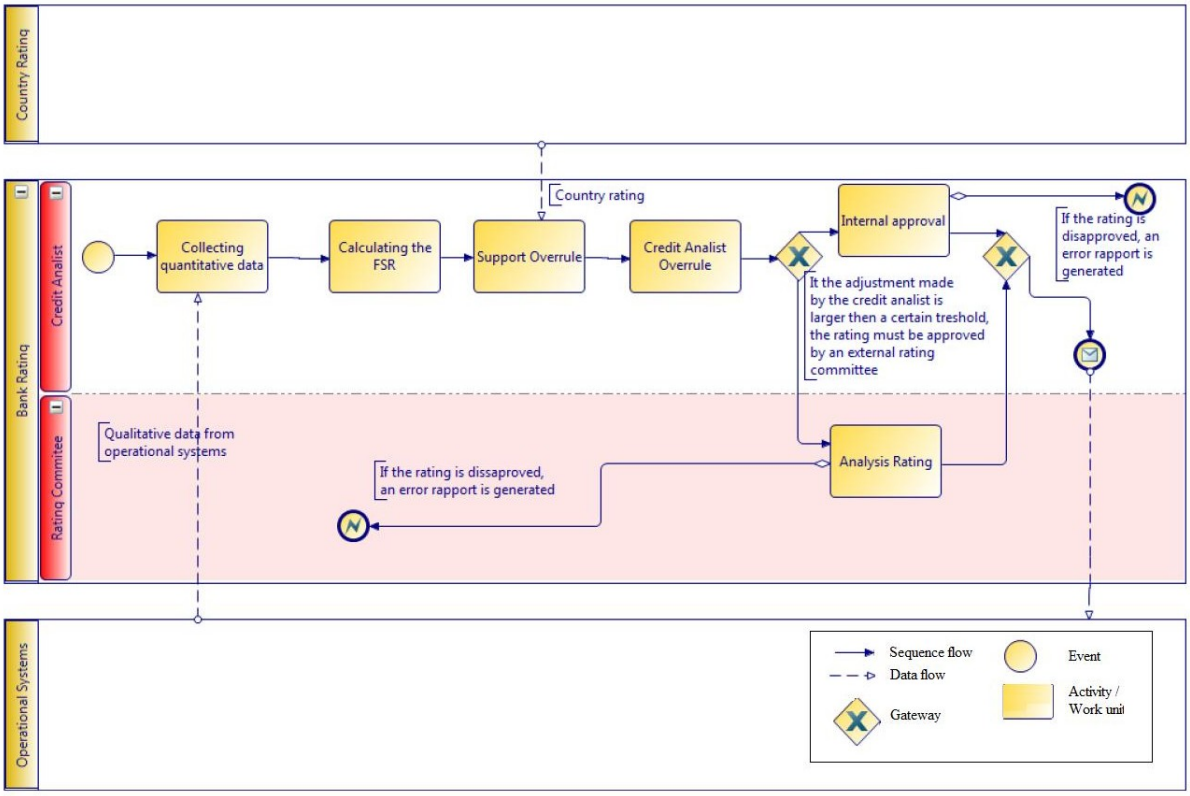

Figure 2: High level BPMN representation of the rating process for banks

be fully understood, however we found some small aspects which could be improved. (cfr. Section 5.3). The additive effect of these elements seems to be underestimated thereby confronting the data users at the end of the chain with poor-quality data.

We defined data quality as comprising six axes, however our experience learns that some axes are difficult to quantify. More specific, respondents found questions related to security and, to a lesser extent, comprehensibility difficult to answer. The first can be explained by the fact that security measures were managed centrally. Business users are often unaware of specific security measures in place. The latter dimension included questions relating to metadata and data ownership. Often, data descriptions were created in an ad-how manner while data ownership was unclear. Not surprisingly, axes (questions) relating to the data itself such as accuracy or completeness were evaluated as more clear by the respondent.

BPMN was found to have some weaknesses; modeling the different sub processes and the data-flows between them showed BPMN to be deficient in some respects. For example, BPMN does not allow to indicate the exact exchange formats used in data flows (e.g. specific excel formats). This is due to BPMN's design for modeling and optimizing processes which focusses less on mapping data-flows.

Furthermore, in our experience, people tend to downgrade own shortcomings and idealize their own part of the process. Their focus is often limited to their job-territory, this is called 'silo-thinking'. Therefore, we believe the questionnaire should be used by an experienced interviewer already familiar with the process as a whole.

Data quality is an important topic in the financial sector as the consequences of poor data quality are high [16, 22]. However, it should be noted that perfect data quality is often undesirable in an operational environment as the associated costs would be prohibitively high [20], thus a reasonable level of data quality should be the focus of a data quality program.

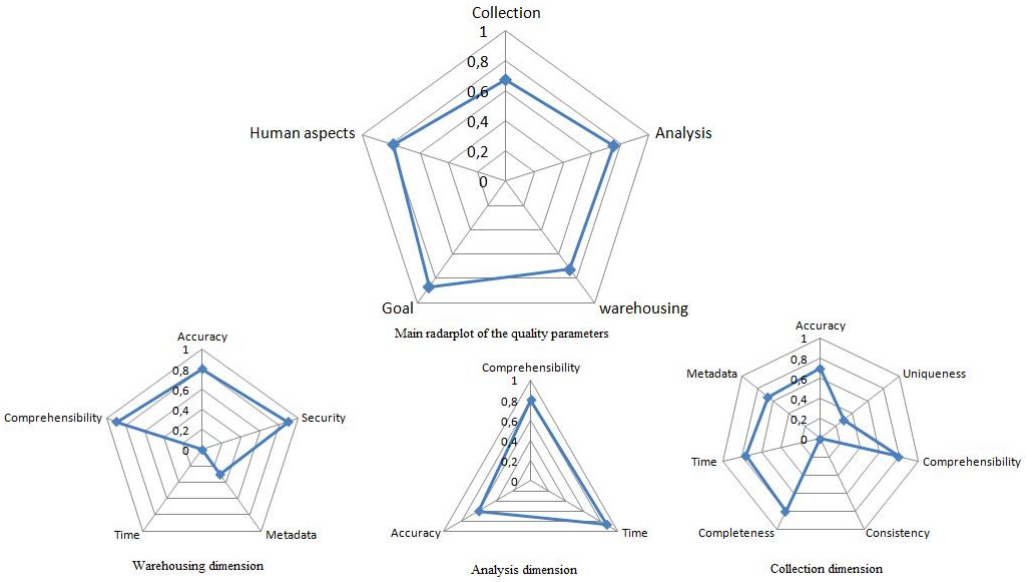

Figure 3: Illustrative radarplots

7. Limitations and future research

Limitations

The methodology used in this study is only applied to one specific process within a financial context. While we believe it is applicable to other processes in other domains, the validity of this claim remains an open question. Related to this observation, the questionnaire used in this study needs to be validated by using it in other contexts than the rating process to which it was applied.

Future research

Data quality is a process that requires on-going attention, as some data quality problems are always to be expected in an operational environment [4, 6]. Therefore, it is necessary to trace the evolution of data quality in a quantitative way to complement the approach adopted in this paper. This requires a more operational definition for each of the six data quality axes defined. While this is straightforward for some axes (e.g. completeness can be measured by the number of missing values), this remains an open question for a number of other dimensions. Currently, a measurement scheme is being implemented within the financial institution. The axes measured are accuracy, consistency, completeness, and time (volatility), using a control chart approach. A control chart will plot a metric versus time, indicating when abnormalities in data quality are experienced and is considered a promising avenue for future research.

8. Conclusion

Data quality is of great importance to the financial sector and is preferably analyzed taking into account its multidimensional nature. In this case study, the rating process of a financial company was investigated using a uniform questionnaire which was used throughout the whole process. In this questionnaire, an ETL (Extract Transfer Load) approach was adopted as data quality issues arise at these points in the process where data is collected, manipulated or stored. The process was analyzed in a semi-quantitative way and represented using BPMN. It was found that overall, data quality was high but the additive effect of some smaller issues resulted in poor data quality at the end of the process thereby illustrating the importance of taking the whole process into account.

One may conclude that a data quality solution requires more than only a set of business rules or a data profiling and cleansing tool. Instead, a process (or company) wide data governance program is needed in which the data quality problems are resolved often using tailored solutions. These solutions may consist of technical elements (e.g. a meta data and master data management) but the human factor in each step of the process

is at least equally important. Also communication and commitment by the management are considered to be crucial elements.

Acknowledgements

This research was supported by the Odysseus program (Flemish Government, FWO) under grant G.0915.09.

Jonas Poelmans is aspirant of the FWO.

References

[1] AHIMA. *Practice Brief: A checklist to Assess Data Quality Management Efforts*, 1998.

[2] B. Baesens. It's the data, you stupid. *Data News*, 2008.

[3] D. Ballou, S. Madnick, and R. Wang. Assuring information quality. *Journal of Management Information Systems*, 20(3):9–11, 2003.

[4] D. Ballou and G. Tayi. Enhancing data quality in datawarehouse environments. *Communications of the ACM*, 42(1):73–78, 1999.

[5] D. P. Ballou and H. L. Pazer. Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2):150–162, 1985.

[6] D. P. Ballou and H. L. Pazer. Cost/quality tradeoffs for control procedures in information systems. *Journal of Management Science*, 15(6):509–521, 1987.

[7] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, 41(3):Article 16, 52 pages, 2009.

[8] C. Batini and M. Scannapieco. *Data Quality: concepts, methodologies and techniques*. Springer, 2006.

[9] L. Berti-Equille. Un état de l'art sur la qualit des données. *Ingénierie des systèmes d'information*, 9(5):117–143, 2004.

[10] G. Castermans, D. Martens, T. Van Gestel, B. Hamers, and B. Baesens. An overview and framework for PD backtesting and benchmarking. *Journal of the Operational research society*, 61(3):359–373, 2010.

[11] European Commission. *Distance Marketing of Financial Services Directive*, directive 2002/65/ec edition, May 2002.

[12] European Commission. *3th Anti-Money Laundering Directive*, directive 2005/60/ec edition, September 2005.

[13] European Commission. *Solvency II directive*. European Union, 2009.

[14] P. Cykana, A. Paul, and M. Stern. DoD guidelines on data quality management. In *Proceedings of the Conference on Information Management*, pages 154–171, Cambridge, 1996.

[15] R. Dijkman, M. Dumas, and C. Ouyang. Semantics and analysis of business process models in BPMN. *Information and Software Technology*, 50:1281–1294, 2008.

[16] L. English. *Improving Data Warehouse and Business Information Quality*. Wiley & Sons, 1999.

[17] E. Gardyn. A data quality handbook for a data warehouse. In *Proceedings of the Conference on Information Management*, pages 267–290, Cambridge, 1997.

[18] J. Grey and A. Szalay. Where the rubber meets the sky: Bridging the gap between databases and science. Technical report, Microsoft Research, October 2004.

[19] M. Helfert and C. Herrmann. Proactive data quality management for data warehouse systems. In *Proc. of the Int. Workshop on Design and Management of Data Warehouses*, Toronto, Canada, 2002.

[20] J. M. Juran and F. M. Gryna. *Quality Planning and Analysis, 2nd ed.* McGraw Hill, New York, 1980.

[21] W. Kim. On metadata management technology: Status and issues. *Journal of Object Technology*, 4(2):41–47, 2005.

[22] G. Kopp. Data governance: Banks bid for organic growth. Technical report, TowerGroup, 2006.

[23] R. Kovac, Y. W. Lee, and L. L. Pipino. Total data quality management: the case of IRI. In *Proceedings of the Conference on Information Management*, pages 63–79, Cambridge, 1997.

[24] K. C. Laudon. Data quality and due process in large interorganizational record systems. *Communications of the ACM*, 29(1):4–11, 1986.

[25] Y. W. Lee, D. M. Strong, B. K. Khan, and R. Y. Wang. AIMQ: a methodology for information quality assessment. *Information & Management*, 40:133–146, 2002.

[26] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466–1476, 2007.

[27] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso, and P. Angeletti. Improving Data Quality in Practice: A Case Study in the Italian Public Administration. *Distributed and Parallel Databases*, 13:135–160, 2003.

[28] Basel Commitee on Banking Supervision. International convergence of capital measurement and capital standards. Technical report, Bank for International Settlements, 2006.

[29] L. Pipino, Y. Lee, and R. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[30] Y. Wand and R. Y. Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[31] R. Y. Wang and D. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, 1996.

[32] S. White. Introduction to BPMN. *BP Trends*, 2004.