Universiteit Leiden

ICT in Business

Data Quality Management

A Solvency II Perspective

Name : S.S. Altinay Soyer
Student-no : s1077260

Date : 27 March 2013

1st supervisor : Peter van der Putten
2nd supervisor: Emiel Caron

MASTER'S THESIS

Submitted to:
Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Table of Contents

Table of Contents 2

Abstract 4

Acknowledgement........ 4

Chapter 1. Introduction........ 5

1.1. Problem Statement 6

1.2. Research Objectives and Contribution........ 6

1.3. Research Methodology 7

1.4. Document Organization 9

Chapter 2. A New Regulation for the European Insurance Companies: *Solvency II* 10

2.1. Solvency II Background........ 10

2.2. Regulatory Framework and Data Quality........ 11

2.2.1. Solvency II and Data Quality Assessment........ 13

2.2.2. Requirements on Data Quality Management 14

Chapter 3. Data Quality Concepts 15

3.1. Data Quality........ 15

3.2. Data Quality Assessment........ 17

3.3. Data Quality Management Methodologies........ 20

3.3.1. Total Data Quality Management........ 22

3.3.2. AIMQ 23

3.3.3. Total Information Quality Management 24

3.3.4. Complete Data Quality Management 25

3.3.5. ORME-DQ 25

Chapter 4. A Data Quality Management System for Solvency II........ 26

4.1. Requirements Engineering 26

4.2. Analysis of DQ Requirements of Solvency II........ 27

4.2.1. Functional Requirements 29

4.2.2. Non-Functional Requirements 31

4.3. Specifications of a Solvency II DQM System........ 32

4.3.1. Data Collection 32

4.3.2. Data Quality Assessment and Improvement........ 35

4.3.3. Data Quality Monitoring ... 39
4.3.4. Data Deficiency Management ... 39
4.3.5. Data Quality Governance ... 40
Chapter 5. Analysis of Data Quality Management Methodologies ... 43
5.1. Comparison of Methodologies ... 43
5.2. Proposed Extensions for Methodologies ... 46
Chapter 6. Field Study at a Dutch Insurance Company ... 48
6.1. Background of the Solvency II Project ... 49
6.1.1. Data Management Team ... 50
6.1.2. Internal Roles and Responsibilities ... 50
6.1.3. Data Quality Assessment ... 51
6.2. A Practical Case ... 52
6.2.1. Data Quality Assessment ... 52
6.2.2. Data Quality Monitoring ... 57
6.3. Recommendations for the Insurance Company ... 59
Chapter 7. Conclusions ... 63
7.1. A DQM System for Solvency II ... 63
7.2. Overview of DQ Concepts and Methodologies ... 64
7.3. Analysis of DQM Methodologies ... 64
7.4. The Practical Case ... 64
7.5. Limitations of the Study and Future Work ... 65
Bibliography ... 66
Appendices ... 72
Appendix I. Interviews ... 72
Appendix II. Mind Map ... 73
Appendix III. Sample Data Dictionary ... 74
Appendix IV. Use Case ... 75
Appendix V. Cost Matrix ... 76
Appendix VI. Impact Classifications ... 77
Appendix VII. R Commands ... 78
Appendix VIII. Excel Functions for the Practical Case ... 79
Appendix IX. Dimension Measurement Techniques ... 80
List of Figures ... 81
List of Tables ... 82

Abstract

Solvency II is a new regulation for insurance companies that operate in Europe. The regulation aims to provide protection to policy holders establishing strict capital requirements and risk management standards for insurance companies. Implementing a proper Data Management structure plays a key role in the regulation as high quality data increases reliability of the required calculations. However, a recent review by the Financial Services Authority[1] showed that insurance companies still have a long way to go in establishing a Solvency II compliant Data Management structure. Since the directive does not attempt to standardize Data Quality Management (DQM), this study aims to provide guidance to insurance companies and other similar enterprises in implementing a suitable DQM structure. First, we analyze the directive's requirements on DQM utilizing Requirements Analysis techniques. Second, we transform these requirements to system specifications of intended DQM system. Finally, we explain the implementation tasks required for each system specification. In addition, we provide a review of available DQM methodologies in the literature to understand whether they could be adopted by insurance companies. The review results in several proposed extensions to the available methodologies in order to achieve Solvency II compliance. As a part of the thesis, a field study is conducted in an insurance company in connection with the company's Solvency II project. Through the field study, practical information on utilizing DQM concepts is obtained as well as information on insurance business and its data sensitivity.

Acknowledgement

I would like to express my sincere gratitude to my advisors Emiel Caron and Peter van den Putten for their continuous support and insightful comments during this study: Peter van der Putten's guidance helped me to find out my way within a large topic and remained focus; Emiel Caron motivated me to complete this study with his result oriented, constructive approach.

Besides my advisors, I would like to thank Ben Berndt and Peter Vlam for proving the field study opportunity. The practical part would not have been possible without their ongoing support and trust. Thanks to Daniel Ritsma for his valuable comments on finding the right direction during the field study, and thanks to Theodoor Rutgers for introducing me to the Data Management group and their activities.

Last but not the least, special thanks; to my husband, who always inspires me, for his patience and encouragement during this long journey; to my children for their unconditional love; to my parents for always being there for me.

[1] The Financial Services Authority (FSA) is an independent non-governmental body, given statutory powers by the Financial Services and Markets Act 2000 - www.fsa.gov.uk

Chapter 1. Introduction

The Solvency II directive is introduced by the European Union to increase the protection of insurance policy holders across the EU. The directive applies to all insurance companies that operate in the EU and it enables a better coverage of all the risk run by an insurance company. The directive introduces new risk management standards for companies in order to guarantee that they can survive during difficult periods such as floods, storms. According to the new rules, insurers are required to hold a certain amount of capital against risks they are exposed to. Whereas current regulative requirements are mainly based on historical data, the new regulation requires consideration of future developments that might affect insurer's financial position [1].

Successful data management establishes the basis of sound risk and capital management for companies. Without having reliable data sources, companies cannot assess their actual status effectively. As a result they cannot represent their actual risk exposure level to the regulatory authorities. In that sense, quality of data used in financial calculations is very critical as it increases the reliability of the results. Consequently, the Solvency II regulation includes specific data quality requirements in order to emphasize the importance of the issue and to set the standards for insurance companies on data management: "*Member States shall ensure that insurance and reinsurance undertakings have internal processes and procedures in place to ensure the appropriateness, completeness and accuracy of the data used in the calculation of their technical provisions* [2]."

But how can companies increase the quality of data they collect and work with? And how can they maintain a high level of quality on a continuous basis? Since data quality is in the center of Solvency II, what are the data quality specific requirements of the directive? How can an insurance company address those requirements with its existing IT infrastructure?

Since the directive was introduced in 2009, many questions such as above are waiting to be answered. Companies are trying to find their way within the regulatory documents and translate the new rules to their own environment with the guidance of consultancy companies. This study aims to answer some of these questions in a structured way and provide guidance to companies that need to be compliant with the directive.

In this chapter, we represent the underlying research questions of the study including why these questions are relevant to investigate. Then, we outline the objectives of the study and the research methodology used trough out the thesis. Figure 1 presents the development of research organized as theoretical and practical parts.

1.1. Problem Statement

This study poses the following research questions:

1. What are the requirements of the Solvency II regulation on Data Quality Management?
2. What are the characteristics (or specifications) of a Data Quality Management system which will be used by an insurance company to address these requirements.

In relation to above questions two sub questions will also be investigated:

3. To what degree do existing Data Quality Management methodologies meet Solvency II requirements?
4. How can we fill the gaps between the regulatory requirements and the available methodologies based on above identified system specifications?

During the research, a field study in a Dutch insurance company took place to understand how insurance companies are coping with the directive's data quality requirements. Also, the field study provided an opportunity for practicing operationalization of some of the identified system specifications.

1.2. Research Objectives and Contribution

Many Data Quality Management (DQM) methodologies are available in the literature with different emphasis and different perspectives [3] [4] [5]. Some of them are designed for a specific data type such as web data or health records, while others are more generic approaches. However, there is no generally agreed and standardized solution for implementing a DQM structure in an enterprise environment. In most cases, companies produce their own solution using bits and pieces from various approaches or commit to a vendor solution.

Solvency II being the first directive to have a clear emphasis on data quality, introduces specific data quality (DQ) requirements and documentation, however without mentioning any specific Data Quality Management Methodology. Also, since organization types and scales vary, the directive does not provide information about the implementation steps that need to be taken to produce high quality data: *" …. More precisely, undertakings shall develop their own concept of data quality starting from a basic interpretation given for the terms 'accurate', 'complete' and 'appropriate'…."* [6]. To be able to develop a custom concept, the companies should derive what is exactly required on data quality from the regulatory documentation and translate those requirements into the requirements for their specific environment.

The objective of that study is providing guidance on Data Quality Management to insurance companies which need to be complied with the Solvency II regulation. First of all, a structured analysis of the requirements on DQ will be performed using the Solvency II regulatory documentation. Although, a specific regulatory document Consultation Paper 43 (CP 43) written towards DQ standards is available, some additional documents such as CP 56 and CP 75 should also be reviewed to derive the complete set of requirements [2] [6] [7]. Secondly, those requirements will be translated to the system specifications of a Data Quality Management System which could be implemented by an insurance company.

Furthermore, while the study examines the well-known methodologies, it will also shed a light on what additional steps need to be taken in utilizing those methods to achieve Solvency II compliance. Additional field study gives the opportunity to analyze the operations of a large insurance company which runs a Solvency II project.

Until now, Solvency II and data quality requirements are mentioned in the several white papers usually written by the consultancy companies [8] [9] [10]. However, none of the available studies are (a) Deriving the specifications of a Data Quality Management system from the Solvency II perspective via structured requirement analysis, (b) Examining usability of the existing methodologies by the insurance companies, (c) Exploring what is specific to Solvency II in relation to the Data Quality Management concepts.

1.3. Research Methodology

Throughout this study a *deductive research* approach associated with a *quantitative analysis* has been used [11]. Via this approach, general statements of the Solvency II regulation have been used to reach conclusions about the specifics of the intended DQM system.

The literature review phase started by examining the academic literature on Data Quality, Data Quality Management and Requirements Engineering. Secondly, Solvency II related literature is reviewed. Available Solvency II literature could be organized in three groups; regulatory documents by the EIOPA[2], white papers from the consultancy companies and academic research papers. Finally, during the field study the internal documents of the insurance company about the Solvency II project are reviewed and several interviews with the project members are conducted. After the initial literature review phase, based on the existing knowledge and findings, the system requirements have been analyzed. Then, those requirements are transformed to the system specifications. As a practical example, two of the specifications have been operationalized via qualitative and quantitative analysis methods using actual data obtained from a Dutch insurance company.

High level findings of the literature review phase are visualized as a MindMap diagram in Appendix II. The MindMap is used as a reference throughout the study; to keep the focus on selected topic areas; to structure the document around the selected topics and to associate reference articles to the selected topics. Figure 1 represents the development of research structure which is divided as a Theoretical Base and a Field Study. The Theoretical Base concludes with System Specifications of the intended DQM system and the proposed extensions for the existing methodologies. The Field Study concludes with recommendations for the Insurance Company (INSC).

[2] European Insurance and Occupational Pensions Authority (former CEIOPS: The Committee of European Insurance and Occupational Pensions Supervisors).

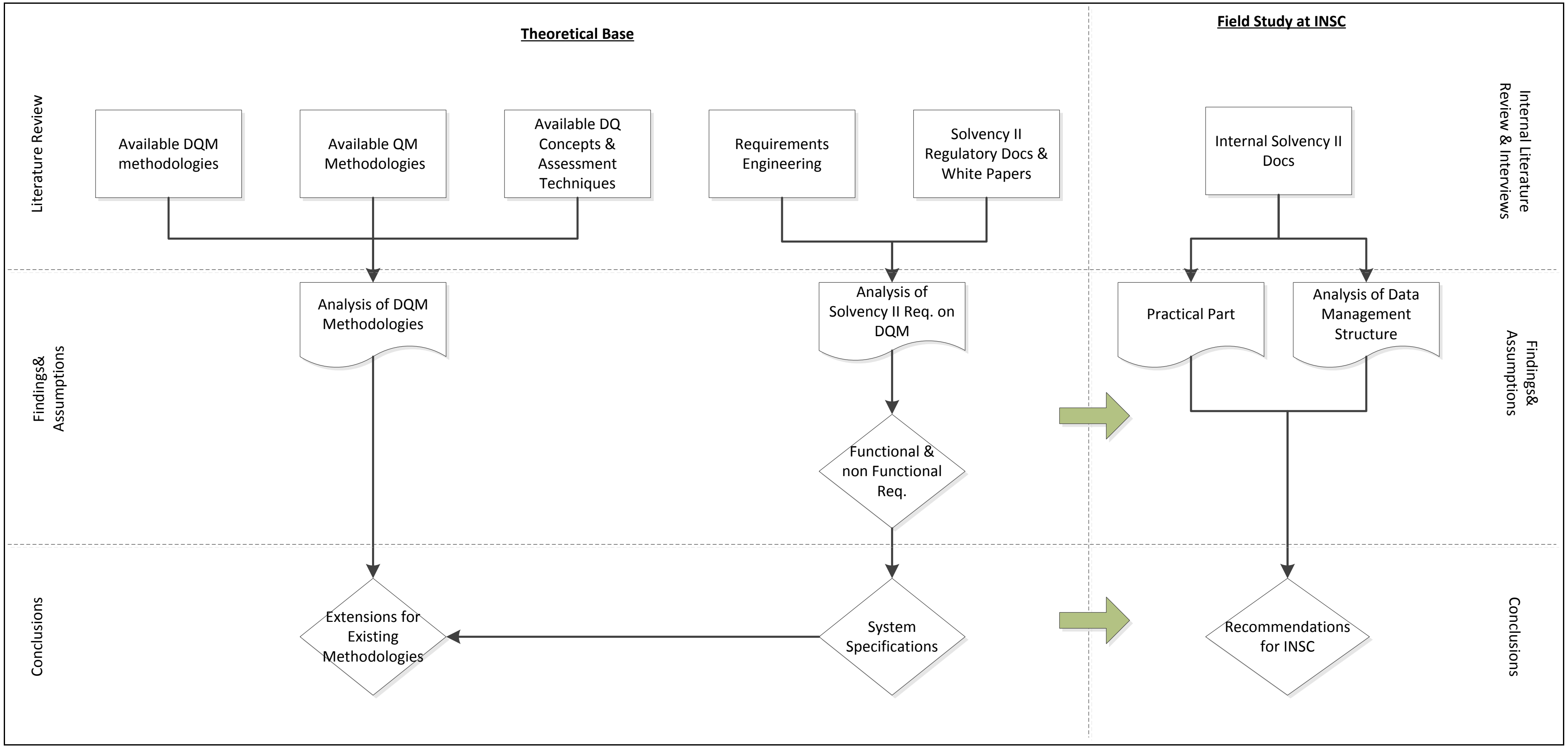
Theoretical Base
Field Study at INSC
Literature Review
Available DQM methodologies
Available QM Methodologies
Available DQ Concepts & Assessment Techniques
Requirements Engineering
Solvency II Regulatory Docs & White Papers
Internal Solvency II Docs
Internal Literature Review & Interviews
Findings& Assumptions
Analysis of DQM Methodologies
Analysis of Solvency II Req. on DQM
Practical Part
Analysis of Data Management Structure
Findings& Assumptions
Functional & non Functional Req.
Conclusions
Extensions for Existing Methodologies
System Specifications
Recommendations for INSC
Conclusions

FIGURE 1. RESEARCH DEVELOPMENT

1.4. Document Organization

The remainder of this study is organized as the following: In Chapter 1, the research problem, relevance of this study and the research method is explained. In Chapter 2 and Chapter 3, the background of the study is outlined under Solvency II and Data Quality titles. In Chapter 4, a Data Quality Management system is proposed via requirements analysis. Chapter 5 includes analysis of the existing DQM methodologies. Chapter 6 consists of the practical part of the study where an insurance company's environment analyzed. And finally in chapter 7, conclusions of the study are outlined.

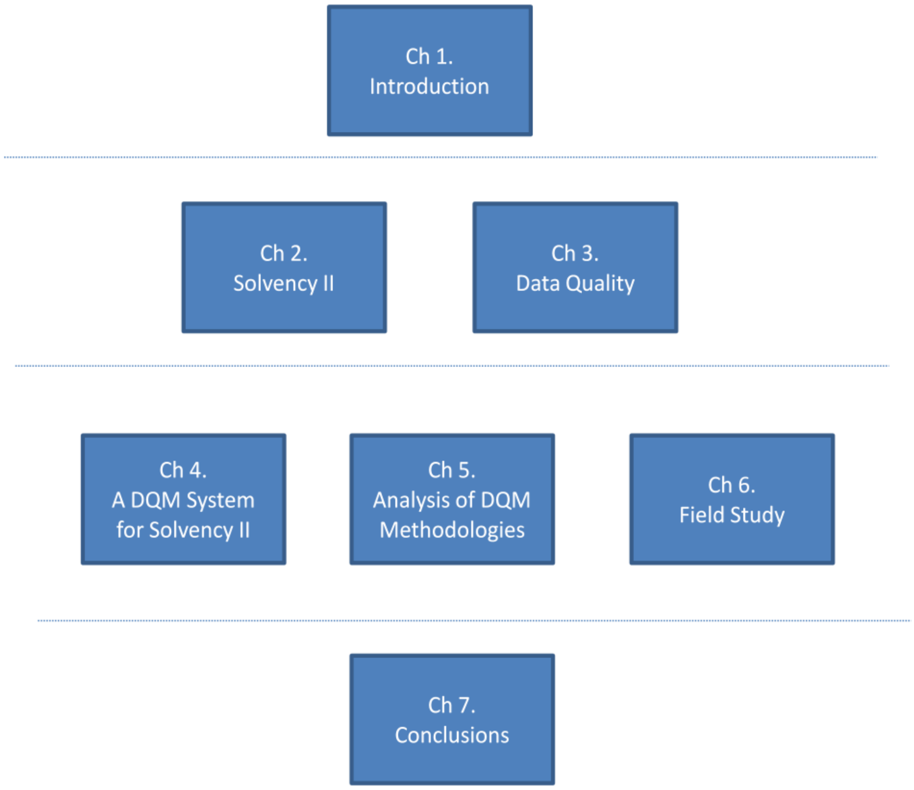

FIGURE 2. DOCUMENT ORGANIZATION

Chapter 2. A New Regulation for the European Insurance Companies: *Solvency II*

In this chapter, information on the Solvency II regulation will be provided in order to understand why such a regulation is needed. Then, the section will continue with explaining the relationship between the regulation and data quality showing that where data quality takes place within the regulatory framework. Finally, Solvency II approach to two important elements of data quality, assessment and management, will be outlined.

2.1. Solvency II Background

Following the recent financial turmoil, in November 2009, the European Union has adopted the Solvency II directive which was under development since 2004 [12]. The development project was managed by EIOPA targeting to unify the insurance market across the European Union under a single regulatory framework [10].

The press release of European Union on the topic summarizes the fundamental reason behind the directive*:*
"The aim of a solvency regime is to ensure the financial soundness of insurance undertakings, and in particular to ensure that they can survive difficult periods. This is to protect policyholders (consumers, businesses) and the stability of the financial system as a whole." [13]

The Solvency II directive is concerned with the amount of capital European insurance companies must hold to reduce the risk of insolvency [14]. It aims to provide protection to policy holders establishing capital requirements and risk management standards that will apply across the EU. It encourages companies to manage risk in a way that is appropriate to the size and nature of their business. Unlike the previous Solvency I directive (1973), it has a risk based approach and moves away from *"one size fits all"* approach [15]. Thus, the insurer should reach a tailored solution in balancing own costs and benefits. Additionally the directive offers incentives to insurance companies for better measuring and managing their risk situation such as expecting lower capital requirements, or lower pricing, etc [7]. Consequently, Solvency II is called as a *principle based* regulation meaning that evolving market practice is the key driver of the regulatory standards rather than imposing the standards from the top which might not be suitable for the market conditions [16].

Initially, the Solvency II directive was planned to become effective by the end of 2012. However, recently the implementation date for the insurers is postponed to 1 January 2014 [17] due to extensive political negotiations among the parties involved [18]. Until now, UK and the Netherlands are named as the best prepared countries for the regulation followed by Germany and Italy [18].

In the meanwhile, the Financial Services Authority (FSA) published results of the first review where they assess whether firms' data management structure complies with Solvency II regulations [19]. FSA used *"external review scoping tool"* which was published in 2011 as a reference during the assessment process [20]. The tool consists of five high level requirements of the directive and their expected controls which will be used by the firms for self-assessment on their data management status. According to the review results, the firms are mainly having difficulties in; implementing organization-wide data governance approach including appropriate roles; identifying and documenting data used in internal model; articulating what 'accurate', 'complete' or 'appropriate' meant in practice; demonstrating the effective operation of data quality checks. The report results show that most of the firms have still long way ahead in achieving Solvency II compliance on data management standards.

2.2. Regulatory Framework and Data Quality

The new solvency system, which consists of three pillars similar to Basel II[3], includes both quantitative and qualitative aspects of risk management (Figure 3). Each pillar focuses on a different regulatory component; respectively capital requirements, risk measurement and management and finally reporting. In the following, the content of these pillars is explained including the data quality concerns of each pillar.

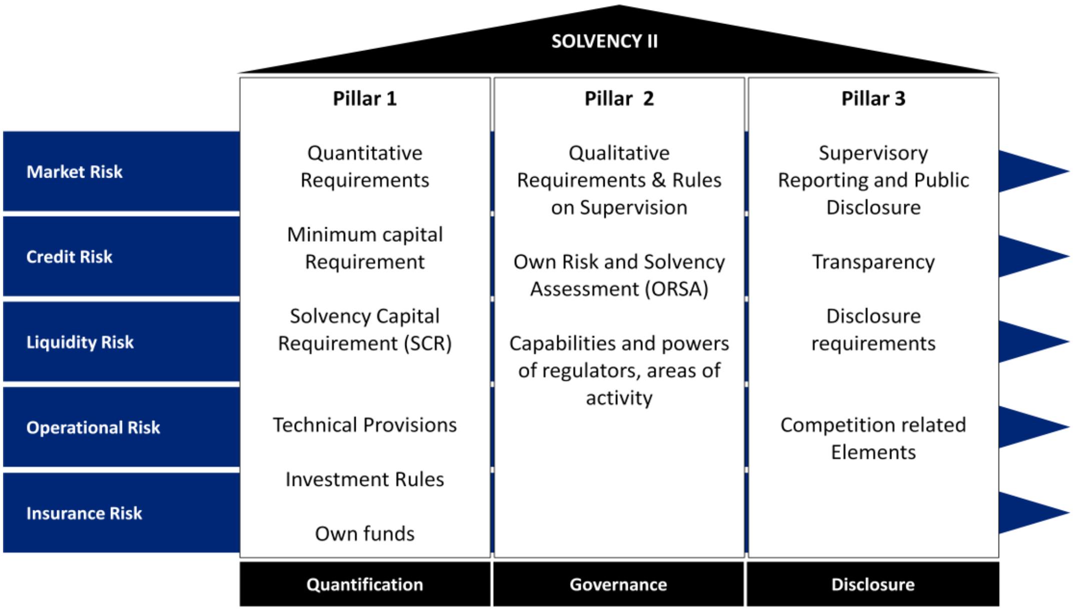
SOLVENCY II
Market Risk
Credit Risk
Liquidity Risk
Operational Risk
Insurance Risk
Pillar 1
Quantitative Requirements
Minimum capital Requirement
Solvency Capital Requirement (SCR)
Technical Provisions
Investment Rules
Own funds
Pillar 2
Qualitative Requirements & Rules on Supervision
Own Risk and Solvency Assessment (ORSA)
Capabilities and powers of regulators, areas of activity
Pillar 3
Supervisory Reporting and Public Disclosure
Transparency
Disclosure requirements
Competition related Elements
Quantification
Governance
Disclosure

FIGURE 3. SOLVENCY II PILLARS [21]

[3] Recommendations published by the Basel Committee on Banking Supervision on banking laws and regulations.

The first pillar deals with the quantitative requirements of the directive involving calculation of *Solvency Capital Requirements (SCR)*[4] and *Technical Provisions (TP)*[5]*.* In order to calculate TP and SCR, firms need to collect disparate data from various types of information sources that often would be dissimilar [9]. However, quality of data used in those calculations is critical as the amount of SCR indicates the availability of sufficient capital for the firm to run its operations. If the firm's available capital is less than the SCR, that may lead to an unacceptable risk for the policy holders and it would be the indication of an early warning for the regulators. Additionally, the reliability of the *internal model*[6] that is used to calculate regulatory capital requirements depends on the quality of the data used for validating the model. As part of *the internal model approval process (IMAP),* an insurer is required to provide evidence of input and output data quality that used in the models [22].

Pillar II consists of the qualitative requirements of the directive that includes an adequate governance system with a proper risk management approach. Usually, a large percentage of the operational risks are caused by poor data quality while running a company's operations. For example, duplicate claims payment, Service Level Agreement (SLA) violations [9] or incorrect policy premium estimation could be related to poor data quality issues. In order to effectively mitigate these risks, insurance companies need to use appropriate controls to detect and prevent data quality issues in operational systems [23].

Pillar III directives are related to *public disclosures* via transparent methods [21]. The companies should provide periodic reports on their operations that include data reconciled with different financial reports to increase the report's reliability [22]. The processes and systems used to generate the reports should be transparent enough to be able to trace obtained data until its source system. Consequently, companies need adequate procedures and systems in place while producing public disclosures.

Solvency II clearly places the emphasis on data governance as an essential part of risk management and the application of consistent standards and definitions across the pillars [24]. The directive is the first regulation that introduces strict data quality requirements for insurers [8]. Acknowledging the importance and criticality of data quality, EIOPA has issued a specific advice on data quality standards for data to be used by various models to calculate TP and SCR. The advice, briefly called CP 43, explains the relationship between high quality of data and decision making in the following statement;
"In general, the more accurate, complete and appropriate the data used in the internal model, the more reliable the resulting model output, and the probability distribution forecast in particular, and the greater the confidence that can be placed in the decisions made on the basis of the model results" [2].

[4] Solvency Capital Requirements (SCR) is amount of capital a firm needs to cover liabilities and provisions for the various types of risks such as underwriting risk, credit risk, market risk, and operational risk [13].

[5] The term 'technical provisions' is an all-embracing term used in the Insurance Accounts Directive (IAD) to cover (for general insurance) provisions for items such as unearned premiums, unexpired risks, claims outstanding (whether or not reported), equalization [87].

[6] The alternative to the standard model for deriving Solvency II capital adequacy requirements.

Although the directive highlights importance of over all data quality, especially data used in the critical calculations (such as SCR) should maintain a high quality standard which is not necessarily an obligation for other data. Therefore, it is possible to conclude that different quality levels are required for different data, based on "use purpose" of data within Solvency II and the highest quality levels are only consideration for a small amount of data.

In addition, as mentioned earlier, being a principle based regulation; the directive does not dictate any specific Data Quality approach but rather sets Data Quality standards for insurers. In a sense, insurers have the freedom to use the most suitable Data Quality concepts for their particular environment as long as they are able to meet the published standards.

2.2.1. Solvency II and Data Quality Assessment

Solvency II directive sets three essential criteria to assess data quality used in the valuation of technical provisions as part of Pillar I. Definitions of these criteria according to Consultation Paper 43 [2] and author's interpretation are given below:

Accuracy
"Data is considered to be accurate if it is free from material mistakes, errors and omissions."
Accuracy of data is mainly related to mitigating data errors caused by human error or system failures. Another cause of data errors is more difficult to resolve; poor system architecture. Usually in an enterprise environment different data systems are used to address different business requirements and the interfaces between those systems are not always automated. Furthermore, data recording should be timely and periodic to obtain accurate data. Organizations should implement cross-checks and internal tests to control accuracy of data.

Completeness
"Data is considered to be complete if it allows for the recognition of all the main homogeneous risk groups within the insurance or reinsurance portfolio."
Completeness of data is related to having sufficient granularity and sufficient historical information available as well as collecting all relevant items for the intended purpose.

Appropriateness
"Data is considered to be appropriate if it is suitable for the intended purpose and relevant to the portfolio of risks being analyzed."
If data is appropriate, it should allow valuation of technical provisions, setting of assumptions and it should be consistent with a prospective view of the behavior of the relevant risks.

According to CP 43, completeness and appropriateness should be assessed *"minimum in portfolio level or where appropriate at more granular level such as individual items"*, where as accuracy should always be assessed in individual item level.

Although the directive requires periodic assessment of above listed criteria, it doesn’t state how often the checks should be performed. While the frequency of the checks is related to the operational structure of the organization, most likely, the frequency should be high enough to capture data deficiencies in a timely manner.

Moreover, the degree of appropriateness, completeness and accuracy should be consistent with *the principal of proportionality* and with *the purpose of analysis*. For instance, less data is expected to be available while evaluating simple risks. On the other hand, where the nature, scale and complexity of the underlying risks is high, companies should pay increased attention to amount and quality of data they used in risk calculations. However, that approach shouldn’t be used as a justification to lower data collection standards, considering that the past data could be relevant in the future for a new business line that the company will enter [2].

2.2.2. Requirements on Data Quality Management

In order to guarantee sufficient data quality on a continuous basis, the directive emphasizes the importance of having a Data Quality Management (DQM) structure in place. As a continuous process, the DQM should compromise the following steps according to the regulatory document CP 43 [2];

1. Definition of the data: A comprehensive list of data required by the provisioning process including detailed description of data items that should be collected.

2. Assessment of the quality of data: Verification of the criteria of appropriateness, completeness and accuracy for the purpose of the analysis. In particular, when data is provided by the third parties, the channels used to collect, store, process and transmit data should also be considered during quality assessment.

3. Resolution of the material problems identified: When such a problem has been identified, the insurer should try to solve the issue within an appropriate timeframe. The root cause of the issue should be traced and solved where it is possible to prevent future repetitions of the similar deficiency. Where a solution is not possible it should be documented including proposed remedies.

4. Monitoring data quality: Data quality should be monitored periodically based on data quality performance indicators identified.

Appropriate processes and procedures should be in place within the organization to achieve the steps listed above. Using those transparent processes and procedures, all collected data should be registered, should maintain sufficient granularity, should be kept for an appropriate period allowing historical analysis and should be assessed periodically. Any adjustment made on the data should be documented as well as its reasons.

Chapter 3. Data Quality Concepts

In this section, current literature on Data Quality (DQ) is introduced. Data Quality Measurement and Data Quality Management are the most examined sub-topics of Data Quality according to the literature. In general, while Data Quality Measurement covers measurement techniques and approaches, Data Quality Management covers management strategies to implement, maintain and improve data quality within an organization. Although DQ has been examined extensively in the literature for quite some time, current research is still far from establishing industry wide standards. Existing studies introduce a wide range of methods and approaches to describe DQ concepts. For instance, identifying and describing DQ dimensions (e.g. accuracy, completeness, currency) is one of the essential steps in DQ measurement and management activities. However, several different approaches are available for this step, such as finding and describing quality dimensions, based on user perspective via surveys [25], considering Information System as a representation of a Real-World System [26] or via intuitive approach namely practical experience [3] [27].

3.1. Data Quality

Before exploring available DQ research, initially we need to understand what data quality is. First, we will start with describing data and quality separately, before moving to the description of data quality.

Data are values of qualitative and quantitative variables, belonging to a set of items [28] . It represents real world objects, in a format that can be stored, retrieved and elaborated by a software procedure [3]. *Information* is defined as data that processed to be useful [29]. In the computing literature, some of the studies use data and information terms inter-changeably. Although there is still an ambiguity around their definitions, a consensus is also available that they are not the same thing [29]. During this study, *data* and *information* are used based on above definitions.

In the literature, several types of *data classifications* are available. The following data types are defined among the others by several researchers [3]:

1. Structured, when data elements have associated fixed structure and resides in a fixed field within a record; such as relational databases, spreadsheets.
2. Semi-structured, when data has a structure which has some degree of flexibility. It is also called schema-less; such as an XML file which doesn't have an associated XML schema file during design.
3. Unstructured, when data is expressed in natural language and doesn't reside in fixed locations; such as word processing documents and e-mail messages.

Data classification is a substantial part of data management activities as it provides information on available data types, data location and access levels. Also knowing available data types would be an input for DQ dimension selection activity and measurement process; different quality measurement techniques are used for different data types.

Definition of *quality* varies and usually depends on the role of people who describes it [30]. Therefore, different definitions of quality are available. According to Oxford Dictionary[7]**,** quality means "*the standard of something as measured against other things of a similar kind; the degree of excellence of something*". Total Quality Management concept describes quality as "*meeting the customer's needs*" or "*satisfying the customer*" with a service oriented approach [30]. In manufacturing, *"a measure of excellence or a state of being free from defects, deficiencies, and significant variations, brought about by the strict and consistent adherence to measurable and verifiable standards to achieve uniformity of output that satisfies specific customer or user requirements"* [31]. And finally, ISO 8402-1986[8] combines both service and product perspectives and defines quality as "*the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs*" [31].

Similar to quality, finding a standard definition of DQ is also difficult. One of the common definitions of *data quality* is "fitness for use" which is also used to define quality in general; *"Data are of high quality if they are fit for their uses (by customers) in operations, decision-making, and planning. They are fit for use when they are free of defects and possess the needed features to complete the operation, make the decision, or complete the plan"* [32].

Although DQ issues could be dated as early as beginning of computer use by organizations, until gaining attention in mid 80s, limited number of studies are produced. According to an investigation conducted by *China National Institute of Standardization* in 2008, number of published academic papers on DQ is increasing since 1981 [33]. In the 70s, data quality issues were mainly addressed in database modeling solutions. When Codd introduced relational database model (1970), he advised about data integrity [34]. Ivonov's doctoral thesis (1972) on *quality control of information* is one of the early studies that recognizes the importance of quality of information for data-banks and management information systems (MIS) [35]. In 1985, Ballou et al. proposed a model to assess the impact of data and process quality in multi-user systems [36]. Another significant academic work is Hansen's master thesis (1991) titled *Zero Defect Data*, where he illustrated the impact of poor data quality in the economy and adapted statistical process control to data quality [37].

[7] http://oxforddictionaries.com/

[8] ISO 8402 – Quality management and quality assurance vocabulary released in 1986 by International Organization for Standardization.

In the 90s, foundation of MIT[9]' *Total Data Quality Management Research Program* accelerated studies with several well-known papers; [25] investigates the user understanding of DQ via structured surveys resulting in a comprehensive dimension list including their definitions; [38] aims to address data consumer's concerns on accessibility and contextual DQ issues; [26] proposes a method to analyze data quality based on ontological concepts; [39] introduces a methodology to define DQ parameters important to users. In the same period, English published his pioneering book where he showed impact of DQ on costs and profitability and suggested practical solutions for organizations to improve DQ [4].

Since the beginning of 2000s, various methodologies and techniques have become available for Data Quality management and measurement. However, we are still far from establishing standards. Batini and Scannapieco's book on Data Quality [3] provides a good overview of the available DQ management methodologies and measurement techniques. The book also introduces a new generic DQ management methodology which is mentioned in the DQM methodologies section of this chapter. Another study that includes majority of DQM methodologies is written with Batini's contribution as well; the paper compares the methodologies in a structured way from different angles [40].

Nowadays, organizations are increasingly searching methods to assess their DQ levels and improve it. Therefore, research in the field is well-appreciated by the software market as well as the academic world. However, software firms still need domain independent standards on data quality to be able to develop their domain specific solutions. Also firms need to use similar standards while adopting a data quality approach. A new ISO[10] standard, ISO 8000, is targeting to address those needs based on the NATO codification system in use for 50 years [41]. Currently, ISO 8000 is partly developed and when it's completed it will provide mechanisms to assure DQ in organizations.

As we mentioned earlier, available DQ research concentrates around assessment or measurement of DQ and management of DQ. In the following two sections, we look at the available studies in these two areas.

3.2. Data Quality Assessment

Data quality is defined as a multi-dimensional concept in several studies [36] [27] [25]. Usually DQ assessment activities start by choosing the quality dimensions which suits the company's specific application. The internal description of the appropriate dimensions should be in line with the company's goals. But what is a quality dimension? Below are some definitions:

"*Properties of data such as accuracy, completeness, timeliness and so on, benefit to decision makers are described as data quality dimensions* [42]. "

"A set of DQ attributes that represent a single aspect or construct of data quality [25]."

[9] Massachusetts Institute of Technology
[10] International Organization for Standardization

Mainly three approaches are available in the literature to identify the dimensions: (1) Theoretical, (2) Empirical and (3) Intuitive. Theoretical approach, considering the information system as a representation of a real-world system, investigates that how data may become deficient during the information product manufacturing process [26]; empirical approach focuses on selecting data quality dimensions interviewing the users [25]; intuitive approach, dimensions are selected for a particular study based on the researcher's experience and intuitive understanding of what attributes are important and align with goals of the study [3].

Using empirical approach, Wang et al. identifies more than 100 dimensions [25]. Then the authors eliminate and consolidate the dimensions based on their importance ranking for users. The final list groups eliminated dimensions into the four categories:

1. *Intrinsic data quality* (accuracy, objectivity, believability and reputation), captures the quality that data has on its own.
2. *Contextual data quality* (value-added, relevancy, timeliness, completeness and appropriate amount of data), considers the context where data is used.
3. *Representational data quality* (interpretability, ease of understanding, representational consistency and concise representation), captures aspects related to quality of data representation.
4. *Accessibility data quality* (accessibility and access security), captures accessibility of data and levels of security.

Solvency II resources mention three of these dimensions as the DQ criteria; Accuracy, Completeness and Appropriateness. *Accuracy* is defined as *intrinsic data quality* in above category list. *Completeness* dimension falls into *contextual data quality* group. *Appropriateness* is not listed. But we place it in the *contextual data quality* category based on its Solvency II definition.

Keeping their Solvency II definitions in mind (Chapter 2), we now present the definitions of the three dimensions in the literature. The definitions vary and there is no agreement on the exact meaning of each dimension, therefore several definitions are available for each dimension. Also, some of the measurement methods for each dimension are mentioned below. The scope of definitions and measurement methods are kept in line with the data types that will be used in the Solvency II directive.

Accuracy

Batini and Scannapieco define accuracy as *"Closeness between a value V and a value V' , considered as the correct representation of the real-life phenomenon that V aims to represent"* [3].

Wang and Strong define as *"The extent to which data are correct, reliable and certified free of error"* [25].

In practical terms, accuracy is defined as the closeness between the data value and the true value. Two types of accuracy are mentioned in the literature: *syntactic accuracy* and *semantic accuracy* [3]. *Syntactic accuracy* means that the considered value might not be correct, but it belongs to the domain of the corresponding attribute [43]. For example, where data value is *Jack* (V) and true value is *John* (V'), Jack is considered syntactically correct as it is a value from *person's name* domain. *Semantic accuracy* means that the data value might be in the corresponding domain, but it is not correct. It is also defined as the closeness of data value V to the true value V'. For example, considering *gender* attribute value, where data value is *female* (V) for John, accuracy would be syntactically correct since *female* is part of the gender domain. But obviously, semantic accuracy would be wrong for the value of *gender* attribute associated with name John.

As could be seen from these definitions, measuring syntactic accuracy is a relatively straight forward activity. We just need to know domain attributes to verify whether a value lies in the domain or not. On the contrary, verification of semantic accuracy is a difficult and complex task [43], [3]. A commonly used technique in verification of semantic accuracy is looking for the same data in different data sources to compare with data in hand.

Completeness

Wang and Strong define completeness as "The extent to which data is of sufficient breadth, depth, and scope for the task in hand" [25].

Wand and Wang define as *"The ability of an information system to represent every meaningful state of the represented real world system"* [26].

In [43], two types of completeness are identified. *Completeness with respect to the attribute values* refers to missing data values. Once missing values are marked, a ratio between missing values and whole values gives the measurement of completeness of data values. Another type of completeness is *completeness with respect to the records*. That means, whether the data set contains necessary information for analysis or not. In other words, data records should include sufficient information to allow analysis.

Appropriateness

The appropriateness dimension is not commonly mentioned in the literature. In [25], the term is used for *appropriate amount of data* dimension. However, the dimension that covers the meaning of appropriateness in [25] seems to be *relevancy*, which is described as applicable, relevant and usable. Additionally, *relevancy* also falls into *contextual data quality* category verifying our previous conclusion on categorization of appropriateness. Unlike accuracy and completeness, measurement of relevancy is less explored in the literature. It seems that contextual dependency of relevancy dimension makes applying a generic measurement technique problematic: While data is highly relevant for one task, it can be irrelevant for another task [44].

After selecting dimensions, both *subjective* and *objective assessment* techniques should be used for measurement [45]. *Subjective assessment* reflects all users' experiences, needs and opinions. Usually, questionnaires are used to understand users' perception of data quality. On the contrary, objective assessment is associated with metrics, statistical analysis techniques and using assessment algorithms. Metrics could be either task-independent or task dependent. While task-independent metrics are used for context –free measurement, task-dependent metrics are developed in a specific application context. Following objective and subjective assessment, the results should be compared; discrepancies should be identified and investigated. In the literature, there are many studies about identifying dimensions. However, measuring dimensions and relating those measurements to standardized metrics are less explored topics and there is strong need to establish a statistical measurement basis for DQ dimensions and indicators. Identifying dimensions is not sufficient, if they cannot be measured.

Pepino et al. proposes one of the following functional forms to develop metrics as a part of the objective assessment of variety of dimensions [45]; *simple ratio*, *min/max operation* or *weighted average*. However, the paper also concludes that "one size fits all" metrics are not the solution. Organizations need to develop and utilize their internal metrics using subjective and objective assessment methods as an ongoing operation. In [46], the authors introduce a metric based approach and describe how DQ metrics could be designed to quantify DQ. Similar to the task dependent and independent measurement described earlier; according to [44], while some data quality dimensions are invariant, some others vary based on context which makes data quality measurement complex. The paper proposes a *dual-process approach* for data quality assessment of both objective (task-independent dimensions such as accuracy, completeness, timeliness) and contextual dimensions (task dependent dimensions such as relevancy, believability). Another example of contextual DQ measurement is [47]. Using content based measurement method; the paper introduces a conceptual measure of business value (intrinsic value) that is associated with the evaluated data.

3.3. Data Quality Management Methodologies

Data quality management (DQM) is described as quality oriented data management. It focuses on collection, organization, storage, processing and presentation of high-quality data [48]. DQM could also be seen of as a specialized version of the existing quality management methodologies towards data management. Those methodologies were developed much earlier than DQM and had a big influence in establishing current DQM concepts. In this section, two of the best known and the most practiced quality management methodologies are briefly explained as they assist us in understanding DQM methodologies explored during this study.

Total Quality Management (TQM) is the earliest quality management approach in the literature. TQM is described as *"an integrated organizational effort designed to improve quality at every level"* [30]. Evaluation of TQM started in 1920s with quality control efforts in production lines. Initially, application of statistical methods to the management of quality was developed by Shewhart, who was a statistician in Bell Labs, to minimize variation in production process which leads to variation in products. After WWII, Japanese manufacturing companies were willing to improve their production quality and they adopted quality control and management methods. In this period, Deming, who was a statistics professor in New York University, assisted many Japanese companies to improve their quality. Interestingly, he pointed out that large majority of quality problems were caused by processes and systems, including poor management, rather than worker error. After Deming, Juran had a big influence on quality management. While Deming stressed an organizational transformation to achieve effective quality management, Juran argued that quality management should be embedded in the organization and shouldn't require any dramatic change. He focused on definition and cost of quality during his studies and was credited with defining quality as "*fitness for use*". He also developed the idea of quality trilogy as a continuous cycle: *quality planning, quality control and quality improvement.* In the 60s, Feigenbaum introduced the concept of *Total Quality Control*. He suggested that quality developments should be integrated throughout the entire organization and management and employees should be committed to improving quality. During the following periods, evaluation of TQM continued with the contribution of several others. TQM has been transformed from being a manufacturing oriented approach to a business management system that is usable by the different industries and its focus extended to embrace quality of the "service" as well as quality of the "product". Today, being an ongoing process TQM is still practiced by many organizations. Similar to Data Quality Management concepts, TQM has also focus on identifying root causes of quality problems and correcting them at the source.

Compared to TQM, Six Sigma is a relatively new concept. It is defined as *"a business strategy that seeks to identify and eliminate causes of errors or defects or failures in business processes by focusing on outputs that are critical to customers"* [49]. It was developed at Motorola in 1986 and gained broader attention when it was adopted by General Electric in 1995. Today, it is used by many different industries as a fact-based, data driven philosophy of quality improvement that values defect prevention over defect detection [50]. Similar to TQM, Six Sigma was also originated from manufacturing processes. Utilizing quality management tools and methods lay at the center of Six Sigma. In [51], it is indicated that Six Sigma emphasizes the importance of decision making based on facts and data, rather than assumptions. Existing DQM methodologies and tools also adopt a similar approach and use statistical methods extensively targeting a realistic measurement of quality and timely detection of defects.

The literature is abundant with methodologies for organizing DQ activities in companies. Among those, the following five methodologies are selected for further analysis in this section (Table 1).

TABLE 1. DATA QUALITY METHODOLOGIES [38] - EXTENDED WITH ORME-DQ

	Acronym	Methodology	Main Reference
1	TDQM	Total Data Quality Management	Wang [5]
2	AIMQ	Information Quality Assessment and Improvement Methodology	Lee et al. [52]
3	TIQM	Total Information Quality Management	English [4]
4	CDQM	Complete Data Quality Methodology	Batini et al. [3]
5	ORME - DQ	ORME - DQ	Batini et al. [53]

TDQM, AIMQ and TIQM are the most known methodologies in data quality literature and it is possible to find the practical examples of their implementations. CDQM is a more recent methodology that uses the building blocks of the previous methodologies as well as addressing their limitations. Only one paper was found about ORME-DQ and no further information is published related to the methodology[11]. However, the methodology was designed to support Basel II regulation which has some similarities with Solvency II on risk management approach. Therefore, ORME-DQ is added to our short-list.

Batini et al.' study is the most comprehensive paper on analyzing several DQ methodologies that we came across during the literature review [40]. It analyzes more than ten methodologies with a focus on DQ assessment and improvement activities. In addition to these, our literature review showed that more methodologies are available. Although some of them, such as TDQM, are practiced more often than the others, it is difficult to call any of the methodologies *de facto standard*. Furthermore, the methodologies are usually written towards a particular application such as ORME-DQ. *China National Institute of Standardization* argues that, no data quality framework is available independent from any particular domain or application [33]; consequently existing frameworks differ in many aspects. On the other hand, two studies categorize TDQM, TIQM and CDQM as *general purpose* methodologies that could be used by different industries [3] [40].

In conclusion, all of the above methodologies are considered to be tightly connected to a specific application which makes standardization of the concepts a challenging task. Even the general purpose methodologies would require modification and customization during implementation phase based on the organization's practices and goals. In the next sections we will look at each methodology in more detail from a Solvency II perspective.

3.3.1. Total Data Quality Management

Total Data Quality Management (TDQM) was introduced at the MIT in 1990s as an extension of Total Quality Management (TQM) to develop a theoretical foundation for data quality. TDQM uses the *information product (IP)* approach inspired by the analogy between manufacturing product of TQM and data. Wang summarizes the purpose of TDQM as *"delivering high quality information products to information consumers"* [5]. TDQM adopts Deming's *"Plan, Do, Check and Act"* from the TQM literature and creates its own *"Define, Measure, Analyze, and Improve"* cycle as a continuous process [53]:

1. *Definition* phase includes identification of data quality dimensions and related requirements.
2. *Measurement* phase produces quality metrics. The feedback provided by those metrics allow for the comparison of the actual quality with the predefined quality requirements.
3. *Analysis* phase identifies the root cause of quality problems.
4. *Improvement* phase focuses on quality improvement activities.

[11] Verified by one of the authors, Prof. Carlo Batini, during our e-mail communication.

Although TDQM partially uses the similarities between a manufacturing product and an information product in defining its concepts, dissimilarities are also mentioned [5]: Data can be utilized by multiple consumers and not depleted, whereas a raw material can only be used for a single physical product. Another dissimilarity arises from timeliness: For instance, we could say that raw material arrived just in time, that would not assign an intrinsic property of timeliness to raw material. Other dimensions, such as the believability of data, simply do not have a counterpart in product manufacturing.

TABLE 2. PRODUCT VS. INFORMATION MANUFACTURING [50]

	Product Manufacturing	**Information Manufacturing**
Input	Raw Materials	Raw Data
Process	Assembly Line	Information System
Output	Physical Products	Information Products

As the first published methodology in DQ literature, TDQM has a long history compared to the other methodologies. Different opinions are available on wide-spread usage of TDQM; although in [53] the authors give examples on "extensive application of TDQM in different contexts", in [54] the authors indicate that the resources on TDQM are scarce due to *"considerable problems in its application".* Based on our research during this study, we agree that the current literature on TDQM is limited. Although many articles refer to TDQM, the most of them do not explain the practical details of the methodology and how to design a DQM system based on TDQM. A few examples on TDQM practices include: a TDQM implementation in a market research company where internal DQ metrics are developed by Kovac et al. [55]; Wijnhoven et al. introduces a "well articulated" methodology as a result of a TDQM implementation where the theory had to be improved [54]; and Nadkarni describes a TDQM implementation in an insurance company [56]. Furthermore, another study extends TDQM by proposing *Information Production Map (IP-MAP)* concept to model Information Products managed by manufacturing processes [57]. Later on, IP-MAP model evolved into IP-UML in order to facilitate modeling of complex systems using processes and actors [58] .

3.3.2. AIMQ

The Information Quality Assessment and Improvement Methodology (AIMQ), has been developed to provide the ability to assess organizations' information quality (IQ) level [59]. That ability would assist organizations in knowing their IQ status and monitoring its improvement over time. Also, the methodology aims to provide a basis for using benchmarking techniques for organizations to compare their IQ level against the others.

AIMQ consists of three components: The first component is called *PSP/IQ model*. This is a 2 x 2 model of what IQ means to information consumers and managers [60]. The four quadrants of the model are used to consolidate dimensions as sound, dependable, useful and usable information. Those quadrants represent IQ aspects relevant to IQ improvement decisions. Each quadrant covers a group of dimensions:

- Sound information: Free of error, concise representation, completeness, consistent representation
- Dependable information: Timeliness, security
- Useful information: Appropriate amount, relevancy, understandability, interpretability, objectivity
- Usable information: Believability, accessibility, ease of operation, reputation

The second component, called as *IQA instrument*, is a questionnaire for measuring IQ according to the defined dimensions which are important to information consumers and managers. The questionnaire scales the dimensions allowing for statistical analysis of each dimension and their aggregations (quadrants).

Finally, the third component consists of two GAP analysis techniques for interpreting the findings of questionnaire for each quadrant and respective dimensions. The first technique, *benchmark GAP*, compares an organization's IQ to a benchmark. This benchmark consists of best-practices of several organizations. The second technique, *role GAP*, measures the distances between assessments of different participants (stakeholders) of an information production system.

AIMQ assesses IQ mainly using questionnaires. According to Batini et al., publications on AIQM focus on assessment activities, however no guidelines and techniques are provided on improvement activities [40].

3.3.3. Total Information Quality Management

Total Information Quality Management (TIQM) methodology (formerly known as Total Quality Data Methodology – TQDM) is inspired by quality management concepts similar to TDQM. Especially Deming Management Method and Keizen had a big influence while establishing the basis of the methodology. It has been initially designed to support data warehouse projects where data from different sources is consolidated into an integrated database [4]. TIQM focuses primarily on management activities that will be performed during the integration of those data sources, in order to make the right choices for the organization. A detailed classification of costs and benefits is provided as part of the methodology. The main goal of the cost-benefit analysis is finding out the most feasible quality improvement activities; such that once they are performed their benefit should exceed their cost.

TIQM consist of 6 process steps [61]:

1. Assess data definition and information architecture quality.
2. Assess information quality.
3. Measure non-quality information costs and risks.
4. Reengineer and correct data.
5. Improve information process quality.
6. Establish the information quality environment.

3.3.4. Complete Data Quality Management

Complete Data Quality Management (CDQM) methodology aims to establish a balance in between completeness and practical feasibility of the DQ improvement process [3]. To achieve the completeness, existing techniques and tools are incorporated into a framework that could be applied to any type of data, such as structured, semi-structured or unstructured. And practical feasibility is achieved via selection of the most appropriate methods for the organization.

The methodology emphasizes a tight connection between DQ measurement, improvement activities and business processes, and organizational costs. It targets to select the best improvement process that maximizes benefits and minimizes costs to the organization. The cost classification technique proposed by the methodology to assess organizational costs is a combined and improved version of the previously introduced cost classifications (i.e. English, Loshin and Eppler-Helfert classifications).

The methodology consists of three phases: *State Reconstruction, Assessment and Choice of the Optimal Improvement*. In the first phase, several matrices are created to represent the relationships among processes, organizational units and databases to understand which organizational units use which databases for which business processes. In the second phase, the new target DQ levels are set in improving process qualities. While setting the new levels, corresponding costs and benefits are evaluated. Then the processes are analyzed to locate the most problematic parts. In the final phase, the optimal improvement process is identified using the inputs of the previous steps and applying a cost-benefit classification to the candidate processes [40]. Batini and Scannapieco, report an implementation example of CDQM in the reorganization of Government to Business (G2B) relationships in Italy [3].

3.3.5. ORME-DQ

ORME-DQ methodology is introduced during the ORME project initiated by the Italian Ministry of Economic Development using CDQM methodology as a reference [40]. The methodology is specialized towards the Basel II regulation in relation to *"data quality and its effects on operational risk"* [53]. Since the low quality of information is treated as an operational risk factor for the banks, understanding actual information quality level of the organization and economic losses caused by poor data is the essential target of using the methodology.

The methodology consists of four phases: *DQ Risk Prioritization, DQ Risk Identification, DQ Risk Measurement and DQ Risk Monitoring*. In the first phase, the relationships between organizational units, processes, services and databases are represented via matrices to provide an overview of data flow, data providers and data consumers. During the second phase, economic losses caused by low data quality are calculated with the help of a cost hierarchy. For each selected cost item, a metric is defined and the corresponding economic value is calculated. In the third phase, appropriate datasets and dimensions are selected to be assessed. Then, the most feasible metric is used for assessment. In the final phase, DQ thresholds are defined to send automated alerts when the target values are exceeded.

Chapter 4. A Data Quality Management System for Solvency II

In the previous chapters, Solvency II regulation is explained from the data quality (DQ) point of view and the data quality concepts are outlined along with the major data quality management (DQM) methodologies. In this chapter, the connection between Solvency II and DQ concepts are realized proposing a DQM system that could be used to achieve Solvency II compliance.

While developing the DQM system, a *systems design* approach used in Systems Engineering, is adopted. In software development, *systems design* is described as follows: "T*he process of defining the components, modules, interfaces, and data for a system to satisfy specified requirements* [62]*".* In practice, that definition could be extended to cover any computer based system design such as implementation of ERP software.

In our case, the entire DQM environment is treated as the *DQM system* which consists of mainly *computer based* (e.g. data collection) parts and partially *non-computer based* parts (e.g. data collection procedures). To be able to design the DQM system, first we need to define system specifications via data quality requirements analysis of the directive: Analysis results constitute the specifications of the intended DQM system. Then, these specifications are explained in detail to give an idea of how the system should be implemented.

4.1. Requirements Engineering

Before performing a DQ requirements analysis on the regulatory documents, some important definitions should be provided.

Requirements Engineering (RE) is a systems and software engineering process which covers all of the activities to understand the requirements of the intended system. Those activities consist of capturing, documenting and maintaining required services, system users, operating environment and associated constraints [63].

Requirement Analysis (RA), as a sub process of the RE activities, covers determining the needs or conditions to produce a new computer based system taking into account various requirements of all stakeholders [63].

In 1977, Ross' well-known study on structured analysis describes systems requirement analysis as follows [64]:

> *"Requirement definition is a careful assessment of the needs that a system is to fulfill,*
> *WHY a system is needed,*
> *WHAT system features will serve and satisfy this context,*
> *HOW the system is to be reconstructed…… ."*

Basically, requirements constitute the specifications for a new system [65]. Functional and non-functional requirements are the common categorization of requirements. In software engineering, while *functional requirements* describe the nature of interaction between the components and their environment, *non-functional requirements* constrain the solutions that might be considered [63].

In more generic terms, functional requirements consist of inputs, outputs, the behavior which describes what a system supposed to accomplish and description of the data that must be managed by the system [65]. The intended behavior of the system may be expressed as services, tasks or functions the system is required to perform [66]. Non-functional requirements impose constraints on the design and implementation such as cost, security or performance requirements.

4.2. Analysis of DQ Requirements of Solvency II

In this section, mainly based on CP 43 document, which is dedicated to *the standards for data quality in calculation of Technical Provisions*, data quality requirements of the Solvency II directive are analyzed using the requirements analysis approach described earlier.

CP 43 is the essential resource to understand the regulation's data requirements. Although CP 43 is written for Technical Provision (TP) calculations, it has been recommended by the regulators that the document should be used as a reference for the entire Pillar 1 which focuses on the quantitative requirements of the directive. The following statement from CP 43 supports this approach: "….*to the extent appropriate, a consistent approach to data quality issues needs to be taken across Pillar 1, without however disregarding the different objectives and specificities of each area.*" Also, it has been stated in CP 56 that CP 43 will be applied, where possible, to the internal model data.

Furthermore, CP 56 and CP 75, which complement CP 43 in calibration of standard formula and operating internal model, are also included in the analysis to understand further data quality requirements.

The Solvency II legislation consists of several levels as represented in Figure 4. Each level addresses different stages of development of the legislation. The data quality related requirements take place in Level 2 as part of the implementing measures.

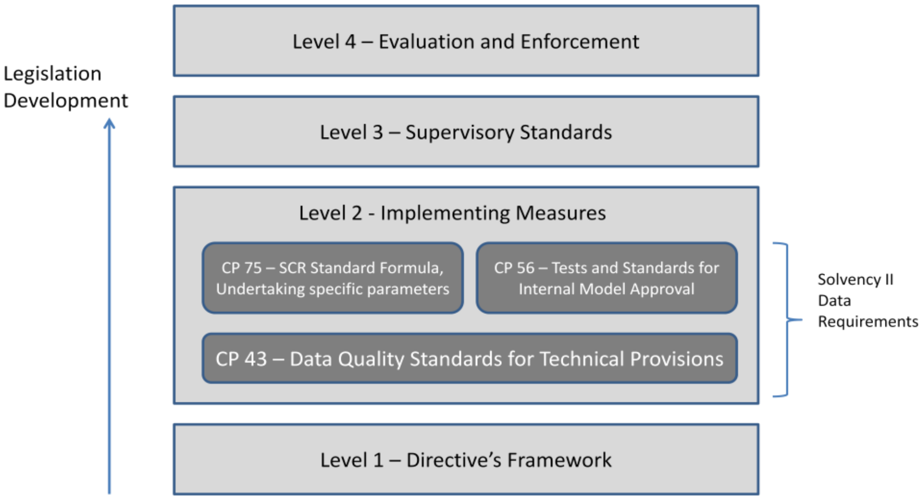
Level 4 – Evaluation and Enforcement
Legislation Development
Level 3 – Supervisory Standards
Level 2 - Implementing Measures
CP 75 – SCR Standard Formula, Undertaking specific parameters
CP 56 – Tests and Standards for Internal Model Approval
CP 43 – Data Quality Standards for Technical Provisions
Solvency II Data Requirements
Level 1 – Directive's Framework

FIGURE 4. SOLVENCY II LEGISLATION DEVELOPMENT AND REGULATORY DOCUMENTS ON DATA REQUIREMENTS

The analysis of these requirements aims to derive *specifications of a Data Quality Management System* which should be used by an insurance company in order to achieve Solvency II compliance. These specifications should be aligned with the mentioned regulatory documentation. Figure 5 shows the steps of the requirements analysis applied. Using common categorization techniques available in the literature, requirement analysis results are organized as *functional* and *non-functional*. Then, these two categories that complement each other are mapped on to the system specifications. *DQM System Specifications* describe an entire DQM system which consists of software, activities, processes and resources.

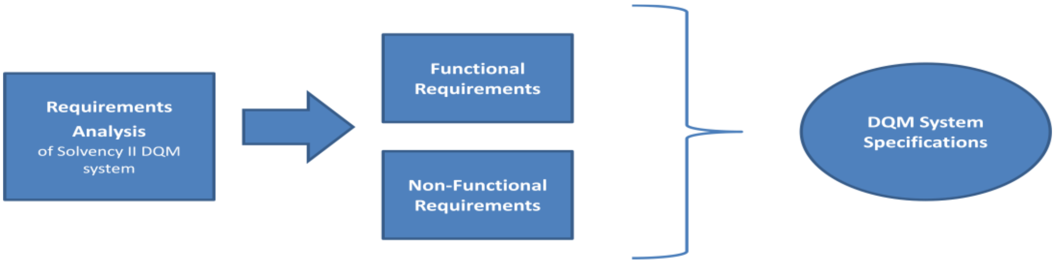
Requirements Analysis of Solvency II DQM system
Functional Requirements
Non-Functional Requirements
DQM System Specifications

FIGURE 5. REQUIREMENTS ANALYSIS STEPS

Requirements analysis process is initiated with answering the questions stated in Ross' System Requirement Analysis definition [64]:

WHY is a new system needed? Insurance companies are responsible for using sufficient and high quality data in the regulatory calculations. How this data is produced and transformed within the data flow should be transparent and traceable from source to target. Additionally, data used in calculations should meet accuracy, completeness and appropriateness criteria. Therefore, insurance companies need to transform their existing information system structure, which is mostly organized around delivering an up and running system, to a quality centric system to achieve the desired regulatory outputs. During the transformation, manual processes should be replaced with more automated processes where possible, which will contribute to the reliability of data.

WHAT system features will serve and satisfy this context? The required features should be organized around two categories: (1) Infrastructure related features such as having appropriate software and hardware in place and using a Data Warehouse system. (2) Governance related features such as having appropriate processes and procedures, an organizational structure, roles and responsibilities in place.

HOW the system is to be reconstructed? A Solvency II project should initiate the implementation of required features incorporating various departments such as Data Management, IT, Risk Management, Business Units and Corporate Governance. Insurance companies need to review existing DQM methodologies, identify the DQ requirements of Solvency II and produce their tailored solution. The solution should use the parts of the existing methodologies which suits the company's specific needs and business activities the most and it should address the regulatory requirements.

4.2.1. Functional Requirements

Figure 6 represents the functional requirements graphically in a *system view*, including required Inputs and Data Sources to run the desired System Functions and the expected system Outputs. This view provides the high-level organization of the entire *Solvency II Data Quality Management System* and the high level sequence of the system parts. The system functions are explained in detail in Section 4.3.

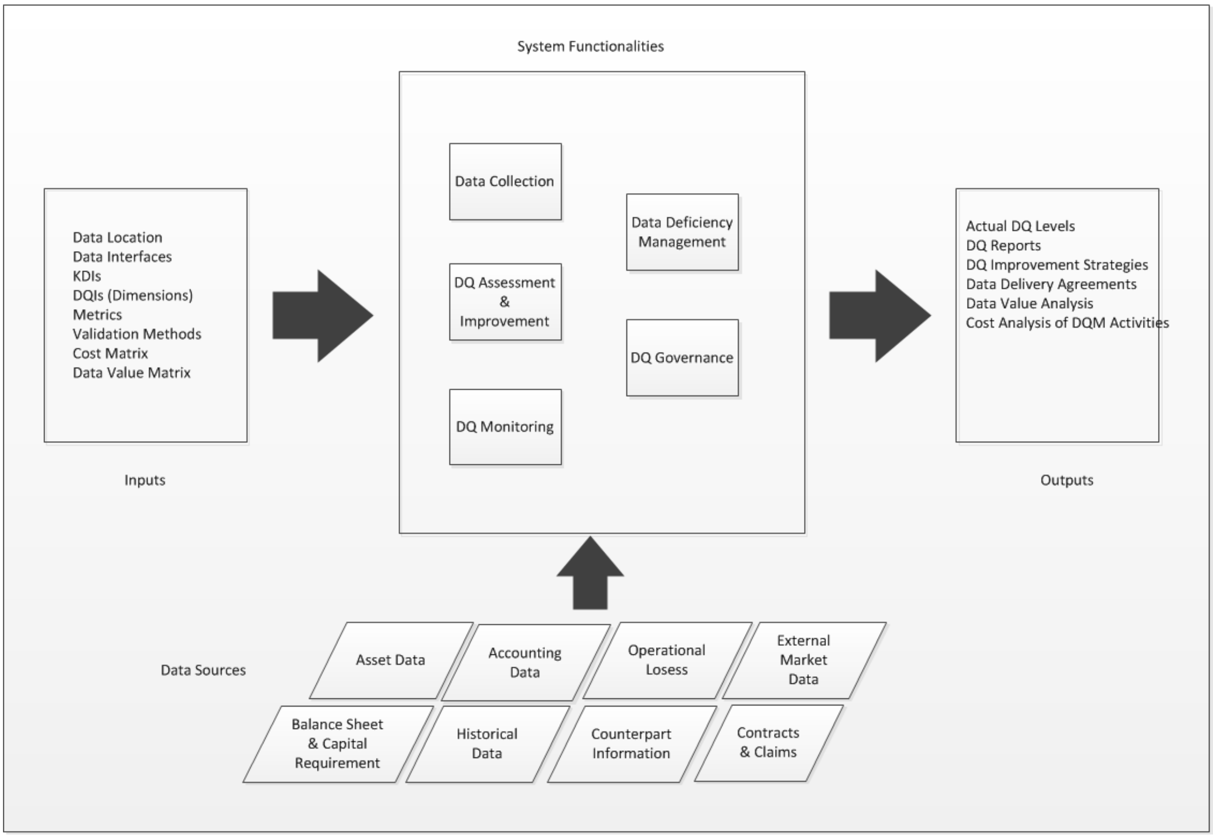

FIGURE 6. DESIGN OF THE INTENDED DQM SYSTEM - SYSTEM VIEW

In Table 3, the System Functions as a part of the Functional Requirements are shown. The table also describes each activity from Solvency II perspective and refers to the corresponding regulatory document.

In addition, *Use Cases* are a good way of visualizing required behavior of the system for a particular scenario [66]. In Appendix IV, a use case graphic of the Data Collection activity is developed using UML[12] representation as an example.

TABLE 3. SOLVENCY II FUNCTIONAL REQUIREMENTS ON DATA QUALITY MANAGEMENT

	Functional Requirements	**Solvency II Description**	**Respective Regulatory Document and Section Number**
System Functions	Data Collection	Collecting data required by Standard Model and Internal Model calculations in a structured way including data definitions and data quality indicators.	CP43 – 1.5, 3.2, 3.76, 3.1.4.2 CP56 – 9.31
	DQ Assessment and Improvement	Identifying quality level of data using transparent assessment methods (both objective and subjective) and considering quality of collection methods or source systems in the assessment process. Based on the assessment results companies continuously work towards improvement of data quality and surrounding process to ensure quality of future data.	CP43 – 3.36, 3.37, 3.1.1, 3.58 CP56 - 5.181, 5.3.3.3 CP75 – 3.21
	DQ Monitoring	Periodical monitoring of data quality based on data quality indicators, quality dimensions and expert judgment.	CP43 – 3.38, 3.80 CP56 – 5.146, 5.3.3.3, 5.3.3.5 CP75 – 3.28, 3.29
	Data Deficiency Management	Solving data deficiencies in a standardized way, in a certain time frame aiming to prevent error re-occurrence via root-cause analysis.	CP43 – 3.1.2, 3.37
	Data Quality Governance	A continuous process of identifying, implementing and updating internal process and procedures, roles and responsibilities required for data quality management activities.	CP43 – 3.1.4, 3.59 CP56 – 4.3, 5.3.3.6, 5.150

[12] Unified Modeling Language

4.2.2. Non-Functional Requirements

Although, a system's utility is described essentially by its functionality, in practice functional and non-functional characteristics complement each other. Usually, system functionality is emphasized over non-functional attributes during the system design. The pressure on delivering a functioning system during a software development project is one of the main reasons of this emphasis [67]. Also, ambiguity of non-functional characteristics such as usability, flexibility, performance, interoperability, contributes to that result. Consequently, different authors propose different definitions of non-functional characteristics. In this section, using a generic description from software engineering, namely "*a non-functional requirement is an attribute of or a constraint on a system*" [68], non-functional requirements of the Data Quality Management system are listed in Table 4. The table also includes the Solvency II definitions of the non-functional requirements referring to the regulatory documents.

Flexibility attribute included in the table below, is not explicitly mentioned in the regulatory documentations. However, flexibility and adaptability of a computer system is an essential requirement in system design, especially considering rapid changes in business activities and corresponding regulatory modifications. Therefore, it is included in the table as a requirement.

TABLE 4. SOLVENCY II NON-FUNCTIONAL REQUIREMENTS ON DATA QUALITY MANAGEMENT

Non- Functional Requirements	Solvency II Description	Respective Regulatory Document and Section Number
Performance	Data deficiencies should be resolved within appropriate time frame. The system should be able to collect data in required granularity. Appropriate amount of historical data should be available.	CP43 - 3.37, 3.26, 3.39
Reliability	Transparent, automated, well-documented system processes.	CP43 - 3.39
Security	Providing security and confidentiality of the information.	CP56 – 4.3 (j)
Usability	Required amount of data (principle of proportionality) should be available. Data should have the best-fit for intended purposes.	CP43 - 3.1.3
Flexibility	The system should be flexible enough to address changes/extensions on regulatory requirements or on business activities.	N/A
Cost	A balance is needed between restricted or comprehensive scope for DQM system to find the optimum cost for insurers.	CP56 - 5.130, 5.131

4.3. Specifications of a Solvency II DQM System

In this section, based on the previous requirement analysis results, high level specifications of a Data Quality Management system, which are applicable by an insurance company, are presented. Each functional requirement - together with group of appropriate non-functional requirements- is transformed to the specifications which constitute a Solvency II Data Quality Management System. These specifications aim to address how the requirements are met including methods and approaches which should be used (Table 5).

In Table 5, the same non-Functional Requirements are used for multiple Functional Requirements. Since interpretation of the same non-Functional Requirement varies among different Functional Requirements. For instance, although *performance* means *"timely error resolution"* for Data Collection function, it means *"periodic quality assessment"* for DQ Assessment & Improvement function. In the following sections, we explain the system specifications in detail including the activities that should to be performed to address each specification.

4.3.1. Data Collection

1. **Extract, Transform and Load (ETL):** This is the initial phase of all data warehousing activities. It consists of extracting data from different data sources, transforming data based on operational needs (cleaning, converting,..etc.). And loading data to the target system (most likely to a Data Warehouse). In this phase, collected data should include Key Data Items (KDIs)[13] which are identified together with business units. In addition, not omitting any relevant data on material information[14] is critical as it would distort the image of the insurer according to CP 43: "*In case of a lack of information, data can be considered as complete only if such deficiency can be justified as immaterial."* Also, *reliability*, of the collection process should be provided by its transparency. The source of data should be traceable by the regulators.

2. **Data Definitions:** In the regulatory documentation, data definition is described as: "*Definition of the data comprises the identification of the needs in terms of data, a detailed description of the items that should be collected and the eventual relations between the different items.*"[CP 43 – 3.34]. This step is aiming to identify the scope of data collection activity as a scope too wide can lead to errors in model calculations.
   The documents that include data definitions should be kept up to date based on changes in the computer systems and model calculations. Therefore, using manual processes in documenting the definitions, such as creating data dictionaries may fail in the long term. Creating an automated process linked to the data warehouse system where all data is collected is beneficial. For instance, adding a definition tag to data items which include *a data definition code*, and only importing data which has the correct tag.

[13] Refers to required individual data items which will be used in Solvency II calculations.
[14] Material Information: Any information about a company or its products that is likely to change the perceived value of a security when it is disclosed to the public.

TABLE 5. SOLVENCY II DATA QUALITY MANAGEMENT SYSTEM SPECIFICATIONS

	Functional Requirement	Non-Functional Requirement	System Specification
1	Data Collection	• **Performance:** Timely error resolution. • **Reliability:** Transparent, documented collection process, traceable data between source and target. • **Security:** Access level permissions on collected data. • **Usability:** Collecting appropriate amount and best fit data. • **Flexibility:** Collection methods which are adaptable to different source systems, different data types. • **Cost:** Cost of collecting restricted data vs. collecting comprehensive data.	1. Extract, Transform and Load (ETL). 2. Collect data definitions of required Solvency II data. 3. Error Correction via Data Deficiency Management. 4. Validation of DQ. 5. Business Unit Sign-Off via Data Delivery Agreement. 6. Storing data to allow historical analysis.
2	DQ Assessment and Improvement	• **Performance:** Periodic quality assessment on prioritized KDIs. • **Reliability:** Transparent, documented assessment process. Use of Expert judgment should be justifiable. • **Usability:** Amount of data should be input for assessment. “Best fit” should be measured via both statistical and contextual measurement, and also expert judgment. • **Flexibility:** New data types should be integrated simultaneously into assessment process when required. Also extensible to new dimensions and metrics. • **Cost:** Cost based evaluation of Quality Improvement activities. Value Based prioritization of KDIs which will be assessed.	1. Periodic data quality assessment based on identified quality dimensions (indicators). 2. Metric development process will be implemented, to measure the following aspects: a) Structural data quality b) Contextual data quality 3. Using Expert Judgment for assessment in a structured way. 4. Adopting Quality Improvement Strategies based on their cost. 5. Development of a Cost Matrix. 6. Development of a Data Value Measurement table.
3	DQ Monitoring	• **Performance:** Continuous quality monitoring. • **Reliability:** Reporting monitoring results, agreement with Business Units on what to monitor. • **Flexibility:** New data types should be integrated simultaneously into the monitoring process when required. • **Cost:** Value Based analysis of items which will be monitored.	1. Continuous data, data flow and data interface monitoring. Careful selection of data items that will be monitored based on Business Unit input and the Data Value table. 2. Reporting monitoring results.

TABLE 5. SOLVENCY II DATA QUALITY MANAGEMENT SYSTEM SPECIFICATIONS

	Functional Requirement	Non-Functional Requirement	System Specification
4	Data Deficiency Management	• **Performance:** Timely error resolution. Timely root-cause analysis. • **Reliability:** Providing “single source of truth” within entire data flow. • **Security:** Only error correction at the source system by the data owner. • **Flexibility:** Easy integration of new data sources and data types. • **Cost:** Cost optimum Deficiency Management using the Cost Matrix.	1. Implementing Data Error Handling process and procedures. 2. Error resolution. 3. Propagating data updates within the data flow. 4. Identifying root-cause.
5	Data Quality Governance	• **Performance:** Keeping process and procedures up to date and aligned with regulatory changes. • **Reliability:** Solid management approach required to avoid organizational uncertainties. • **Flexibility:** Agile governance to remain compliant with regulatory changes and new business activities. • **Cost:** Seeking optimum cost for each governance activity.	1. Monitoring regulatory changes. 2. Identifying required process and processes, roles and responsibilities. 3. Identifying Data Quality related risks via Risk Management. 4. Identifying Data Quality related costs via Cost Schema.

3. **Error correction via Data Deficiency Management:** In some cases, collected data may not meet the standards of identified criteria. The reasons for such deficiencies are as follows [CP 43]:
   a) Reasons related to the nature or size of the portfolio (such as having limited amount of historical claims data).
   b) Reasons related to deficiencies in the undertakings' internal processes of collecting, storing or validating data quality (such as IT mistakes, high cost of collecting and maintaining data).
   c) Reasons related to deficiencies in the exchange of information with business partners in a reliable and standardized way.

   In correcting those errors, having a standardized approach utilized via Data Deficiency Management is important to have successful error correction process.

4. **Validation of Data:** Validation of Internal Model is explained in detail in CP 56: *"Validation process should not only be applied to calculation of SCR (Solvency Capital Requirement), due to the broad scope of Internal Model used for SCR calculation, it should also be applied to qualitative and quantitative processes of the model including data"*. However, insurance companies are responsible for implementing their own way of data validation and making clear for the regulators what standards are used for the validation. The author proposes the following activities for the validation process:
   - Validating the quality via application of quality criteria and respective dimensions.
   - Expert judgment is also recognized as an important tool in Solvency II. Especially where the collected data is not sufficient for risk assessment, validation of data via expert judgment is a beneficial aid for the risk evaluation process [CP56 - 5.3.3.5].
   - Storing and maintaining data an appropriate amount of time. Historical data allows validation of actual data.
   - Via reconciliation of data with other reports that used for different purposes [CP43 – 1.3].
   - Comparing internal data with data provided by external resources.

5. **Business Unit Sign-Off via Data Delivery Agreement:** Once the data is delivered to a centralized system, such as a data warehouse, a Data Delivery Agreement (DDA) should be signed between data owner (Business Unit representative) and data system owner (IT or Data Management representative) to verify content and quality of delivered data.

6. **Storing data to allow historical analysis:** Collected data should be stored for an appropriate amount of time to allow historical analysis which is used as a validation method as well.

4.3.2. Data Quality Assessment and Improvement

1. **Periodic data quality assessment:** Data quality assessment activities should be performed on a regular basis on KDIs using standardized metrics. For effective assessment, KDIs should be prioritized according to their value for the Solvency II calculations (for instance; how their absence or poor quality would affect the corresponding calculation result). *A Data Value Matrix* should be used in prioritizing KDIs in order to assess their quality.

The Solvency II directive introduces three criteria (dimensions) to assess data quality: Accuracy, Completeness and Appropriateness. However, definition of each criterion is notably wide which makes measuring with a single metric impossible. Therefore, the author proposes transforming each criterion to an extensive dimension list to be able to measure the various requirements stated in CP 43 (Table 6).

In dimension selection process, the intuitive approach described earlier is adopted. The main reasons for this; (1) High level dimensions are already given by the Solvency II directive (criteria). (2) Solvency II definition of each criterion includes clues of dimensions which should be used. For instance, the following statement indicates *currency* dimension; "*recording information should be in a timely manner*". (3) Advised assessment levels of the criteria require using specific dimensions for measurement, such as *granularity* dimension. (4) Characteristics of the data (transactional) that is be used in Solvency II require using specific dimensions for measurement, such as *volatility* dimension.

Majority of these dimensions are introduced earlier in DQ literature by the well-known studies [25] [26], except *proportionality* dimension. In Table 6, definitions of the proposed dimensions from academic literature and the Solvency II documents are shown. Since the proposed dimensions are not explicitly mentioned in the regulatory documents, they do not have exact definitions in the Solvency II documents. Therefore, their Solvency II definitions are interpreted by the author.

2. **Metric development process:** Internal metrics should be developed based on data type and quality dimensions. Metrics should measure:
   a) Structural data quality: Absolute standard measurement, disconnected from a specific usage [69]. This approach refers to an objective measurement of quality using statistical techniques.
   b) Contextual data quality: Data quality assessment based on intrinsic values of data, such as purpose of data, conceptual business value associated with the data and specialties of the decision-maker. Therefore, it refers to a subjective measurement of quality.

3. **Expert Judgment:** Use of expert judgment for Internal Model calculation is outlined in a specific policy [CP56 – 5.3.3.5]. According to the policy, *"In general, the more data quality and data availability is compromised, the greater the extent to which undertakings rely on expert judgment".* However, the regulators also recognize that, even in situations where a lot of data is available about the risk, there is still need for expert judgment. For example *"in selecting the data to use; selecting the time period of the data; adjusting the data to reflect current and future conditions; adjusting for outliers and adjusting industry data to reflect the insurer's circumstances".* Therefore, expert judgment is actively encouraged by the regulators including the following recommendations for the insurers [CP 56] : (1)
   - *Based on data monitoring results, document all instances in which data quality may be compromised;*
   - *Justify, explain and validate the use of expert judgment when related to data;*
   - *Document the inputs and assumptions on which expert judgment is based, as well as the methodology applied in the generation, use and validation of expert judgment.*

TABLE 6. DEFINITIONS OF THE PROPOSED DIMENSIONS

SII Criteria		Proposed Dimension	Interpreted Definition from Solvency II [CP43]	Definition from Data Quality Literature
Accuracy	1	Syntactic Accuracy	Data free from material mistakes, errors and omissions.	Closeness of a value v to the elements of corresponding definition domain D [3].
	2	Currency	The recording of information should be performed in a timely manner.	How promptly data is updated [3].
	3	Traceability	The insurer should be able to demonstrate usage of data through operations including cross-checks.	Ability to verify the history, location and usage of an item by means of recorded identification [70].
	4	Credibility	Judgment of trustworthiness of data based on analysis of the underlying liabilities, the company and portfolio's experience and relevant qualitative information such as consistency with available market data.	Refers to subjective and objective components of the believability of a source [45].
	5	Consistency	Integrity of the same internal data within different points of time. And alignment of internal data with external data.	Data consistency is combination of validity, accuracy, usability and integrity of related data between applications and across an IT enterprise [3].
	6	Volatility	Using up-to-date information is essential[15].	The length of time data remains valid or frequency with which data vary in time [3].
	7	Timeliness	Calculation of best estimate should be based on up-to-date information.	How current data for the task at the hand [3].
Completeness	8	Completeness	The data covers all the main homogeneous risk groups in the liabilities' portfolio.	The extent to which data are of sufficient breadth, depth and scope for the task at hand [1].
	9	Granularity/ Depth of Data	The detail level of information should be such that it allows for identification of trends and understanding of behavior of underlying risks. Also it should allow for application of adequate provisioning methodologies[16].	Granularity of data refers to scale or level of detail in a set of data [31].
	10	Historical Data	The available, reliable, historical records for a data item.	Past periods data. Usually used for forecasting future data or trends [31].
	11	Proportionality	While it would be expected that less data is needed to evaluate simple risks, more data should be available where the nature, scale and complexity of the underlying risks is high.	Properly related in size, degree or other measurable characteristics [71].
	12	Variety of data / Heterogeneity	How heterogeneous the portfolio is[17].	Complexity or variability of data [72]. Data distributed in various resources and represented with different formats [73].
Appropriateness	13	Relevancy	Data suitable for intended purposes such as relevant to the portfolio of risks being analyzed.	Data which is applicable to the situation or problem at hand that can help solve a problem or contribute to a solution [31].
	14	Semantic Accuracy	Consistency of data when it is compared to different data sources.	The closeness of the value v to the true value v' [3].
	15	Amount of Data	Quantity of data used in calculation[18].	Quantity of data takes place in data sets.

[15] This dimension is not completely included in the regulatory text as it is more related to understanding data characteristics. But it is necessary to measure to provide information on how often data changes within specific time period.

[16] For instance, if run-off triangles are used to calculate the best estimate, it is necessary to record separately all payments and the date at which the payment was made, instead of just the total amount paid [2].

[17] More heterogeneous the portfolio is, the more detailed the data should be.

[18] In case of a lack of information, data can be considered as complete only if such deficiency can be justified as immaterial. The assessment should also include an analysis of whether the undertaking's information is comprehensive and a relative comparison with other data for similar lines of business and/or risk factors [2].

4. **DQ Improvement:** Improvement of DQ is one of the continuous activities in DQM system. Once DQ level is identified via assessment, data items should be prioritized to improve based on their DQ level and business value (using a Data Value Matrix, see step 6). In this step, choosing appropriate improvement strategy considering (a) Cost of the improvement activity and (b) Organizational goals is essential.

5. **Development of a Cost Matrix:** An internally developed cost matrix will be used for the following purposes within the data quality management system: (a) Evaluation of quality improvement activities, (b) Cost optimum Deficiency Management (c) Cost optimum DQ Governance.

   Inspired by the well-known cost matrices (English classification, Loshin Classification, Eppler - Helfert classification and their comparative classification in [3]) an example Cost Matrix developed towards Solvency II based on the previous requirement analysis results (Appendix V). Due to large variations in organizational goals, focus and business activities among the insurers, customized solutions are needed in this step. Therefore the insurers should consider their organizational needs and emphasis in adopting such a cost matrix.

6. **Data Value Measurement:** Measuring quality of all data items generated for Solvency II and monitoring their quality continuously are costly activities. Organizations need to prioritize data items subject to the regulation to implement cost effective and efficient measurement and monitoring processes. In an enterprise environment, many different as well as overlapping data items are critical for various business units. Therefore, enterprise level, structured data value measurement is required for an appropriate prioritization.

   Our definition for *Data Value Measurement* is; to *understand which business processes use a particular data item, how important the item is for that process and how its absence affects the process.*

   Considering the intrinsic characteristics of value measurement, it is difficult to represent data value in a tangible, comparable way for the decision makers. Therefore, the data value should be associated with its impact on corresponding business process as in *"Business Impact Classification of Poor Data"* approach introduced by Loshin [74]. Base on Loshin's classification, an example impact classification table is developed in line with the Solvency II system processes (Appendix VI). Using the example as a reference, companies should develop a custom impact scale for each impact category.

4.3.3. Data Quality Monitoring

1. **Data, Data Flow and Data Interface Monitoring:** DQ monitoring as a system function includes data flow and data interface monitoring as well as monitoring individual data items. Data flows and interfaces are the critical elements of a data quality management system as much as data items, as they play an important role in data generation and transformation.

   According to the regulatory documentation, restricting data in monitoring minimizes the costs. Therefore knowing what to monitor is beneficial for the organization. However, organizations need to establish a balance between restricting data and using comprehensive data to prevent having incorrect results in the monitoring activity [CP 56 – 5.3.3.1]. A value based evaluation of data items is the right solution in selecting suitable monitoring scope that will benefit both the organization and the monitoring results.

   Although most of the data items and data flows are specific to each organization, some generic items and flows are derived below. Below information should be used as a reference in defining which data items, data flows and interfaces to monitor within the entire Solvency II process.

   - Solvency II data sources: Life/non-Life/Health Contracts & Claims, Asset Data, External Market Data, Accounting Data, Counterparty Information, Operational Losses, Solvency Balance Sheet, Capital Requirements (Figure 6).
   - Data sources used for activities such as Economic Scenario Generation, Life Liabilities Model Points, Assets Model Point, Experience Analysis and Assumptions.
   - Activities providing input for the regulatory calculations: e.g Cash Flow Projection Life, Cash Flow Projection Assets, Cash Flow Projection non-Life/Health.
   - Calculation results consolidated and aggregated under appropriate risk groups.
   - Mandatory internal and external reports.

2. **Reporting:** Monitored DQ results should be reported in different organizational levels to create awareness of DQ levels of different data items and surrounding data flows and processes. Some of these reports should be used in order to prove the transparency of data flows to the supervisory authorities.

4.3.4. Data Deficiency Management

1. **Data Error Handling:** Process and procedures should be implemented to determine how the errors will be logged, and resolved or escalated if they cannot be resolved. Service Level Agreements (SLAs) should be developed towards standardization of error handling.

2. **Error Resolution:** Based on the SLAs between data owner and IT department, who is responsible for error handling, error resolution should be implemented in a timely manner. Error resolution should always be done by the data owner at the source system to comply with *Single Source of Truth (SSOT)* principle.

Considering the complexity of current enterprise level systems, deploying single source of truth (SSOT) based system designs becomes important. The term *SSOT* is used in Data Modeling and Data Warehousing terminology and refers to storing every data element exactly once and having links to that data element within the enterprise by reference only [75]. Therefore, when the data element is updated, the updates will be propagated.

3. **Data Updates:** Following error correction at the source system, the corrected data should be propagated in the entire data flow to prevent discrepancies between source and target systems. Those updates should be monitored and reported in order to provide regulatory evidence on the data update activity.

4. **Root-Cause Analysis:** Root- cause analysis should be used to identify the reasons of the deficiencies and to eliminate the actual reasons of the error rather than just focusing on the error itself. This preventive approach provides benefits in the long term and mitigates data deficiencies.

   Additionally, during error resolution and prevention, the Cost Matrix introduced earlier should be used to compare the cost of activities and to select the ones with the best cost/benefit ratio. However, in some cases although the error prevention activity is costly, the organization might still need to implement the solution to remain compliant with the regulation. As it is highlighted in CP 43 – 3.22, where the data deficiency is related to insufficient internal processes, the insurer should take appropriate measures to remedy the situation in due course.

4.3.5. Data Quality Governance

Insurers need to establish their own policies on Data Quality, approved by the senior management [CP 56 – 5.150]. Another statement in CP 56 in regard to the data policy is that the policy should be agreed with the supervisory authorities as a part of the internal model approval process. Also, major policy changes shall always be subject to prior supervisory approval.

Based on above information, companies need to develop their custom Data Quality Management structure and supporting procedures carefully considering that some parts of the practice will be subject to the supervisory authority’s approval.

In this section, Data Quality Governance functionality of the system is explained in detail introducing a practical governance model (Figure 7). DQ Governance functionality describes how DQ activities are governed towards Solvency II compliance in an organization, what kind of management structure is required, which roles should be established with which responsibilities, and which policies and procedures should be implemented.

Data Governance Team

The internal definition of "data quality" should be done by the *Data Governance* group who is responsible for identifying all data management related strategies of the organization. That definition should be connected with the organizational level goals and strategies. Some examples of the data management related strategic decisions are; current and future regulatory requirements company needs to comply with; how to use available data for business related decision making to improve the performance of existing business activities as well as to enter new business lines, etc. Assessment of the organizational risks carried because of poor data should be the part of general risk management activities in order to provide input to the strategic decision making process. Also, required high-level policies and procedures, roles and responsibilities should be identified by this group. High level executives such as COO, CIO should be members of the group to guarantee organization wide management support.

Data Management Team

Data Management team is in the middle layer in this multi-tier management structure. While performing all data management related activities (such as implementing policies and procedures identified by governance group, assigning roles and responsibilities, identifying and monitoring DQ attributes on high level), it should also establish communication with the business units who produce and use data. The team should realize user awareness on data quality concepts organizing trainings for business units which will provide a long term increase on data quality. On the other hand, business units should also provide full management support to Data Management team as well as timely input on changing business activities. A Data Quality Service Level Agreement which would be signed between Data Management team and each business unit should formalize their interaction as well as data content and its quality level which will be delivered by the business unit.

Additionally Data Management team should work closely with IT services and provide organizational requirements of Data Quality service on behalf of the business units and Data Governance groups. In return, IT services should implement required infrastructure, most likely a data warehouse system, and provide technical support of the infrastructure.

External Parties

External parties refer to several non organizational groups which interact with the DQM system. Some of them provide input to the system such as introduced regulations by the regulators or external data by the data providers. Some of them use the system outputs such as supervisory authorities, shareholders and customers who access various reports.

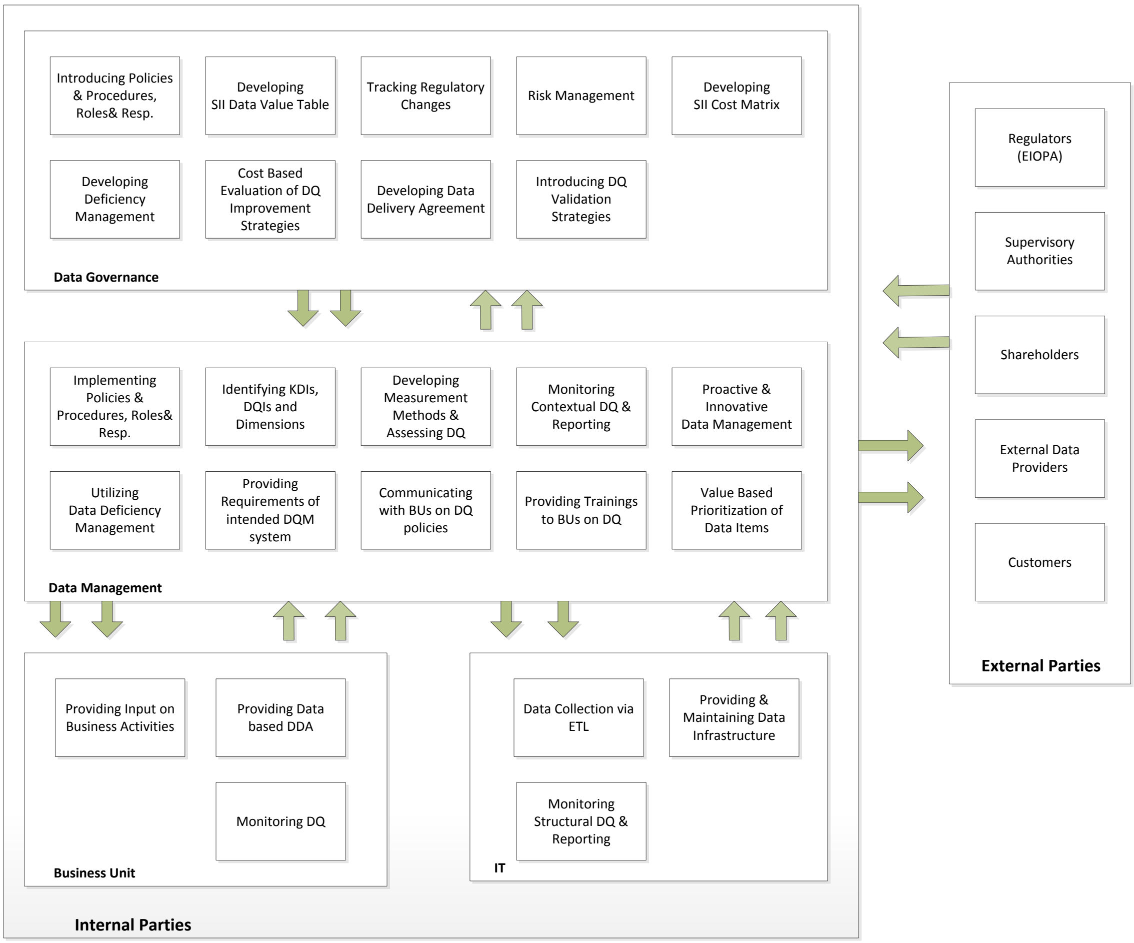
Introducing Policies & Procedures, Roles& Resp.
Developing SII Data Value Table
Tracking Regulatory Changes
Risk Management
Developing SII Cost Matrix
Developing Deficiency Management
Cost Based Evaluation of DQ Improvement Strategies
Developing Data Delivery Agreement
Introducing DQ Validation Strategies
Data Governance
Implementing Policies & Procedures, Roles& Resp.
Identifying KDIs, DQIs and Dimensions
Developing Measurement Methods & Assessing DQ
Monitoring Contextual DQ & Reporting
Proactive & Innovative Data Management
Utilizing Data Deficiency Management
Providing Requirements of intended DQM system
Communicating with BUs on DQ policies
Providing Trainings to BUs on DQ
Value Based Prioritization of Data Items
Data Management
Providing Input on Business Activities
Providing Data based DDA
Monitoring DQ
Business Unit
Data Collection via ETL
Providing & Maintaining Data Infrastructure
Monitoring Structural DQ & Reporting
IT
Internal Parties
Regulators (EIOPA)
Supervisory Authorities
Shareholders
External Data Providers
Customers
External Parties

FIGURE 7. SOLVENCY II GOVERNANCE MODEL

In summary, initially centralized, later on a combined (centralized and distributed) management strategy should be adopted in Solvency II Data Governance model. Centralized management which will be provided by Data Governance Group is needed especially at the beginning, to ensure defining a single strategy across business units, establishing organization wide standards and gaining executive level management support. Then an organization wide Data Management Group should be utilizing and implementing those strategies which are introduced by the centralized management. However, once data quality concepts are understood by the organizational layers, centralized and distributed approaches should be combined in order to give a certain level of autonomy to the business units in maintaining their own data quality.

Chapter 5. Analysis of Data Quality Management Methodologies

In this section, some of the well-known DQM methodologies are analyzed from the Solvency II perspective. While answering the research questions 3 and 4, the aim of the analysis is to investigate the usability of a methodology for the Solvency II DQM system. Thus, the target is to find the degree of adoptability of the methodology by an insurance company in implementing a Solvency II compatible DQM system.

5.1. Comparison of Methodologies

In Table 7, the phases and high level activities of the methodologies are outlined. These activities are evaluated from the Solvency II perspective in the last column. Although some limitations are identified as a result of evaluation, it is hard to reject the use of any of the methodologies for a Data Quality Management System implementation during a Solvency II project, since none of them have significant incompliance issues with the regulatory requirements based on available information.

TABLE 7. ANALYSIS OF DQM METHODOLOGIES

	Phases [40] [53]	**General Description**	**Evaluation from SII Perspective**
TDQM	1. Definition a. Data Analysis b. Data Requirement Analysis c. Process Modeling 2. Measurement: Measurement of Quality 3. Analysis: Identification of error causes 4. Improvement: Selection of strategies and techniques	1. Uses the principals of TQM 2. Adopts Information Product (IP) approach 3. Introduces IP-MAP language for information process modeling , IP-MAP has been extended towards UML 4. Practical experiences are available from various businesses 5. Roles for each phase are also provided 6. Provides guidelines on how to implement the methodology 7. Adopts survey based dimension identification approach 8. Focuses on operational side of DQ activities	1. Dimension identification process should combine regulatory requirements and user opinion 2. Evaluation and comparison of improvement activities should be based on a Solvency II Cost Matrix 3. Measurement of Data Items should be prioritized based on their impact on Solvency II data process
TIQM	1. Assessment a. Data Analysis b. DQ Requirement Analysis c. Measurement of quality d. Evaluation of costs 2. Improvement a. Identification of error causes b. Design of data improvement solutions c. Process control d. Process redesign 3. Improvement Management & Monitoring	1. Introduced to support data warehouse projects 2. Assumes that while data is consolidated at data warehouse, errors and heterogeneities are eliminated 3. Provides extensive Cost-Benefit analysis from managerial perspective for DQ improvement 4. Focuses on economical side of DQ activities	1. Error correction might be required at any stage of data flow including data warehouse 2. Data Items should be prioritized based on their Impact on Solvency II data process
AIMQ	1. Measurement 2. Analysis and Interpretation of Assessment	1. Focuses benchmarking for quality assessment as an objective and domain independent technique 2. GAP analysis is used for benchmarking 3. Introduces PSP/IQ model to classify dimensions according their importance from user's perspective	1. Focuses assessment part, no information is provided on improvement activities 2. Dimension importance ratings are done via surveys which is not sufficient when the requirement is to comply with regulations 3. Assessment is done based on benchmark analysis which requires information on other company's best practices. That assessment technique could be used just as addition to solid statistical measurement methods which should be implemented

TABLE 7. ANALYSIS OF DQM METHODOLOGIES

	Phases [40] [53]	**General Description**	**Evaluation from SII Perspective**
CDQM	1. State Reconstruction 2. Assessment a. Data Analysis b. DQ Requirement Analysis c. Measurement of quality 3. Improvement a. Identification of error causes b. Selection of strategies and techniques d. Evaluation of costs	1. The most recently developed generic methodology 2. Provides normalization techniques to improve DQ by comparing data 3. Addresses instance level heterogeneity 4. Includes Cost-Benefit analysis	1. Measurement of Data Items should be prioritized based on their impact on Solvency II data process 2. SSOT principal should be integrated within Error Localization and Correction activities
ORME - DQ	1. DQ Risk Prioritization a. Reconstruct the state building correlation matrixes 2. DQ Risk Identification a. Loss event profiling and evaluation of economic losses b. Selection of critical processes c. Selection of critical data sets and data flows 3. DQ Risk Measurement a. Qualitative and Quantitative DQ assessment b. Approx. evaluation of loss events 4. DQ Risk Monitoring a. Evaluate DQ dimension values periodically based on target values	1. Developed to support Basel II’s operational risk evaluation approach. 2. Uses CDQM methodology s a reference.	1. Development of organizational level best practices should be integrated into Risk Measurement phase 2. Only methodology that includes selection of critical data sets and data flows based on their Risk level

5.2. Proposed Extensions for Methodologies

As mentioned earlier, since all analyzed methodologies require some level of customization during implementation, any of the methodologies can be implemented during a Solvency II project. However, all methodologies need customization to align with the regulatory requirements. Table 8 represents the proposed extensions to the methodologies compared in Table 7. The first column of the table includes corresponding functional requirement for each extension to visualize the location of the activity within the Solvency II Data Quality Management system.

TABLE 8. PROPOSED EXTENSIONS TO METHODOLOGIES

Solvency II DQM System-Functional Requirement	Extension
Data Collection	1. Validation of DQ
Data Quality Assessment & Improvement	2. Extension of the dimension list via Intuitive Approach 3. Developing metrics for DQ assessment based on the dimension list 4. Structured expert judgment 5. Solvency II – Cost Matrix 6. Selection of appropriate data items
Data Quality Monitoring	7. Location of quality checks in data flow 8. Data Tags 9. Error capturing at the earliest possible stage
Data Deficiency Management	10. Resolution of Deficiencies 11. Change Management 12. Single Source of Truth

Some of the proposed extensions are explained in Section 4.3, such as Validation of DQ, Structured Expert Judgment, Cost Matrix and Single Source of Truth. Therefore, they are not explained in this section again. Detailed information on the remaining *"extensions"* is listed below:

Extension of Dimension List: The methodologies include several dimensions described in DQ literature, especially based on Wang and Strong's study [25]. Since included dimensions don't cover all the dimensions we proposed, extension of the dimension list is required for each methodology.

Interestingly, according to Batini et al. [40], only CDQM methodology among the other well-known methodologies is extensible to include new dimensions and metrics additional to the dimensions were given as a part of the methodology. However, we don't see any clear evidence of that statement within the article and no indication of such limitation considering the published practices of the methodologies. Therefore, we conclude that the proposed dimensions could be used by the all well-known methodologies we analyzed.

Metrics: The methodologies don't introduce metrics to measure all dimensions proposed earlier, as the metrics could vary based on data type, internal definition of the dimension and organizational goals. Some of available metrics for the three Solvency II criteria are introduced in Chapter 3. For the extensive dimension list proposed earlier, organizations need to develop their internal metrics based on their internal targets. However, while developing customized metrics, a combination of subjective (user opinion) and objective (statistical approach) measurement techniques should be considered. Appendix IX represents some of available measurement techniques that could be used for the proposed dimensions.

Selection of Data Items: According to Solvency II documentation, evaluation of the three criteria should be done at a fine level of granularity such as “individual item level” (Table 7). The methodologies mention database and data flow level DQ assessment; however how the data items will be selected to be assessed is not addressed.

In systems specifications section, a method to measure data value is proposed based on impact analysis of poor data: *Data Value Measurement*. The value of data for organization provides input to data item selection process in order to assess quality only for the most valuable data.

Quality Checks: Data starts its long journey with a Policy Administration system at an insurance company. Since large companies usually use legacy policy administration systems, data has to be extracted and transformed in order to be imported into contemporary software systems. Due to complexity of data flows, multiple quality checks should take place aligned with both data processes and business activities. Quality checks should be done *after* each significant data process such as “extracting data” as well as *before* every critical business activity such as preparing Technical Provisions in order to guarantee its quality for the activity.

Data Tags: The origin of data used by actuaries is sometimes unknown, as the actuaries use mixed data sources both internal and external. Therefore, using a data item level tagging system could be beneficial to trace the data. A similar approach was previously recommended by Wang et al. [39].

Error capturing: While implementing a quality monitoring structure for the entire data flow, capturing data deficiencies at the earliest stage, such as Policy Administration System database, should be the target. Considering similarities between data production and manufacturing; if quality issues are captured in the earlier phases of the production cycle, the defects will be less costly to fix in the long run since the inspection, rework and rejects will be avoided.

Resolution of Deficiencies: The problems on verification of data quality criteria should be resolved within an appropriate time frame and any data limitation should be documented properly including description of remedies and assignment of responsibilities according to Solvency II. Clearly an error logging and monitoring system is required. CDQM mentions “error correction” as an improvement activity in the context of process improvement with no reference to the location of the activity. To address these issues, as a practical approach, monitoring data quality deficiencies and implementing their resolution should be integrated into Change Management and Help Desk Problem Management systems that are available in most of the companies’ IT management structure.

Change Management: In some cases, adjustments could be applied to data by actuaries to improve goodness of fit according to the regulation. However, the record of these changes should be kept in a Change Management system.

Chapter 6. Field Study at a Dutch Insurance Company

In addition to the theoretical part described in the previous sections, nine-month field study was conducted in 2012 at one of the largest insurance companies in Europe located in the Netherlands. The Insurance Company (INSC) is part of Dutch Financial Services Group (DFSG) which operates internationally in the banking and insurance sector. DFSG also owns an insurance business unit which needs to comply with the Solvency II regulation. Consequently, INSC runs a Solvency II project in parallel with DFSG's insurance department. Mainly using Data Management standards set at the corporate level by the task forces, INSC has been developing its custom approach to Data Management with help from several consultancy companies. The names of the two organizations will not be disclosed in this report due to the confidentiality of the information collected.

The field study aims to obtain practical information regarding the implementation of the Data Quality Management concepts in a real environment. It also provides an opportunity to understand the insurance business, its data sensitivity, and how insurance companies are coping with the Solvency II requirements.

Furthermore, this part of the research has contributed to the overall objective in order to provide realistic and applicable guidance to insurance companies for Solvency II compliance, rather than just providing theoretical inputs. As a result, the requirements analysis results presented in Chapter 4 are utilized by the Insurance Company in the implementation of their Data Quality Management structure.

The following sections combine both the author's observations of the company's Solvency II project and a review of internal policies developed for Solvency II compliance. The chapter also includes a data analysis section to give a practical example of the measurement and monitoring system specifications described earlier in Chapter 4.

In total 12 interviews were conducted at INSC. The list of interviewee roles and sample interview questions are available in Appendix I. The first interviews were mostly aiming at understanding the Solvency II project organization and what the project is expected to deliver from a Data Management perspective. Later on, interviews became more focused on Data Quality aspects of the project. All interviews were semi-structured; prepared questions were used as well as unplanned questions based on the interviewee's feedback. Notes were taken during the interviews and documented right after the interviews. No structured analysis was applied to the interview notes since the aim was to collect information on the company's operations in relation to Solvency II.

In addition to the interviews, a second valuable information source was the company's intranet. The company has several Share Point sites used as the document repository for the Solvency II project. These sites are quite up to date and they are used on a daily basis to communicate project developments to the project members.

6.1. Background of the Solvency II Project

The Solvency II project was initialized by establishing several task forces at the corporate level to address the different requirements of the directive. The task forces are aiming to concentrate the knowledge around the various Solvency II topics bringing internal specialists and external consultants together. Two task forces mentioned below, in regard to their contribution to Data Management, started to work around early 2011.

Data Architecture, task force is aiming to design conceptual and logical data models based on *IBM – Insurance Information Warehouse* (IIW) tool's concepts and definitions. Although IIW will not be used as the data warehouse software, its modeling standards are selected to define the data architecture due to the predefined Solvency II content included in the design module of the tool. Currently the team is in the process of translating the definitions of the IIW data model into the SAP Business Warehouse (SAP BW) that will be used as the data warehouse software. Additionally, the data architecture task force is responsible of creating the standards of the Data Dictionary[19] that will include the technical details of all Solvency II data items. The team is also documenting the Solvency II data interfaces and the data flows [76].

Data Quality, task force created the first DQ policy describing the governance and DQ framework in 2011 at the corporate level. The team also analyzed the current maturity level of Data Quality within DFSG. For 2012, the team was planning to achieve total Solvency II compliance by rolling out DQ activities for the remaining processes and data. The team is responsible for specifying process flows within the SII scope, setting the standards of the Data Directory[20], defining Data Quality Indicators (DQIs), and introducing DQ measurement and reporting solutions [77].

[19] A repository of information containing the unique definitions of data [79].

[20] An inventory of all data used within the SII chain along with its characteristics, the processes they are used for and the controls applied on those data. Has a link with the Data Dictionary, which provides the unique definitions for the data in the Data Directory [79].

6.1.1. Data Management Team

The Data Management team has been brought together to address the Solvency II data quality requirements within INSC. Currently the team performs totally in the context of the Solvency II project. However, in the future the scope of the team may be expanded to include other data elements from different business units.

This field study started soon after the team started its operations with many uncertainties regarding the scope and responsibilities. Initial difficulties were in creating a company-wide awareness of DQ and Data Management and defining areas of responsibility. Following the corporate level *Data Governance and Quality Management Policy* published in May 2012, INSC's Data Management team started working on a policy where the team goals and activities will be outlined. The section below is taken from the draft version of the policy [78]:

"The ultimate goal of Data Management is to ensure that the data in the organization is trusted and reliable, and is a true asset to the organization. It achieves this by implementing measures and controls, which ensure the quality of the Data as well as the quality of the data integration process."

According to the same document, the team's activities will concentrate around three topics:

- Data Governance: Implementing clear roles and responsibilities for data owners and data stewards. Using Data Delivery agreements between data suppliers and data receivers.
- Data Definitions: Creating Data Dictionaries to guarantee uniform understanding of data within the enterprise.
- Data Quality: Measuring data quality to make sure it complies with the standards.

6.1.2. Internal Roles and Responsibilities

Below are the internally identified roles and responsibilities at INSC in regard to Data Management [78].

Data Governance Board is chaired by Data Management Office and is responsible for compliance with the regulatory and corporate level practices. Data owners are the members of the board to support continuous improvement of Data Management activities.

A Data Owner is an individual responsible for collecting data and keeping this data accurate and complete. The Data Owner should ensure that the data and processes within his scope have sufficient quality. The quality of data should be assessed and preserved within the entire process by the data owner with support of Data Steward and Data Custodian.

Data Steward is the representative of data owner who knows the business value of data. He reports and resolves DQ issues, ensures DQ policy compliance, and sets DQ requirements.

Data Custodian is responsible for the technical environment and database structure, ensuring that data remains unaffected in storage and accessible by the authorized Data Owners and Data Stewards (usually this role is fulfilled by IT Operations).

Data Management Office is responsible for defining and implementing Data Management policy and administration of Data Delivery Agreement, Data Definitions and Data Flows, and Data Quality Indicators.

6.1.3. Data Quality Assessment

The following definitions are obtained from the corporate level policy document to give an idea of how Solvency II criteria are interpreted within the organization [79].

"Accuracy refers to whether the data correctly records the business object or event it represents. It has two requirements: (1) it must be the right value and (2) it must represent the value in a consistent form with all other representations of the same value."

"Completeness refers to whether the data set contains all required elements."

"Data are considered to be appropriate if they are suitable for the intended use, and relevant to the portfolio of risks being analyzed. Appropriateness also refers to the robustness of the data - whether it has sufficient granularity to identify trends, and to provide a full understanding of the behavior of underlying risks."

In addition to the Solvency II' given criteria, the following two criteria are also defined by DFSG:

"Data is considered consistent (integral) when similar metadata is used for data used by users or processes. Moreover, there should be links between related data allowing for reconciliation."

"Data must also be accessible and available to the different stakeholders. In addition, it is also required that data is auditable and any modifications to data need to be traceable and a clear link from source to output must be available."

Assessment of above quality criteria is done via *Data Quality Indicators (DQIs)*. DQIs are defined as "*data controls that are used to measure the quality of data items"* [79]. DQIs are used to detect data quality issues on *Key Data Items (KDIs)* and to resolve the problems at an early stage in the process. DQIs should be defined with cooperation of data owners and users, and should be documented in a *Data Directory* by the data owner. INSC's DQIs consist of four levels:

1. Level 1 - Definition: Ensure each data item is clearly defined including its granularity and flagged as optional or mandatory.
2. Level 2 - Identification: Control that all mandatory fields are filled.
3. Level 3 - Validity: Control that the data field has an acceptable value with right format and within right range.
4. Level 4 - Reasonableness: Identify deviations from expected values via comparison with historical data, benchmarks and other data sources.

As the first step of the assessment process, Data Management team assists the business units, who will provide data for the Solvency II calculations, in preparation of the Data Dictionaries. A sample Data Dictionary is shown in Appendix III. DQI levels are included in the Data Dictionary for each data item to describe how they will be interpreted for specific data.

6.2. A Practical Case

In this section, two of the DQM system specifications identified earlier in Chapter 4 are operationalized in INSC's environment to provide a practical example. First, *Syntactic Accuracy* dimension is measured on the insurance data as a *Data Quality Assessment* practice. Second, application of data quality checks are exercised within the data flow as a *Data Quality Monitoring* practice.

6.2.1. Data Quality Assessment

In Chapter 4, extension of three Solvency II criteria with a comprehensive dimension list is proposed (Table 6). In this section, one of these proposed dimensions, *Syntactic Accuracy*, is assessed. The data obtained from Fire Insurance[21] business unit is used for the measurement. Fire Insurance business unit operates as part of the Non-Life Insurance group within the organization.

In Table 6, two descriptions of *Syntactic Accuracy* are provided:

Interpreted Definition from Solvency II [CP43]	**Definition from the Literature**
Data free from material mistakes, errors and omissions.	Closeness of a value v to the elements of corresponding definition domain D [3].

Combining both definitions, during the assessment we search for the following; *whether the individual data item differs from the pre-defined quality standards or it meets the standards. If it doesn't meet the standards, data is said to be in ERROR state, otherwise it is TRUE.*

[21] *Brand Verzekering* (Dutch)

Then, *Simple Ratio* functional form is used similar to the *free-of-error* measurement described in [45]; ratio of TRUE data items to the whole data items gives us *Syntactic Accuracy* rate of the individual data items within the data set.

The assessment is done in two steps: First, using freeware R application (v2.12.0) developed for statistical data analysis; we analyze the data set. This activity helps us to understand data characteristics, such as average (*mean*), to what extent data differs from the average (*standard deviation*), does it include any unexpected values (*outliers*). Second, we combine our findings from the statistical analysis results and from data consumers (users) in setting *quality standards* for each data item. Then we apply the standards to the data set in order to see the degree to which it meets the standards, meaning how *syntactically accurate* it is.

Statistical Analysis of Sampling Data

The insurance data containing both text and numerical values is stored in databases or spreadsheets. Each data element such as policy number, policy holder information or policy premium reside in a fixed field either within a relational table structure or within a flat table. Therefore, the insurance data maintains the characteristics of the *structured data* which allows us to apply traditional statistical methods.

During the study, extracted files from the *Policy Administration system* (*VTA*)[22] residing on the mainframe are used. A data file is extracted by a *SAS*[23] script on a monthly basis to transfer data to the different application databases. An extracted file consists of approximately 600,000 rows and more than 300 columns. Each row includes all recorded information about one customer's fire insurance policy, such as policy number, start and end date of the policy, policy premium, etc. In this practice, after the extraction, actual policy numbers are replaced with the unique dummy numbers due to the confidentiality of data.

Before the analysis, the majority of the columns are eliminated on the sampling file based on the *Key Data Items (KDIs)*. KDIs are the important data items that are used in Solvency II model calculations. The table below shows the list of KDIs on the subject data. Each column title corresponds to a variable field used within the VTA (policy administration system) software. Also row numbers are reduced to 390,000 for the analysis due to computer system limitations.

[22] VTA : Verzekering Technische Administratie (Dutch) is the Policy Administration system where all insurance policy entries take place.

[23] Statistical Analysis System – SAS Institute Inc.

TABLE 9. SOLVENCY II - KEY DATA ITEMS FOR FIRE INSURANCE DATA

	Column Name	Description
1	polisnr	Insurance Policy Number. Unique number for the each client.
2	ingwyjr	Start Year. Beginning year of the policy. Three digits numerical data. Due to design of the legacy application three digits used instead of four. Therefore, it has to be converted to the four digits year format after extracted.
3	ingwymnd	Start Month. Beginning month of the policy. Two digits numerical data.
4	ingwydag	Start Day. Beginning day of the policy. Two digits numerical data.
5	term	Term Number. The period of time that an insurance policy provides coverage. Two digits numerical data which takes one of the following values 0, 1, 3, 6, 12, 60, 120.
6	standpni	Insurance Premium non-indexed. The premium amount is not linked to a financial index. Numerical data that could take decimal or negative values.
7	standpi	Insurance Premium indexed. The premium amount is linked to a financial index. Numerical data that could take decimal or negative values.

Statistical data analysis using R software is performed only on the two *premium columns* as data values in these columns has significant variation compared to other columns. Insurance premiums are in the ratio scale; therefore we can measure the average (mean) as well as the middle observation (median) [80].

Indexed Premium
Min. 0.00 and Max. 5799.00
Mean **37.83**
SD **81.34022**

Non-Indexed Premium
Min. - 379.80 and Max. 245,800.00
Mean **68.59**
SD ***683.1179***

R commands used for the analysis are listed in Appendix VII. After loading the data set called *Brandmaster* to the application (**read.table**), the number of the loaded rows is checked (**nrow**) and data set is reviewed to verify correct load (**edit**). Then *mean, median, min and max values* of the columns are calculated (**summary**), see above results. These values showed that *Indexed premium* doesn't take any negative value and changes between 0 and 5799. On the contrary, *non-Indexed premium* takes negative values and is distributed between -379.80 and 245,800.

Then, we check *Standard Deviation* for both columns (**sd**). *Standard Deviation* represents how much dispersion from the average (*mean*) exists on a data set. A low *standard deviation* indicates that the data points tend to be very close to the mean; high *standard deviation* indicates that the data points are spread out over a large range of values [81]. We can see that both *Premiums* have high SD which indicates that data sets spread in a large range.

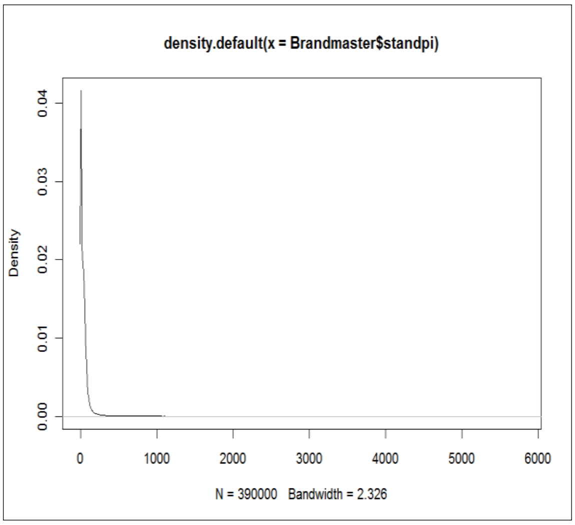

FIGURE 8. DENSITY PLOT OF INDEXED INSURANCE PREMIUM

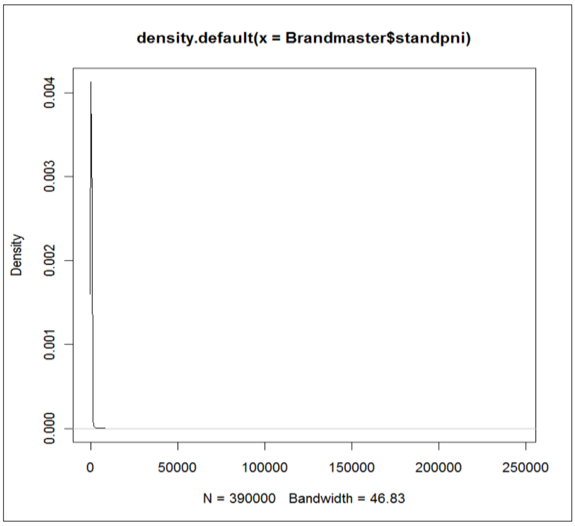

FIGURE 9. DENSITY PLOT OF NON-INDEXED INSURANCE PREMIUM

Figure 8 and 9 are the density graphs of both columns. Both graphs indicate a positively skewed data set where *the mean* will be greater than *the median*. Those graphs also show that the data is concentrated at the lower end of the range, which means there are more data items that take a low value. Value of the *mean* is pulled upwards by the few very high data values which indicate outliers.

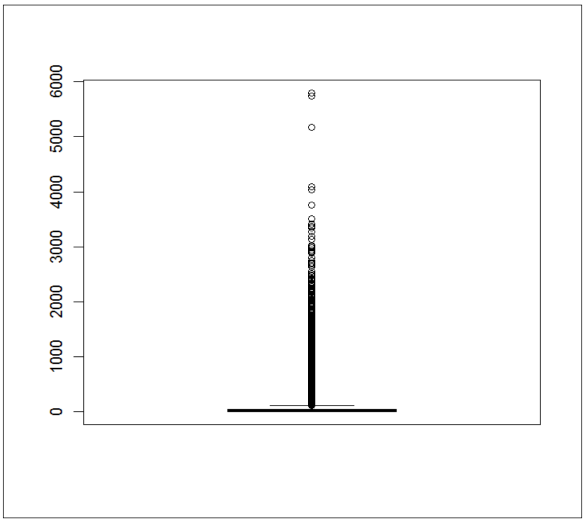

FIGURE 10. BOXPLOT FOR INDEXED PREMIUM

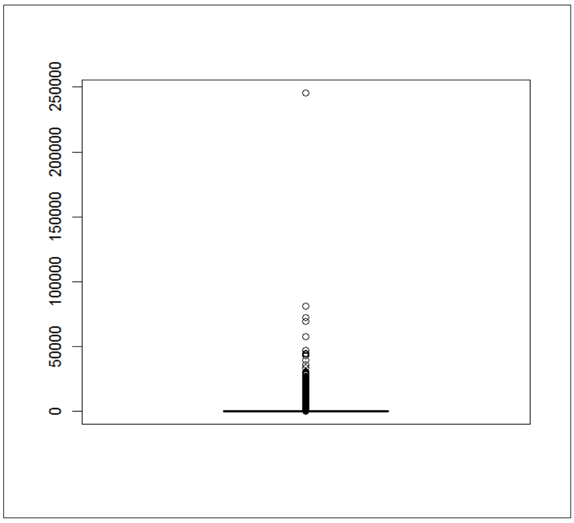

FIGURE 11. BOXPLOT FOR NON-INDEXED PREMIUM

Figure 10 and 11 show *boxplot* of both premiums where we can clearly see the outliers for both data sets (only for positive values). These outliers are the extreme values that could be used as a reference while setting thresholds for the data set.

Applying Quality Standards on Sampling Data

TABLE 10. ORIGINAL DATA FILE

polisnr	ingwyjr	ingwymnd	ingwydag	term	standpni	standpi
a5403	110	3	11	12	0	52.25
a5404	111	1	21	60	73.17	0
a5405	109	9	25	12	-34.03	184.22
a5406	111	6	1	12	0	38.58
a5407	109	10	1	12	0	53.58
a5408	104	10	1	12	0	68.01
a5409	110	9	1	12	0	39.96
a5410	108	5	15	12	0	41.37
a5411	110	9	20	12	132.16	0
a5412	110	2	18	12	18.72	0

Table 10 represents a small part of the original data file. For each column, the following *quality standards* are developed to assess the quality: Data Type, Blank/Filled and Condition. We search these *standards* in the original data file to calculate Syntactic Accuracy. We also add a threshold column to Table 11 based on the outliers we identified in the Boxplots (Figure 10 and Figure 11).

TABLE 11. QUALITY STANDARDS

	Column Name	Data Type	Blank/Filled	Condition	Threshold
1	ingwyjr	Positive Integer	filled	ingwyjr+1900≤current year	N/A
2	ingwymnd	Positive Integer	filled	≤12	N/A
3	ingwydag	Positive Integer	filled	≤31	N/A
4	term	Positive Integer	filled	only fixed values: 1, 3, 6, 12, 60,120	N/A
5	standpni	Rational number	filled	standpni and standpi cannot be 0 at the same time	25000
6	standpi	Rational number	filled	standpni and standpi cannot be 0 at the same time	5000

For further analysis, above quality standards are transformed to Microsoft Excel functions. When the functions are applied to Brandmaster data set, each field generates a TRUE or FALSE value. See MS Excel function list at Appendix VIII - Table 15.

After applying predefined *quality standards* as Excel functions, an Excel table consisting of TRUE and FALSE values is generated. If data item does not match the respective quality standard, the function generates a FALSE value, otherwise it generates a TRUE value. The thresholds are not included within the Excel functions, therefore they don't affect TRUE/FALSE values. Instead, we apply thresholds to the dataset separately, to see the exact number of data that exceeds the thresholds.

Results

The Syntactic Accuracy rate of the data items is calculated using *Simple Ratio* below:

Number of TRUE Items / Total Item Number

Table 12 shows that the large majority of the records meet the predefined quality standards, therefore their *Syntactic Accuracy* rate is very close to 1, which indicates a high accuracy. Additionally, we can see the number of records exceeding the thresholds is very low considering the amount of data analyzed. The records above the threshold and FALSE values should be examined for a root-cause analysis.

TABLE 12. ASSESMENT RESULTS

ingwyjr	ingwymnd	ingwydag	term	standpni	standpi	
389986	390000	390000	389995	389997	389997	Total Number of TRUE values
14	0	0	5	3	3	Total Number of FALSE values
				25	3	Number of values exceed the threshold
0.999962	**0.999997**	**0.999997**	**0.999985**	**0.99999**	**0.99999**	**Syntactic Accuracy Rate (Total TRUE values/Total values)**

In INSC's environment, this example corresponds to the assessment of four levels of Data Quality Indicators (DQIs) described in Section 6.1.3. However, Level 4 has been included only partially (as thresholds) since the author did not have access to the historical data or other data sources that should be used for reconciliation.

6.2.2. Data Quality Monitoring

In DQM System Specifications section, the importance of monitoring data flows as well as data items is mentioned. Now we give an example of the data flow monitoring activity using Fire Insurance business unit's generic data flow. This data flow is valid for the majority of business units at INSC that are responsible for a specific type of insurance such as traffic insurance. In this example, we propose inserting several *data quality check points* to the data flow to increase the quality of the final data product.

VTA is the initial application which policy administrators use to record the customers' insurance policy related information. Different departments use different VTA applications and interfaces; some of these legacy applications are developed internally and still in use even though that specific type of insurance policy is not sold anymore. They are still in use, because they provide access to the existing policy holder's data. Some other types of VTA applications are customized "off the shelf" enterprise software such as Peoplesoft and SAP. Mainly VTA data is stored in a mainframe environment and internally developed SAS programs extract the data from the mainframe.

Figure 12 represents the data flow for Fire Insurance business unit beginning with policy entrance into the VTA application and finalizing with data import to the Business Data Warehouse. Below are the suggested process steps where data quality checks should be applied:

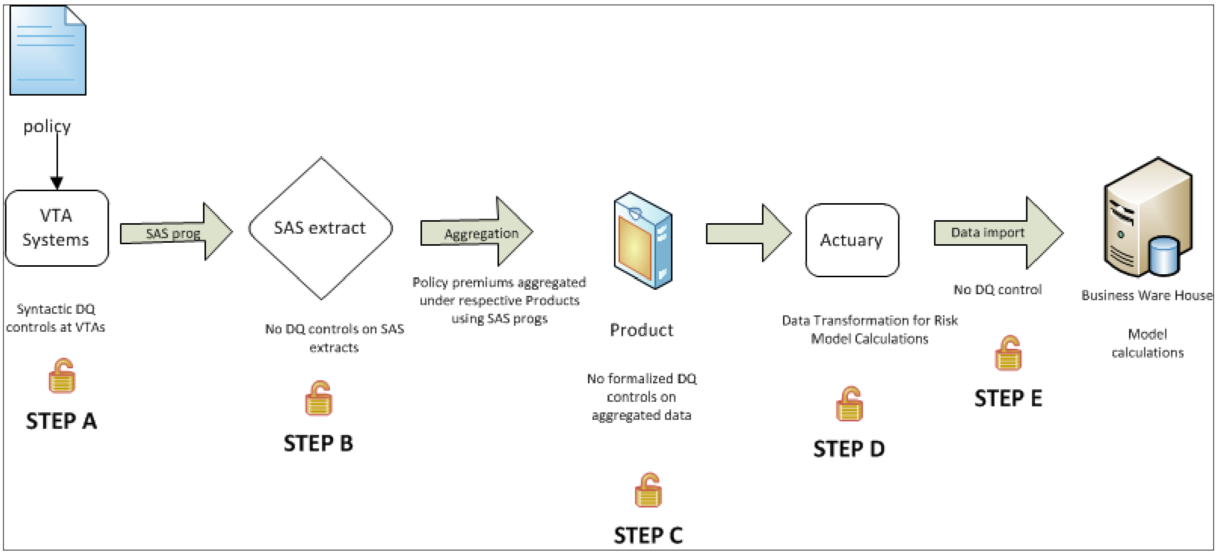
policy
VTA Systems
SAS prog
SAS extract
Aggregation
Policy premiums aggregated under respective Products using SAS progs
Product
Actuary
Data import
No DQ control
Business Ware House
Model calculations
Syntactic DQ controls at VTAs
STEP A
No DQ controls on SAS extracts
STEP B
No formalized DQ controls on aggregated data
STEP C
Data Transformation for Risk Model Calculations
STEP D
STEP E

FIGURE 12. FIRE INSURANCE BUSINESS UNIT DATA FLOW

- **Step A:** Usually VTA applications have their own quality controls at syntactic (structural) level to make sure correct type of data is entered in a specific field, such as an enforced data format in a date field or a drop down menu for gender selection. However, these controls do not prevent creating a record with missing information: If the information is missing, either the record is created with a blank field (if the system allows) or some standard value is entered in the field, such as using *male* for the unknown gender. Furthermore, semantic errors are also difficult to prevent as they are more difficult to capture compare to syntactic errors. For instance using 1915 as a birth year instead of 1951 for a policy holder.

- **Step B:** Most of the SAS scripts used to extract data are developed a while ago and their content and exact purpose are not documented. Currently within Solvency II context, some of these scripts are revisited to document their content and create an information repository. Technically, these scripts could also be used to check especially syntactic data quality of the extracted data which can provide an additional quality control.

- **Step C:** Using SAS programs, extracted files are aggregated under Products which provides a higher level data view. At this stage, no automated quality check is available. Again, structural quality checks could be embedded into the SAS programs which run the aggregations.

- **Step D:** The Product level view of data is used by Actuaries[24]. Actuarial function is an important part of Solvency II legislation and providing high quality data would increase the accuracy of estimations they have to calculate. However, more transparency is required also for the actuarial calculations which are usually not automated and done based on individual spreadsheets. At this step, with cooperation of the actuaries, well designed quality checks should be integrated into the data flow.

- **Step E:** Currently all data coming from the business units, that will involve Solvency II calculations, are imported into the data warehouse without any elimination or quality checks. After the Data Management team completes working with the business units and finalizes the data dictionaries, they will be able to provide feedback to the data warehouse administrators on data quality requirements and data scope. Then adequate quality checks should be embedded into the import process to avoid: (1) loading unnecessary data to the data warehouse, (2) loading poor quality data to the data warehouse.

Also, to implement optimized quality checks, the location of the checks should be well-defined and should correlate with the business value of the activities. That means, the location where we insert a quality check should have importance within the business process that uses the data.

6.3. Recommendations for the Insurance Company

As the final part of the field study, a list of recommendations for INSC is developed. The recommendations are derived from the observations of practices as well as from the analysis of the available policies.

1. Just like all changes in organizations, embedding a Data Quality Management system into the organizational processes requires a strong management support and ownership. Management support provides rapid spread of information within the company about the new goals. Then the new goals should be translated into actual tasks and targets for employees. Resistance to change is a natural instinct in organizations, but awareness and ownership are the best remedies. Unless the new process is owned by employees, who work in different organizational layers, it cannot be successful.
   INSC needs to pay special attention to providing information on Data Management activities and targets to the employees. For instance, another quality program called SAFE has been

[24] Actuaries analyze important data such as mortality, sickness, injury and disability rates and use that information to aid those involved with insurance. An actuary is responsible for collecting the data to forecast future risks and see how these predictions will affect various aspects of insurance [88].

running in order to comply with quality requirements of Dutch National Bank (DNB) for some time. However, employees are not sure that what they need to do differently on data quality for Solvency II compliance.

2. Producing quality data requires a team effort and contributions from all parties involved in data production processes. High level of user awareness of overall data quality should be achieved via creating clear data process flows. As a result, the individuals can realize the value of their activities and the data they produce within the entire data process flow. Rather than having the impression of working on isolated islands, users would be aware of their contribution within the whole system.

3. The high level data flow of the Solvency II regulation is shown below. Using VTA data, Cash Flow Model calculations are done, then Risk Models are applied to the calculation outcomes, and finally SRC is calculated:
   VTA → Cash Flow Model →Risk Models → SRC (Solvency Risk Capital)[25]

   On some occasions, Cash Flow system is not running properly due to the amount of blank fields in the corporate client's data. Consequently, the INSC is seeking methods to increase the amount of data that comes from the corporate clients with the pressure of Solvency II deadlines. The "quick and dirty" solution in practice is filling the blank fields with the standard values as those fields usually have no significant importance for the model calculations (such as the gender field). It seems that the solution is helping to increase the amount of processed data and resulting in more reliable model calculations. However, the solution is not improving overall data quality and the business unit might need the correct values of those fields for a new business activity or extension of the regulations in the future.
   The correct solution would be fixing the root cause of the issue: Why are those fields blank? Could personnel awareness be improved to make sure all fields are completed in the future? If those fields are not required, could we eliminate them from the interfaces? While implementing a temporary solution for the blank fields, companies should also make plans for revisiting those records to implement a permanent solution.
   Furthermore, quick-fixed data spreads through the interfaced systems and ends up in the data warehouse conflicting with *"data warehouse only contains trusted data* [78]" principal.

4. It has been indicated by the regulators that the Solvency II regulation will be extended in the future with several complementary policies. Thus, agility of Data Quality Management structure is paramount: Implemented structure should be flexible enough to extend or modify its data scope when it is needed. Automated processes are the essential basis of flexibility. Ross et al. argue in their well-known book the benefits of *foundation for execution*[26]: "*Digitizing core business processes makes the individual processes less flexible while making the company more agile*" [82]. Interestingly, once business processes are

[25] Capital required to finance the consequences of business risks.

[26] Foundation for Execution: Automating a company's core capabilities via IT infrastructure and digitized business processes.

digitized and automated, they provide better feedback on business activities and they do not require management attention as much as before. Consequently, management may spend more time on innovation.

Currently, many processes need to be automated at INSC's environment. In many cases, SAS extracts generated from the mainframe are transferred to different systems manually, such as MS Excel, and then modified. For instance, Actuarial Reporting activities (AV [27]) require using SAS extracts that include VTA data. The accountants prepare Technical Provisions (TP) using the extracts in MS Excel and add columns to the original data file, if they capture any inconsistent data. Automation of such processes using an adequate reporting system for AV would provide two benefits aligned with Solvency II: transparent processes and reliable data.

5. Well-established, clear procedures and processes throughout the organizational layers are required to prevent any conflicts among the parties involved. Especially before introducing the INSC's Data Management Policy, the Data Management team was pointed for all data related issues within the organization. This approach created tremendous amount of work load and pressure on the team. In fact, data should be owned and maintained by data producers and users: the business units. Data Management's role should be assisting them in achieving high quality data while remaining compliant with the regulations. Additionally, Data Management team creates the communication environment between different business units who work for the same goal: Solvency II compliance.

   Later on, introducing policies and roles helped to Data Management team in sharing responsibilities with different stake holders such as business units and IT department. However, these policies should be published, kept alive and modified as required.

6. Data Management activities require implementing processes and procedures as well as assigning roles (described as a part of DQ governance). However, especially monitoring and assessment activities require intensive IT involvement. Therefore, starting to work with the IT department in early stages of the project is recommended for the success of the project. At INSC, IT participation to Data Management team activities was established at a later stage in the project. Therefore, Data Management team members experienced difficulties in collecting information on data flows and creating data dictionaries, as they were not well aware of the capabilities of the IT infrastructure.

7. Consistent definitions and practices across the organization are needed to avoid any confusion. For instance, according to the documents generated by Data Architecture and Data Quality task forces, *Data Dictionary* and *Data Directory* terms seem to be separate aspects of the data definition. In INSC, *Data Dictionary* refers to an information repository that includes "technical information" on data such as data format, reference table, etc. within the database and *Data Directory* refers to an information repository that includes "functional information" on data such as use purpose, owner, etc. However, in the literature Data Dictionary is described as "*a centralized information repository on data such as meaning, relationships to other data, origin, usage, and format* [83] ", while Data Directory

[27] AV: Actuaris Verslaglegging

refers to just *a digital folder used to organize data*. Additionally, Data Directory as a term is not included in INSC's data management policy, although it has been used in practice [78].

8. INSC's Data Management Policy doesn't give separate definitions of the Solvency II criteria (accuracy, completeness and appropriates) and uses DFSG's Data Management Policy as the basis for the definitions [79]. Although DFSG's policy takes some part of the definitions from the regulatory documents, it seems like there is still confusion on clarifying their context.
   For instance, *appropriateness* dimension is introduced as being related to the granularity of data in DFSG's policy document. In CP 43, *appropriates* indicates how the portfolio is representative and suitable for the analysis. And representativeness is explained as "*being consistent with prospective view of relevant risks"* rather than granularity. Instead, granularity aspect takes place within *completeness* dimension: "D*ata is considered to be complete if it has sufficient granularity to allow for the identification of trends and the full understanding of the behavior of the underlying risks".*

9. In Chapter 5, analysis results of the methodologies showed that there is no perfect-fit methodology for a Solvency II project and all methodologies require some level of adjustments during implementations. However, selecting a methodology as a starting point during a DQM system implementation is essential. It clearly provides several benefits for organizations:

   a. Standardization of concepts and definitions across the organizational layers.
   b. Ability to compare with available methodologies or proprietary methodologies introduced by the consultancy companies.
   c. Ability to compare candidate DQ measurement software against the selected methodology.
   d. Ability to set clear organizational direction and targets on DQM strategies.

   At INSC, no methodology is adopted as a standard approach and Data Management team is navigating through the DQM knowledge introduced by the consultancy companies.

10. INSC is considering to purchase a Data Quality tool to be implemented as part of the data warehouse. During selection process, a candidate tool should be compared against: (1) Available DQ methodologies, (2) INSC's DQM structure, to identify possible gaps. The DQM methodology residing behind the tool may not be addressing the INSC's internal requirements. Also, the capability of the obtained methodology from the software company could be limited by the tool's capabilities [61].

Chapter 7. Conclusions

Solvency II is a new regulation for the European insurance companies and has clear emphasis on Data Quality (DQ). Therefore, clarifying its Data Quality Management (DQM) requirements and translating these requirements into system requirements is beneficial for companies. However, academic studies in the field are limited since the regulation has a short history.

This study is initiated to analyze the Solvency II directive's requirements on DQ. Furthermore, the study is also targeted to examine available literature on DQ from the Solvency II perspective in order to provide an overview of available DQ concepts and data quality management (DQM) methodologies. This information can be used as a reference for companies that need to implement a DQM structure to comply with the directive.

Main findings of the study are outlined in the following sections. Although the research reached its aims, there were some unavoidable limitations that are listed in the final section.

7.1. A DQM System for Solvency II

Throughout this study, we use requirements analysis techniques and systems design approach in order to find specifications of a DQM system that could be used to achieve Solvency II compliance. Table 5 includes analysis results and specifications of the intended system. Requirements analysis is useful in order to pinpoint the exact requirements within the complex regulatory documentation. Systems design approach helps to translate these requirements into an understandable and structured system view. Thus, insurance companies can easily adopt these outcomes or use them as a reference. This conclusion was verified at the insurance company (INSC), when Data Management team adopted some parts of the requirement analysis results of the study.

In Chapter 4, we outlined the activities to address each specification. These activities are explained in detail with realistic tasks considering an actual company's business environment and IT infrastructure.

7.2. Overview of DQ Concepts and Methodologies

In Chapter 3, an overview of DQ concepts and DQM methodologies is provided as a result of literature review phase of the study. The selected concepts and methodologies are represented in line with data quality related information obtained from the Solvency II documentations. Therefore, companies could compare the similar concepts between DQ literature and Solvency II to understand regulatory interpretation: such as meaning and scope of the Solvency II criteria (accuracy, completeness and appropriateness) in DQ literature. This information can also be used by companies in development of suitable measurement methods of the Solvency II criteria.

In addition, overview of existing methodologies provides information on what methodologies can be used in implementation of a DQM structure within a company. Since no methodology is mentioned in the regulatory documents, companies need to develop a methodology themselves or adopt an existing methodology to their environment.

7.3. Analysis of DQM Methodologies

As we mentioned earlier, according to the regulatory documents, companies need to develop their internal DQM solution. After providing an overview of existing methodologies in Chapter 3, in Chapter 5, we investigated usability of these methodologies by companies instead of developing a totally new DQM structure. Since this option could provide several benefits for companies, we analyzed the methodologies from the Solvency II perspective (Table 7).

The analysis concluded that none of the methodologies are a total misfit for Solvency II. However they all need some Solvency II specific extensions. The proposed extensions are shown in Table 8. The extensions and related activities are explained in detail in Chapter 5.

7.4. The Practical Case

Working in a practical case provides many benefits in development of DQM system specifications, rather than only working with the regulatory documents. As a result, the derived system specifications are practical enough to be applied by companies. In Chapter 6, two of the system specifications are operationalized in order to represent how the specifications should be applied to an actual company's environment.
At the end of Chapter 6, we introduce some recommendations for INSC's Data Management activities. However, these recommendations could easily be generalized for the industry. Probably the majority of the large insurance companies are experiencing similar difficulties in dealing with the Solvency II regulation. Some recommendations are also applicable outside the insurance sector, where data quality is critical to a company's operations.

7.5. Limitations of the Study and Future Work

The initial target of the study was to analyze the methodologies against each system specification derived earlier as a result of the requirements analysis activity (Chapter 4). However, limited information is available, especially on practices of the methodologies. Therefore, only a high level analysis is performed (Table 7). Possible reasons for lack of information as follows:

- Organizations that use these methodologies prefer to keep “the lessons learnt” for themselves due to confidentiality.
- Since no data is available from multiple firms using the same methodology, it is hard to generalize the methodology in order to be used by various firms in different business lines. Therefore the firms are reluctant in adopting these methodologies by themselves.
- Due to dissimilarities between the firms, their goals, industries and how they operate, a lot of customization is needed in implementation of any methodology; the firms need to generate their terminology, decide about the dimensions to measure, develop measurement methods, monitoring and reporting strategies etc. Most of the time, required tools should be internally developed based on organization’s specific needs.
- Large organizations usually work with consultancy companies in implementing a data quality management structure and usually consultancy companies prefer to introduce their proprietary methodologies which are not always publicly available.

Furthermore, the practical part (Chapter 6) remains limited to one insurance company and one business unit. Therefore, we propose the following, as an extension of this study: First, all proposed system specifications should be operationalized within multiple business units of a company to come up with company’s best practices. Second, the number of insurance companies should be increased to practice application of the system specifications. Findings of these practices result as the industry wide best practices in the field. Finally, these best practices are used to identify industry standards and to compare different companies’ environments.

In addition, one of the difficulties during the study was finding the way within the comprehensive regulatory documentation which has its own terminology and includes many citations to other regulatory documents. Considering many industries are becoming increasingly regulated, the number of the regulations a company needs to comply are quite high. Hamdaqa et al.’s article on *Citation Analysis to Facilitate the Understanding of Regulatory Documents*, points out the same problem [84]. Therefore, structured analysis of regulatory documents could appear as a new research field in the future.

Bibliography

[1] D. Thienpont, "EUROPA - Internal Market - Single Market News - Edition 46," 2007. [Online]. Available: http://ec.europa.eu/internal_market/smn/smn46/docs/insurance_en.pdf.

[2] CEIOPS, "CP43 - CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: Technical Provisions - Article 86 f, Standards for Data Quality," 2009.

[3] C. Batini and M. Scannapieco, Data Quality - Concepts, Methodologies and Techniques, Springer, 2006.

[4] L. P. English, Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits, Wiley, 1999.

[5] R. Y. Wang, "A Product Perspective on Total Data Quality Management," *Communications on the ACM,* 1998.

[6] CEIOPS, "CP56 - CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: Articles 120 to 126, Tests and Standards for Internal Model Approval," 2009.

[7] CEIOPS, "CP75 - CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: SCR standard formula," 2009.

[8] M. Neri, "Meeting the Data Quality Management Challenges of Solvency II," 2011. [Online]. Available: www.moodysanalytics.com.

[9] A. Dutta, S. Linsley and M. Edenroth, "Effective Data Quality Management: The Path to Solvency II," 2011. [Online]. Available: www.infogix.com.

[10] C. Sounders and G. Olsen, "Solvency II and Data: Myths and Misunderstandings," 2011. [Online]. Available: www.ibm.com.

[11] P. Ghauri and K. Gronhaug, Research Methods in Business Studies, 4th Edition, FT Prentice Hall, 2010.

[12] "Solvency II - Introducing Solvency II," [Online]. Available: https://eiopa.europa.eu/activities/insurance/solvency-ii/index.html.

[13] "Solvency II: FAQs," European Union, 2007. [Online]. Available: http://europa.eu/rapid/press-release_MEMO-07-286_en.htm?locale=en.

[14] E. P. Release, "Solvency II: EU to take global lead in insurance regulation," 2007. [Online]. Available: ec.europa.eu.

[15] "On the Proposal for Guidelines on ORSA," EIOPA, 2012.

[16] J. Smith, B. Meaney, G. Olsen, R. Jones and P. Havelock, "Solvency II: Enabling Transformation Trough Regulation," 2009. [Online]. Available: https://www-304.ibm.com/easyaccess/fileserve?contentid=178908.

[17] "Solvency II - News," The Financial Services Authority (FSA), [Online]. Available: http://www.fsa.gov.uk/solvency2.

[18] C. Parra-Serrano, "Solvency II: The Waiting Game," 2012. [Online]. Available: http://www.postonline.co.uk/post/feature/2219416/solvency-ii-the-waiting-game.

[19] F. S. Authority, "Solvency II: Internal Model Approval Process Data Review Findings," 2012.

[20] F. S. Authority, "External Review Scoping Tool," 2011.

[21] "Solvency II Requirements," Deloitte , [Online]. Available: http://www.deloitte.com/view/en_GB/uk/industries/financial-services/issues-trends/solvencyii/solvencyiirequirements/index.htm.

[22] "Taking-up and pursuit of the business of insurance and reinsurance (Solvency II)," European Union, 2008. [Online]. Available: http://europa.eu/legislation_summaries/internal_market/single_market_services/financial_services_insurance/l22030_en.htm.

[23] CEIOPS, "CP37 - Advice for Level 2 Implementing Measures on Solvency II on: The procedure to be followed for the approval of an internal model," 2009.

[24] "Building a better Solvency II Solution with IBM Insurance Information Warehouse," 2011. [Online]. Available: ftp://ftp.software.ibm.com/software//data/sw-library/industry-models/IIW_SolvencyII_Whitepaper.pdf.

[25] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems,* 1996.

[26] Y. Wand and R. Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM,* 1996.

[27] T. Redman, Data Quality for the Information Age, 1997.

[28] A. Mitra, "Classifying Data for Successfull Modelling," 2012. [Online]. Available: www.dwbiconcepts.com.

[29] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. van Liere, K. L. Ma, W. Ribarsky, G. Scheuermann and D. Silver, "Data, Information, and Knowledge in Visualization," *Computer Graphics and*

Applications, IEEE, 2009.

[30] R. D. Reid and N. R. Sanders, Operations Management, 4th Edition, Wiley, 2009.

[31] "Business Dictionary," WebFinance, Inc., [Online]. Available: http://www.businessdictionary.com/definition/quality.html.

[32] J. M. Juran and A. B. Godfrey, Juran's Quality Handbook, 5th Edition, New York: McGraw-Hill, 1999.

[33] "Discussion on Data Quality," China National Institute of Standardization, 2009. [Online]. Available: http://metadata-standars.org/Document-library/Documents-by-number/WG2-N1301-N1350/WG2_N1346_discussion_on_data_quality.pdf.

[34] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM,* 1970.

[35] K. Ivanov, "Quality Control of Information: On the concept of accuracy of information in data-banks and in management information systems. Doctoral dissertation.," The University of Stockholm and The Royal Institute of Technology, 1972. [Online]. Available: http://www8.informatik.umu.se/~kivanov/diss-avh.html.

[36] P. D. Ballou and H. L. Pazer, "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science,* 1985.

[37] M. D. Hansen, "Zero Defect Data, Tackling the Corporate Data Quality Problem, Master Thesis," Massachusetts Institute of Technology, 1991. [Online]. Available: http://dspace.mit.edu/handle/1721.1/13812.

[38] D. M. Strong, Y. W. Lee and W. R. Y, "Data Quality in Context," *Communications of the ACM ,* 1997.

[39] R. Y. Wang, H. B. Kon and S. E. Madnick, "Data Quality Requirements Analysis and Modelling," in *Ninth International Conference on Data Engineering*, Vienna, 1993.

[40] C. Batini, C. Francalanci, C. Cappiello and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys,* 2009.

[41] P. R. Benson, 2008. [Online]. Available: http://www.oilit.com/papers/Benson.pdf.

[42] R. Y. Wang, V. C. Storey and C. P. Firth, "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering,* 1995.

[43] M. R. Berthold, C. Borgelt and F. Höppner, Foundations of Intelligent Data Analysis: Making Practical Sense of Real Data, Springer, 2010.

[44] G. Shankar and S. Watts, "A Relevant, Believable Approach for Data Quality Assessment," in *8th International Conference on Information Quality*, 2003.

[45] L. L. Pipino, Y. W. Lee and R. Y. Wang, "Data Quality Assessment," *Communications of the ACM,* 2002.

[46] B. Heinrich, M. Kaiser and M. Klier, "How to measure data quality? – a metric based approach," in *28th International Conference Information Systems (ICIS)*, Montreal, 2007.

[47] A. Even and G. Shankaranarayanan, "Value-Driven Data Quality Assessment," in *10th International Conference on Information Quality*, 2005.

[48] B. Otto, K. Wende, A. Schmidt and P. Osl, "Towards a Framework for Corporate Data Quality," in *18th Australasian Conference on Information Systems*, 2007.

[49] R. D. Snee, "Why Should Statisticians Pay Attention to Six Sigma?," *Quality Progress,* 1999.

[50] D. W. Benbow and T. M. Kubiak, "The Certified Six Sigma Black Belt Handbook," American Society for Quality (ASQ), [Online]. Available: http://asq.org/learn-about-quality/six-sigma/overview/overview.html.

[51] J. Antony, "Some Pros and Cons of Six Sigma: An Academic Perspective," *TQM Magazine,* 2004.

[52] Y. W. Lee, D. M. Strong, B. K. Kahn and R. Y. Wang, "AIMQ: A Methodology for Information Quality Assessment," *Elsevier Information and Management,* 2002.

[53] C. Batini, D. Barone, M. Mastrella, A. Maurino and C. Ruffini, "A framework and a methodology for data quality assessment and monitoring," in *12th International Conference on Information Quality*, 2007.

[54] F. Wijnhoven, R. Boelens, R. Middel and K. Louissen, "Total Data Quality Management: A Study Of Bridging Rigor And Relevance," in *15th European Conference on Information Systems*, St. Gallen, 2007.

[55] R. Kovac, Y. W. Lee and P. L. L, "Total data quality management: the case of IRI," in *Conference on Information Quality*, 1997.

[56] P. Nadkarni, "Delivering Data on Time: The Assurant Health Case," in *11th International Conference on Information Quality*, 2006.

[57] G. Shankaranarayanan, R. Y. Wang and M. Ziad, "Ip-Map: Representing the manufacture of an information product," in *The 2000 Conference on Information Quality*, 2000.

[58] M. Scannapieco, B. Pernici and E. Pierce, "IP-UML," *Information Quality,* 2005.

[59] Y. W. Lee, D. M. Strong, B. K. Kahn and R. Y. Wang, "AIMQ: A Methodology for Information Quality Assessment," *Elsevier, Information & Management,* 2002.

[60] B. K. Kahn, D. M. Strong and R. Y. Wang, "Information Quality Benchmarks: Product and Service Performance," *Communications of the ACM,* 2002.

[61] L. P. English, "Total information quality management: A complete methodology for IQ management," *DM Review,* 2003.

[62] "Systems Engineering / System Design and Development," 2009. [Online]. Available: http://www.mitre.org/work/systems_engineering/guide/se_lifecycle_building_blocks/system_design_development/.

[63] G. Kotonya and I. Sommerville, "Requirements Engineering with Viewpoints," *Software Engineering Journal,* 1996.

[64] D. T. Ross, "Structured analysis (SA): A language for communicating ideas," *IEEE Software Engineering,* 1977.

[65] J. Mylopoulos, "Information Systems Analysis and Design / VIII. Requirements Analysis (powerpoint slides)," 2004. [Online]. Available: http://www.cs.toronto.edu/~jm/340S/Slides2/ReqA.pdf.

[66] R. Malan and D. Bredemeyer, "Functional Requirements and Use Cases," 1999. [Online]. Available: http://www.bredemeyer.com/pdf_files/functreq.pdf.

[67] L. Chung and J. do Prado Leite, "On Non-Functional Requirements in Software Engineering," *Springer, Conceptual modeling: Foundations and applications,* 2009.

[68] M. Glinz, "On Non-Functional Requirements," in *15th IEEE International Requirements Engineering Conferenc*, 2007.

[69] A. Even and G. Shankaranarayanan, "Value-Driven Data Quality Assessment," in *10th International Conference on Information Quality*, 2005.

[70] "ISO 9001:2000 - Quality management systems - Requirements," International Organization for Standardization, 2000.

[71] "The Free Dictionary," [Online]. Available: http://www.thefreedictionary.com/proportionality.

[72] H. Li and J. F. Reynolds, "On Definition and Quantification of Heterogeneity," *Oikos,* 1995.

[73] B. C, B. D, C. F and G. S, "A Data Quality Methodology for Heterogeneous Data," *International Journal of Database Management Systems,* 2011.

[74] D. Loshin, "Evaluating the Business Impacts of Poor Data Quality," [Online]. Available: http://www.sei.cmu.edu/measurement/research/upload/Loshin.pdf.

[75] M. Chisholm, "There is No Single Version of the Truth," 2006. [Online]. Available: http://www.information-management.com/issues/20061201/1069851-1.html?zkPrintable=1&nopagination=1.

[76] "Taskforce Data Architecture v3 - Corporate Insurance Solvency II program".

[77] "Solvency II Program - DQ Rollout Plan 2012/2013".

[78] "INSC Data Management Policy and Process Guide, Draft v0.18".

[79] "DFSG's Data Governance and Quality Management Policy v1.3".

[80] P. Ghauri and K. Gronhaug, Research Methodologies in Business Studies, 4th Edition, Prentice Hall, 2010.

[81] I. Diamond and J. Jefferies, Beginning Statistics, An Introduction for Social Scientists, Sage Publications, 2009.

[82] J. W. Ross, P. Weill and D. Robertson, Enterprise Architecture As Strategy: Creating a Foundation for Business Execution, Harvard Business Press, 2006.

[83] IBM, IBM Dictionary of Computing, 10th Edition, McGraw-Hill, 1993.

[84] M. Hamdaqa and A. Hamou-Lhadj, Citation analysis: An Approach for Facilitating the Understanding and the Analysis of Regulatory Compliance Documents, Sixth International Conference of Information Technology, 2009.

[85] M. Scannapieco, P. Missier and C. Batini, "Data Quality at a Glance," *Datenbank-Spektrum,* 2005.

[86] P. Atzeni and V. De Antonellis, Relational Database Theory, Benjamin-Cummings Publishing Co., Inc., 1993.

[87] "GIM6000 - Technical provisions," UK Goverment - HM Revenue & Customs, [Online]. Available: http://www.hmrc.gov.uk/manuals/gimanual/gim6000.htm.

[88] "The Role and Responsibilities of an Actuary," 2006. [Online]. Available: http://www.exforsys.com/career-center/career-tracks/the-role-and-responsibilities-of-an-actuary.html.

Appendices

Appendix I. Interviews

Below is the role list of the interviewees at the Insurance Company. Some of the interview questions are listed below (Chapter 6).

Interviewee role list:

Functional Designer (Responsible of several software installed on the mainframe environment)
Information Analyst (Currently writing functional designs of SAS extracts)
Data Warehouse Architect
Data Modeling team member
Business Architect (Working on conceptual and logical data model on IIW and technical model on SAP BW)
Several Data Management team members
Corporate Clients business unit employee
Senior Accounted responsible of Corporate Clients
A specialist on Business Value Chain

Sample Questions:

What is your background?
What is your role within the Solvency II project?
Which activities you need to perform to full-fill this role?
Which difficulties you face while performing your tasks?
What is the relationship between your tasks and data quality?
Did you consider data quality before the Solvency II project during your business activities?
What is your opinion on data quality level at INSC?
What do you think on data quality activities at INSC? Are they sufficient, what could be improved?

Appendix II. Mind Map

Mind Map is used during the initial phase of the study to structure the research process (Chapter 1).

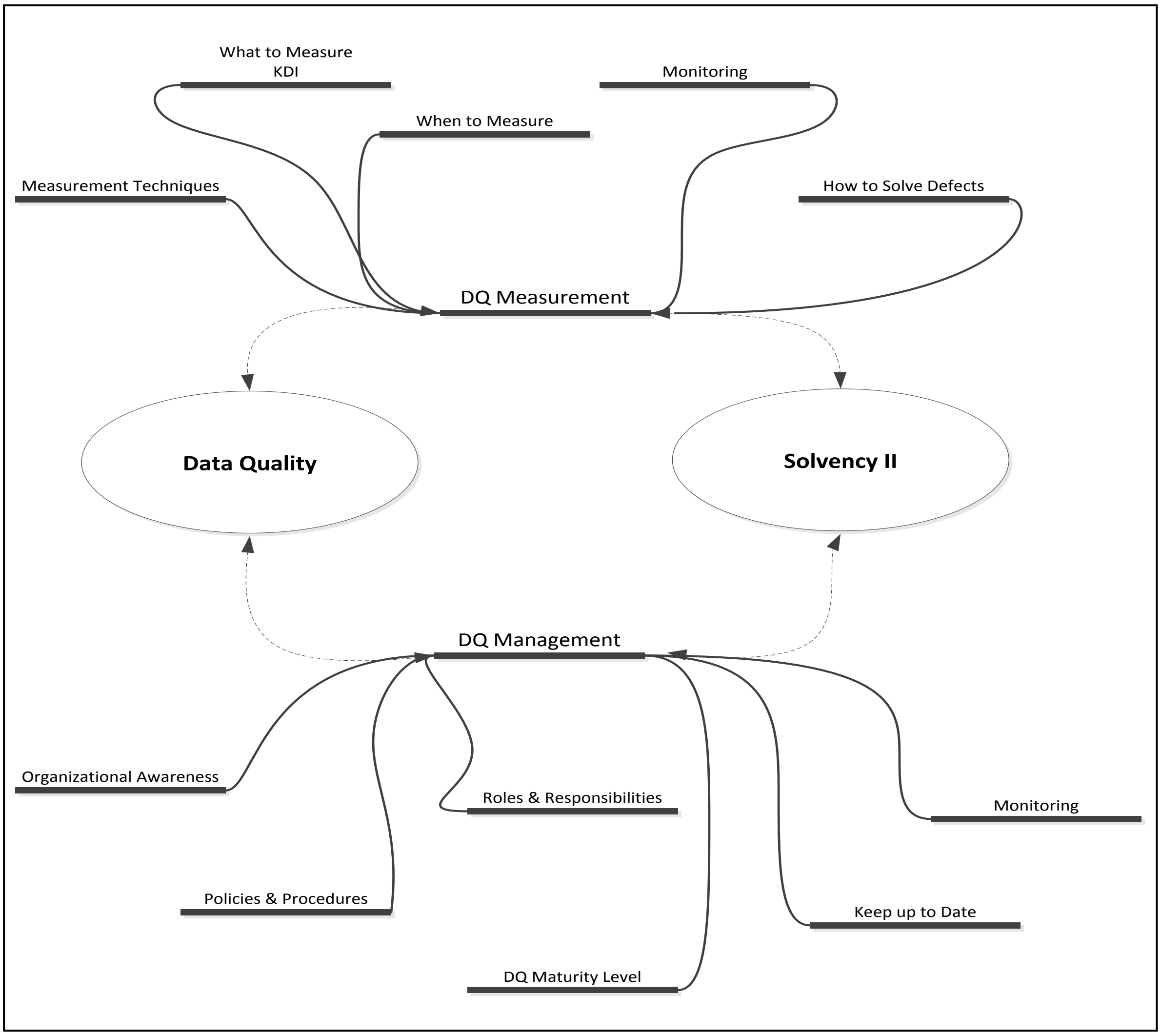

FIGURE 13. MINDMAP

Appendix III. Sample Data Dictionary

Table 13, represents a part of the Data Dictionary used at the Insurance Company where the field study took place (Chapter 6).

TABLE 13. DATA DICTIONARY EXAMPLE (OBTAINED FROM THE INSURANCE COMPANY)

Column Name	Column Definition
Solvency II Domain	The SII Domain provides information where the data item is located within the SII chain. Possible domains are: Insurance/Mortality, Insurance/Morbidity, Insurance/P&C, Business/Political risk, ..etc.
Process ID	The process ID refers to the unique process in which the data item is located.
Data flow ID	The data flow ID is the reference to the data flow diagram and the exact location where the data item is used as input data in a data set, see also E below.
Source System	The source system provides information of the origin of the data. The source might be a DB, a spreadsheet, an application, etc.
Data set ID	A data set ID refers to the data set which includes the data item.
Data set description	The description should allow the user to understand the data set in the process flow and data flow.
Data item	Name of the data item.
Data item definitions	The basic definitions of data used for the overall SII reporting i.e. a unique definition attributed to data item to describe the content of such item.
Key Data Item	Indicates with Y/N whether the information is a Key Data Item. A Key Data Item is considered Key if: o it is an input data o it is part of a material Subprocess (sub processes are: Business persistency uncertainty risk , market FX risk, business persistency volatility risk...). Materiality is defined if the Subprocess contributes more than 1% of the SCR contribution o The attribute impacts SCR calculations: only the attributes where the information impacts the SCR calculation are considered key.
Data type	Only for KDIs. The data type provides the type of the key data item type.
Data format	Only for KDIs. Defines the allowed format of the data item. This information is used to set DQI 3.
Control ID	Only for KDIs. If a control exists, either an automated control, documented manual control, etc. the control ID refers to the existing control ID. The sheet "Existing control details" include detailed information about the existing controls.
Control Description	Only for KDIs. The control description should allow the user to understand what the control is testing. Further details such as recurrence, person who perform the test, documentation etc. should be indicated in the sheet "Existing controls details".
Data Owner	Only for KDIs. The field data owner indicates the position of the data owner, rather than a specific name of an employee. However, it must be possible to allocate the position to one specific person. If this is not possible further details are necessary to allow the identification of the data owner.

Appendix IV. Use Case

A use case represents a way of using the system and respectively the required behavior of the system in that particular scenario. Thus, use cases capture who (actor) does what (interaction) with the system, for what purpose (goal), without dealing with system internals [66]. Uses cases are wide-spread practices of explaining functional requirements graphically. Therefore, one of the functional requirements, Data Collection, is expressed as a Use Case scenario in Figure 14 using UML representation (Chapter 4).

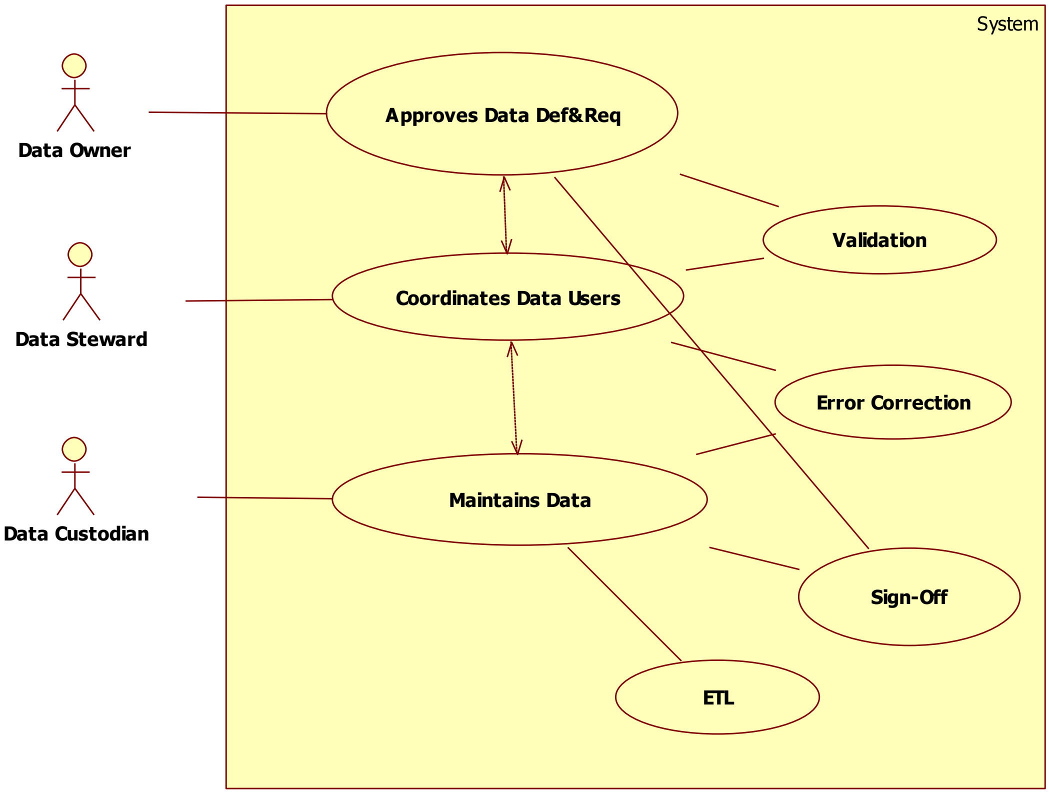

FIGURE 14. UML - USE CASE SCENARIO OF DATA COLLECTION

Appendix V. Cost Matrix

An example cost matrix that could be used for Solvency II is shown below (Chapter 4).

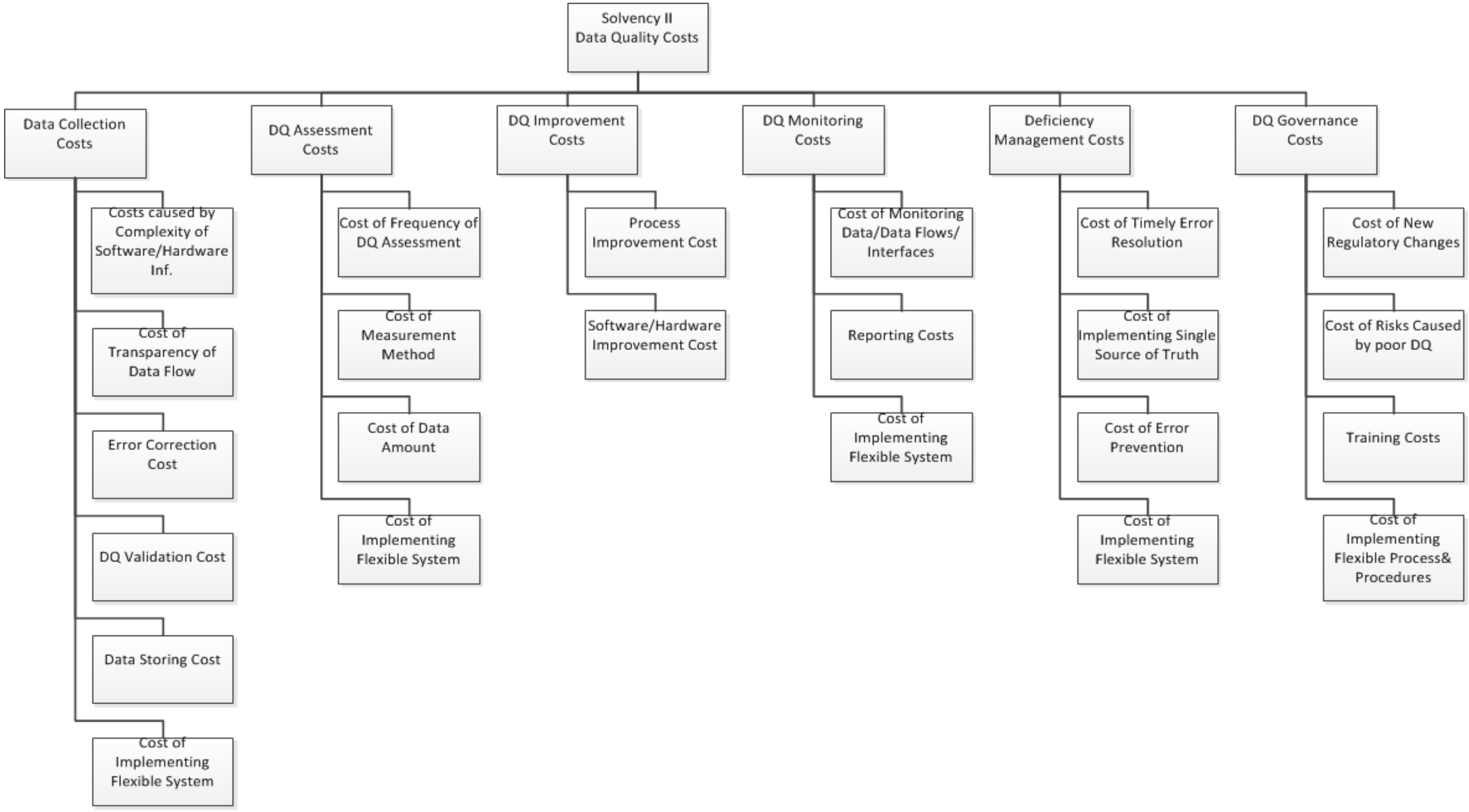
Solvency II Data Quality Costs
Data Collection Costs
Costs caused by Complexity of Software/Hardware Inf.
Cost of Transparency of Data Flow
Error Correction Cost
DQ Validation Cost
Data Storing Cost
Cost of Implementing Flexible System
DQ Assessment Costs
Cost of Frequency of DQ Assessment
Cost of Measurement Method
Cost of Data Amount
Cost of Implementing Flexible System
DQ Improvement Costs
Process Improvement Cost
Software/Hardware Improvement Cost
DQ Monitoring Costs
Cost of Monitoring Data/Data Flows/ Interfaces
Reporting Costs
Cost of Implementing Flexible System
Deficiency Management Costs
Cost of Timely Error Resolution
Cost of Implementing Single Source of Truth
Cost of Error Prevention
Cost of Implementing Flexible System
DQ Governance Costs
Cost of New Regulatory Changes
Cost of Risks Caused by poor DQ
Training Costs
Cost of Implementing Flexible Process& Procedures

FIGURE 15. SOLVENCY II COST MATRIX

Appendix VI. Impact Classifications

Table 14, includes example impact classifications for *Data Value Measurement* activity explained in Chapter 4.

TABLE 14. SOLVENCY II – IMPACT CLASSIFICATIONS FOR DATA VALUE MEASUREMENT (INSPIRED FROM [70])

Impact Class	Sample Impacts of Poor Data Quality in a Solvency II System
Financial	• Regulatory fines, penalties. • Miscalculation of operational risk which would result to higher capital requirements. • Missing the incentives offered by the regulators for better measuring and managing.
Operational Efficiency	• Inability of streamlining data generation process. • Inconsistency between data flows and business flows which would result complexity in DQ checks.
Confidence	• Distorted organizational image.
Satisfaction	• Dissatisfaction of the internal and external stake holders (customers, employees, shareholders and regulatory/supervisory authorities)
Risk	• Increasing the organizational Risk Exposure.
Regulatory Compliance	• Lack of transparency in data flows. • Inability of justification of used quality measurement methods. • Inability of justification of using expert judgment. • Incompliance with the Principal of Proportionality.

Appendix VII. R Commands

Below is the list of R commands used in Chapter 6 for statistical analysis.

```
> Brandmaster<-read.table("C:\\Brandmaster.csv", sep = ",", dec = ".", header = TRUE)

> nrow(Brandmaster)
[1] 390000

> edit(Brandmaster)

> summary(Brandmaster$standpi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    2.64   22.28   37.83   48.11 5799.00

> summary(Brandmaster$standpni)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  -379.80      0.00      0.00     68.59      0.00 245800.00

> s=Brandmaster$standpi
> sd(s)
[1] 81.34022

> t=Brandmaster$standpni
> sd(t)
[1] 683.1179

> plot (density(Brandmaster$standpi))

> plot (density(Brandmaster$standpni))

> boxplot(Brandmaster$standpi, Brandmaster$standpni)
```

FIGURE 16. R COMMANDS

Appendix VIII. Excel Functions for the Practical Case

MS Excel functions used in Chapter 6 for Syntactic Accuracy measurement are shown below.

TABLE 15. EXCEL FUNCTIONS

	Column Name	
1	ingwyjr	=AND(B2>0,B2+1900<= 2012,IF(ISBLANK(B2),FALSE,TRUE),IF(ISNUMBER(B2), TRUE, FALSE),IF(B2 = INT(B2), TRUE, FALSE))
2	ingwymnd	=AND(C2>0,C2<= 12,IF(ISBLANK(C2),FALSE,TRUE),IF(ISNUMBER(C2), TRUE, FALSE),IF(B2 = INT(C2), TRUE, FALSE))
3	ingwydag	=AND(D2>0,D2<= 31,IF(ISBLANK(D2),FALSE,TRUE),IF(ISNUMBER(D2), TRUE, FALSE),IF(D2 = INT(D2), TRUE, FALSE))
4	term	=AND(E2>0,IF(ISBLANK(E2),FALSE,TRUE), OR(E2={1,3,6,12,60,120}))
5	standpni	=AND(IF(ISBLANK(F2),FALSE,TRUE),OR(IF(AND(F2=0,G2<>0),TRUE,FALSE),IF(AND(F2<>0,G2=0) ,TRUE,FALSE),IF(AND(F2<>0,G2<>0),TRUE,FALSE)))
6	standpi	=AND(IF(ISBLANK(G2),FALSE,TRUE),OR(IF(AND(F2=0,G2<>0),TRUE,FALSE),IF(AND(F2<>0,G2=0) ,TRUE,FALSE),IF(AND(F2<>0,G2<>0),TRUE,FALSE)))

Appendix IX. Dimension Measurement Techniques

Some of the measurement techniques that could be used to measure the dimensions proposed in Chapter 4 are shown below.

TABLE 16. DIMENSION MEASUREMENT TECHNIQUES

		Dimension	Measurement Technique
Accuracy	1	Syntactic Accuracy	Use comparison functions to measure the distance between the value v and the true value d, such as *Edit Distance* [85] Alternative method is available at [46] as correctness measurement
	2	Currency	= Age +(DeliveryTime – InputTime) Age: How old the data unit DeliveryTime: The time the information product is delivered to customer InputTime: The time the data unit is obtained [3]
	3	Traceability	Data items should include an identification tag that includes its location, history and usage
	4	Credibility	Individual's assessment (such as actuary) of the credibility of the data source, comparison to a commonly accepted standard, and previous experience [45]. Alternative method is available at [44]
	5	Consistency	Data Edits[28] for non relational data [3]. Integrity Constrains for relational data [86].
	6	Volatility	No need of introducing specific metrics for it as it inherently characterizes types of data [85] High Volatile -> Data must be current Low Volatile -> Currency less important
	7	Timeliness	Ranges from 0 to 1, max {0,1 - (currency/volatility)} [3] Alternative method is available at [46]
Completeness	8	Completeness	Simple ratio: incomplete values/ all values [45] More information is available at [3]
	9	Granularity/ Depth of Data	To find out whether data has sufficient granularity, it should be compared to required level of detail
	10	Historical Data	Simple Ratio: Amount of available historical data available to amount of expected historical data
	11	Proportionality	Simple Ratio: The ratio of the number of data items available for calculation to the number of data expected to be available (data expected to be available increases if the risk level is high)
	12	Variety of data / Heterogeneity	The number of data sources used to generate the data, that information could be included in identification tag which will be used for traceability
Appropriateness	13	Relevancy	Dual – process approach introduced at [44]
	14	Semantic Accuracy	Use yes/no question – the corresponding true value has to be known [3].
	15	Amount of Data	Simple Ratio: The ratio of the number of data units provided to the number of data units needed, and the ratio of the number of data units needed to the number of data units provided [45]

[28] Data Editing: Task of detecting inconsistencies by formulating rules that must be respected by every correct set of answers.

List of Figures

Figure 1. Research development 8
Figure 2. Document Organization 9
Figure 3. Solvency II Pillars [21] 11
Figure 4. Solvency II Legislation Development and Regulatory Documents on Data Requirements.... 28
Figure 5. Requirements Analysis Steps........ 28
Figure 6. Design of the Intended DQM system - System View........ 29
Figure 7. Solvency II – Data Governance Model........ 41
Figure 8. Density Plot of Indexed Insurance premium Figure 9. Density Plot of non-Indexed Insurance premium 55
Figure 10. Boxplot for indexed premium Figure 11. Boxplot for non-indexed premium........ 55
Figure 12. Fire Insurance business unit data flow 58
Figure 13. MindMap 73
Figure 14. UML - Use Case Scenario of Data Collection 75
Figure 15. Solvency II Cost Matrix 76
Figure 16. R commands 78

List of Tables

Table 1. Data Quality Methodologies [38] - Extended with ORME-DQ 21
Table 2. Product vs. Information Manufacturing [50].......... 23
Table 3. Solvency II Functional Requirements on Data Quality Management.......... 30
Table 4. Solvency II Non-Functional Requirements on Data Quality Management.......... 31
Table 5. Solvency II Data Quality Management System Specifications.......... 33
Table 6. Definitions of the Proposed Dimensions 37
Table 7. Analysis of DQM Methodologies 44
Table 8. Proposed Extensions to Methodologies.......... 46
Table 9. Solvency II - Key Data Items for Fire Insurance Data.......... 54
Table 10. Original Data File 56
Table 11. Quality Standards 56
Table 12. Assesment Results 57
Table 13. Data Dictionary Example (Obtained from the insurance company).......... 74
Table 14. Solvency II – Impact Classifications for Data Value Measurement (INSPIRED FROM [70])... 77
Table 16. Excel Functions 79
Table 17. Dimension Measurement Techniques.......... 80