CRANFIELD UNIVERSITY

MARIA J. CARRION

DESIGN AND IMPLEMENTATION OF AN AUTOMATED DATA QUALITY CONTROL SYSTEM IN THE FINANCIAL SECTOR

SCHOOL OF APPLIED SCIENCE
Knowledge Management for Innovation

MSc
Academic Year: 2012 -2013

Supervisor: Dr. Jörn Mehnen
September 2013

CRANFIELD UNIVERSITY

SCHOOL OF APPLIED SCIENCE
Knowledge Management for Innovation

MSc

Academic Year 2012 -2013

MARIA J. CARRION

Design and Implementation of an Automated Data Quality Control System in the Financial Sector

Supervisor: Dr. Jörn Mehnen
September 2013

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science

ABSTRACT

Lending Controls play an important role in ensuring that banking product control standards are understood and applied. The Data Quality and Structural Systems team continuously develop and deliver operational controls within the Core Banking Platform in order to improve the performance of the business and to minimise risk and costs. The identification of existing internal controls and reporting of Data Quality in a bank and their link with automation requirements and software process improvement are in the focus of this research. Therefore, the aim is to develop an automated process and software tool that helps to improve and control internal banking processes as well as manage the data quality assessment and monitor the retrieval of the pending position status of loan processes in banks. The methodology followed is based on Software Re-engineering such as Agile Methodology and Extreme Programming. The automation of Data Quality Internal Controls has proven to be a suitable process to improve the quality of data, minimising the errors and reducing the time as well as the management of the customer impact.

Keywords:

Internal Controls, Lending Controls, Agile methodology, Software Improvement, Software automation

ACKNOWLEDGEMENTS

I would like to thank my academic supervisor Dr. Jörn Mehnen, for his invaluable guidance and constructive suggestions to accomplish the academic requirements of the thesis.

I would also like to offer my special thanks to Claire Baaye and Gary Rodwell at Santander bank Plc. As my industrial supervisor, Claire has been extremely helpful and always available for advice and Gary (El Guru del VBA) has been tremendously generous with his expertise and endless knowledge during the technical element of this project.

Finally, I wish to thank my colleagues, Amina, Ana, Mike and Paula for their constant support and patience throughout my study.

The target journal is “Journal of Data and Information Quality (JDIQ)”

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IT	Information Technology
IC	Internal Controls
SOX	Sabanes-Oxley
SCM	Supply Chain Management
BI	Business Intelligence
SPI	Software Process Improvement
KPI	Key Performance Indicators
DQSS	Data Quality and Structural Systems
VBA	Visual Basic for Applications

1 INTRODUCTION

The financial sector is a complex environment with a lot of uncertainties. There are complicated operations with a high level of risk and complex concepts to understand financial activities. Concretely, banking operations involve complex mathematical calculations. Usually there is a core Banking Platform that automatically calculates all of these operations. However, the task of checking any possible errors that a system can commit is typically not widely automated. Behind each financial operation there is an algorithm that needs to be understood. In order to analyse the concept and process of all operations, the specifications of the product and the conditions must be well understood.

In this particular case the products to be analysed are lending products such as mortgages and loans. The definition of mortgage according to [23] is: "*A secured loan against a property".* According to [20] the definition of loan is: "*The act of giving money, property or other material goods to another party in exchange for future repayment of the principal amount along with interest or other finance charges*". Each lending product has different components such as the fees, capital or interest. Therefore every component has a different behaviour depending on the specification of the mortgage or loan. In banks there is a complex system that calculates the processes according to the conditions defined by the contract of the product: each loan repayment to pay, each date to settle or the amount of fees to pay. Therefore, the process to control the data is one of the main issues that this study is focusing on. The DQSS (Data Quality and Structural Systems) team in bank used in the case study if this research is responsible for the key Parameters of the Structural Systems in Santander's Core Banking Platform and Data Quality and Control of the Accounts and Customers for the whole bank considered in the case study.

The bank that participated in this research is an international retail and commercial bank with locations such as Europe, and America. It is the second largest bank in Europe and with more than 1500 branches and 25 million customers. The business tool orientation is retail, which is a type of protection to

the changes internationally considered from the construction of different structures between commercial and investment banking [5].

The big managers of banks consider pockets of risks that may affect the development of a business and therefore what risks are involved. This means that the quality of risk management is a priority in the performance of a bank. One of the most crucial concerns is that of the credit risk and whether a company would be able to compete in the high end market or end with bankruptcy [6].

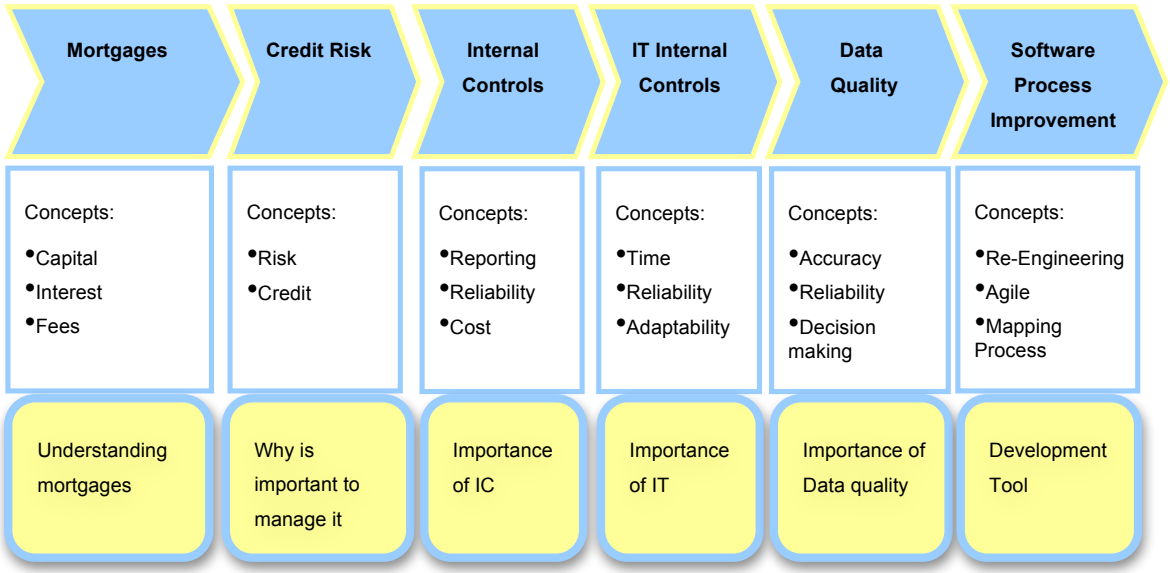

Figure 1 - Literature Review Map

In a banking environment there are several variables that enable the success or failure of an activity. An example of this can be considered to be risk. Risk is defined by the Oxford's English dictionary as: "*a situation involving exposure to danger*" [30]. However financial risk is: "*The probability that an actual return on an investment will be lower than the expected return*." [11]. Hence managing risk becomes a highly important issue to be included into a company's agenda because an effective control of risk must be reached. Bad management of risk can cause severe problems to the cash flow and solvency of a bank. In fact, bank insolvency has been an important problem all around the world in the last decades [10]. [2] defines risk management in business as a "*core activity in asset allocation conducted by banks, insurance and investment companies, or*

any financial institution that evaluates risks". Specially, credit risk is currently one of the key factors to efficiently manage and accomplish the requisites of high competence in banking. It has been usually defined as default risk, i.e. "*the risk of loss from a borrower failure to repay the amount owed (principal or interest) to the bank on a timely manner based on a previously agreed payment schedule*" [33]. Therefore, the study of credit risk becomes important and its analysis is an important problem challenged by banks because it is one of the main types of risk that initiate bankruptcy [22]. According to [10], there are basically four factors involved in it: loan pricing, specific credit limits, security to reduce risk, and the diversification. There are different models for determining banking credit risk, but the majority of which are optimisation models about asset-liability proportion that are under the conditions of capital requirements of Basel II. Basel II consists of extensive set of supervisory standards to improve risk management practices, which are structured along three mutually reinforcing elements. The most relevant element is the one that addresses minimum requirements for credit and operational risks. The principal objective is the stability of the international banking system via better risk management. This is done by imposing regulatory capital requirements more in line with current bank good practices. [33].

There are several bases to help manage the credit risk and one of them lies on the correct data extracted from the reports that describe certain processes within a company. In order to control this data, there is a tool commonly used in Financial Reporting: Internal Controls. Internal Controls (IC) are important to avoid management failures. [39] states that control is a practice established to guarantee that business processes are carried out reliably, securely, with the proper agreement and in a prearranged manner. Moreover, in financial reporting, IC are recognized as being fundamental to excellent information systems and high-quality financial information [3].

An analogy can be used to get a more thorough understanding of IC. This involves the consideration of weaknesses to measure and analyse the control. A material weakness in internal control is defined as ‘‘*a significant deficiency, or*

combination of significant deficiencies, that results in more than a remote likelihood that a material misstatement of the annual or interim financial statements will not be prevented or detected" [29]. A hazard such as this highlights the importance of detecting mistakes in order to maintain good management of IC. Moreover due to the defects of financial reporting in recent years, Sabanes-Oxley (SOX) was legislated to improve reliability; therefore, most of the controls adopted pursuant to the act are concerned with the timeliness, accuracy and integrity of financial data [39].

The effectiveness in the use of the IC over information is included in the quality of information systems and it has an effect on main decisions in businesses [18]. In addition to addressing the issues of inaccuracy of financial data, the IC can be extrapolated to control the quality of any kind of data. Therefore one of the current main issues nowadays is the relation between IC, data quality and Information Technology (IT). It has been reported that IT competence was an important differentiator in banking that were performing well in the 1980s [25].

A recent white paper by the SANS institute notes "*while the topic of information security is not specifically discussed within the text of the act, the reality is that modern financial reporting systems are heavily dependent on technology and associated controls*" [34], [35]. It is dependent because IT is the tool to enable a report to be written within the company in terms of time and accuracy. In general, today, enterprises are very dependent on IT for the management of business processes. [34] states that, due to the interconnected of IT systems and infrastructure, the growing complexity and constantly changing threat and regulatory environments, IT IC need to be implemented to mitigate those risks. Basically IT IC are used to capture, manage and record raw transactional data of economic events and encourage the elaboration of financial reports [34]. Furthermore, the majority of the reports that consider the issue of IT IC have reached the same conclusion: companies with IT control deficiencies report considerably more IC deficiencies than non-IT companies [13]. In addition [19] stated that firms unable to implement adequate controls over IT systems and not making those important investments, control weaknesses will cause much

bigger expense in identifying weaknesses and assessing related risks. Even though SOX requires the assessment of IT IC, compliance should be considered as a way to help the business manage its different risks, and quality IT IC should be recognised as being financially beneficial and not as mere additional expenses with any benefit. Firms should therefore carefully determine the level of IT IC needed to secure its IT systems and infrastructure, and thus viewing that effective IT controls are needed to fully understand the potential of IT investments whilst mitigating connected risks [34]. With regard to the bank environment, the IT and IC are related to the data quality. The similarities between financial reporting and data quality reporting are in terms of reliability and accuracy of the information ready to use.

Data is very important in modern day systems. When used in the correct context, it translates to information which can be used to lead strategies in order to compete in the market. Therefore data quality is highly necessary for a company to be considered a strong competitor. Quality has been defined as "T*he conformance to standards"* [4], [32]. Statistical quality control is a process that ensures that output meets standards. This means setting standards, deriving statistics, collecting data and taking corrective actions when statistics point out that the process is out of control [16], [21]. Data quality is one of the critical problems facing organizations today. As management becomes more reliant on information systems to complete visions, data quality becomes a bigger issue in organizations. Issues with data quality may cause real losses, both economically and socially [12]. This could damage the relationship with the customer.

Information quality management is a complex process and research in which interest is growing exponentially. Data quality has economic wide benefits. As a matter of fact, studies have highlighted the importance of information quality to a wide range of business domains, including Enterprise Resource Planning (ERP), Data Warehousing, Supply Chain Management (SCM), advanced Data analytics/mining, and product data management [7]. However, there is not one single definition of data quality. Accuracy, consistency, timeliness, relevancy,

completeness, and fitness for use are among the variables most frequently used to define it. [24], [12]. Potential issues of data quality may include aspects such as migration problems and systems integration, age and redundancy or relevance and the temptation to collect data among others. [26], [17], [24]. However, future work is needed to extend the area of data sharing, as this capability continues challenging even the most advanced organizations due to the importance of data sharing. Data sharing is considered as a link between data quality and organisational learning, and that data quality deals the sharing to performance relationship [27].

Due to the importance of data quality and the knowledge of how to process it, the software used to storage, process and analyse it has become increasingly important. One of the major tools used to process and analyse data is the spreadsheet. 80% of people are currently exposed to the use of spreadsheet which is now an essential tool for modern businesses. Currently, thanks to the spreadsheet's ability to store and analyse essential business data, spreadsheet technology let the firms reduce hardware and software costs, so even in the smallest company, spreadsheet applications are now common place [28]. The propagation of spreadsheet packages is triggered by reporting limitations associated to existing transactional and Business Intelligence (BI) systems. However, manual data analysis and transformations of data are the primary reasons for bad information quality whilst using spreadsheets [7]. This means that using a tool with repetitive tasks and manual adjustments could cause human errors that are difficult to detect. Hence there is a necessity for a system that helps to automate the IC in a company that would assist in minimising the human errors and hence reducing the time for each task. Furthermore, in a business environment there is demand for constant improvement to adapt the software to the new requirements that are introduced. To improve the efficiency and effectiveness of a software development and to develop software products, there is a systematic approach named Software Process Improvement (SPI) [36]. There are different approaches to provide criteria, measurements and strategies to assess improvement initiatives. [14] conclude that the effect of

changes in the requirements process can be observed and measured at different levels:

1. Endeavour and quality of requirements related activities in the requirements phase

2. Project success in terms of budget, meeting time and scope constraints

3. Product success in terms of achieving both the customers' and the company's expectations

4. Company success in terms of market strategies and product portfolio

5. The influence on society

Another concept that could be used is in this context is Software Re-engineering. According to [31], "*Re-engineering is the examination, analysis and alteration of an existing software system to reconstitute it in a new form, and the subsequent implementation of the new form.*" Therefore the main objective is to understand the existing developed software (function, architecture, design, and implementation), comprehend the structure and requirements of the desired software, and understand the relationship between the AS-IS software and the TO-BE. With all this information, the next step is to re-implement it improving the performance of the system fulfilling the new requirements set by the analysis.

As time goes by software loses its efficiency and it needs to be adapted to the dynamic business environment, therefore re-engineering is considered. Agile methodology is recently used in software development as an alternative to the traditional methodologies. Agile software development follows the idea of continuous re-engineering by the Extreme Programme methodology [1]. It is a different method used to develop IT projects. As a lightweight development method, the main basis of this methodology is to develop and implement software in short incremental iterations. It can be developed as part of functions to implement specific performance of a system, be re-engineered to improve it, and subsequently develop the next part of the requirements later on. Agile

development speeds up the delivery of business value. Through a process of continuous planning and feedback, this method is able to ensure this value is continuing to be maximized during the development process. As a result of this iterative planning and feedback loop, teams are able to continuously align desired business needs with the delivered software, easily adapting to changing requirements throughout the process [37]. Nevertheless, this method does not fit in a standard classical software development methodology (waterfall and v-model) [15]. The main principles of the Agile methodology are: Customer oriented, short time scale deliver, bringing business people and developers together, and face-to-face conversation to build a good environment and support. In addition to these, attention to technical excellence and good design, sustainable development and simplicity should be considered [9]. Extreme programming is one of the most commonly implemented methods to apply the Agile methodology. The project can be developed with four simple steps: Planning, design, coding and testing. It is a dynamic method that allows for the constant reception of feedback because of the focus on customer demands [8].

In brief, this study contributes to the research about the issues of IT IC of data quality and the necessity of automating them to reduce the human errors and hence improve the value of the output. Sometimes these IT IC are difficult to implement manually and highly time consuming, therefore there is scope for the application of an automated system for data quality IC. The improvements of the automated process can be based on the requirements of the users and measured against the criteria that have been previously ascertained.

The literature review has the purpose of revising the same concept of IC in financial reporting already applied to the Data Quality department. Automated IC about quality of data with reports as the output of the analysis that inform about the data stored in the database. Having the same structure of taking care of the accuracy and correctness of the data, allows to be assured the highest quality of data. The relation with the IT and the IC means relevant reduction in human errors and speediness of the analysis of the process. Finally, the last step of automating the process can be completed based on Agile methodology

[9]. After the research in the literature, Agile is thought to be the most suitable software development methodology for this project based on the main Agile Principles mentioned above. The project is characterised by fast delivery due to the short frame, continuous changing requirements, co-operation with business experts and constant testing. Moreover, the project did not include defined phases, require formal documentation or approvals methods for delivery, therefore to use standard software delivery approach using Waterfall method was discarded.

The identification of existing internal controls and reporting of Data Quality in a bank, and the link with automation requirements and software process improvement, were the focus of this research. Therefore, the gap found in the literature is about IC to ensure the quality of the data, reporting about this data quality status and its automation. There are already IC about financial reporting enacted by SOX as mentioned earlier, but there is no literature on using internal Data Quality Controls and Reporting to help ensure quality of data and be implemented through an automated process. Due to that, there is a necessity to control that the data stored and used is reliable because it could be an enormous cost for a company and even more if errors on data have customer impact. Also decision making is another reason why data quality becomes important.

Consequently the **aim of this research project** is to develop an automated process and software tool that help improve and control internal banking processes as well as managing the data quality assessment and monitoring the retrieval of the pending position status of loan processes in banks.

The **objectives of this research** are as follows:

1. Identify the most common practice and perform a literature review to acquire a deeper understanding of the related banking products, control processes in business, existing IT control models, data quality, software improvement methods and procedures

2. Requirements analysis of the current banking products and processes through interviews and literature review to establish a clear view of the required improvements and applications to be developed.
3. Develop code using existing client software to automate the bank's daily IC in order to analyse data quality from the database.
4. Ensure the quality of data by implementing the automated process and software developed reducing the time employed and minimising the errors on data.
5. Develop a software tool to retrieve the pending position status of a loan. This application should reduce the errors that originate in the core platform with high relevance due to the large customer impact that it can have.

The scope of the project is the development of a software tool that includes features to manage the existing IC, improving and automating the control process efficiently using software re-engineering standards and Agile methodology. This research also covers the improvement of data quality and the monitoring of the retrieval of the pending position status of a loan.

The structure of the paper is as follows. Section 2 describes the methodology. Section 3 presents results and validation of the proposed solution. Section 4 outlines the discussion, conclusion and future work.

2 MATERIALS AND METHODS

The methodology was set up thoroughly to match the aim and objectives established in the previous chapter and it was structured accordingly following extreme programming methodology:

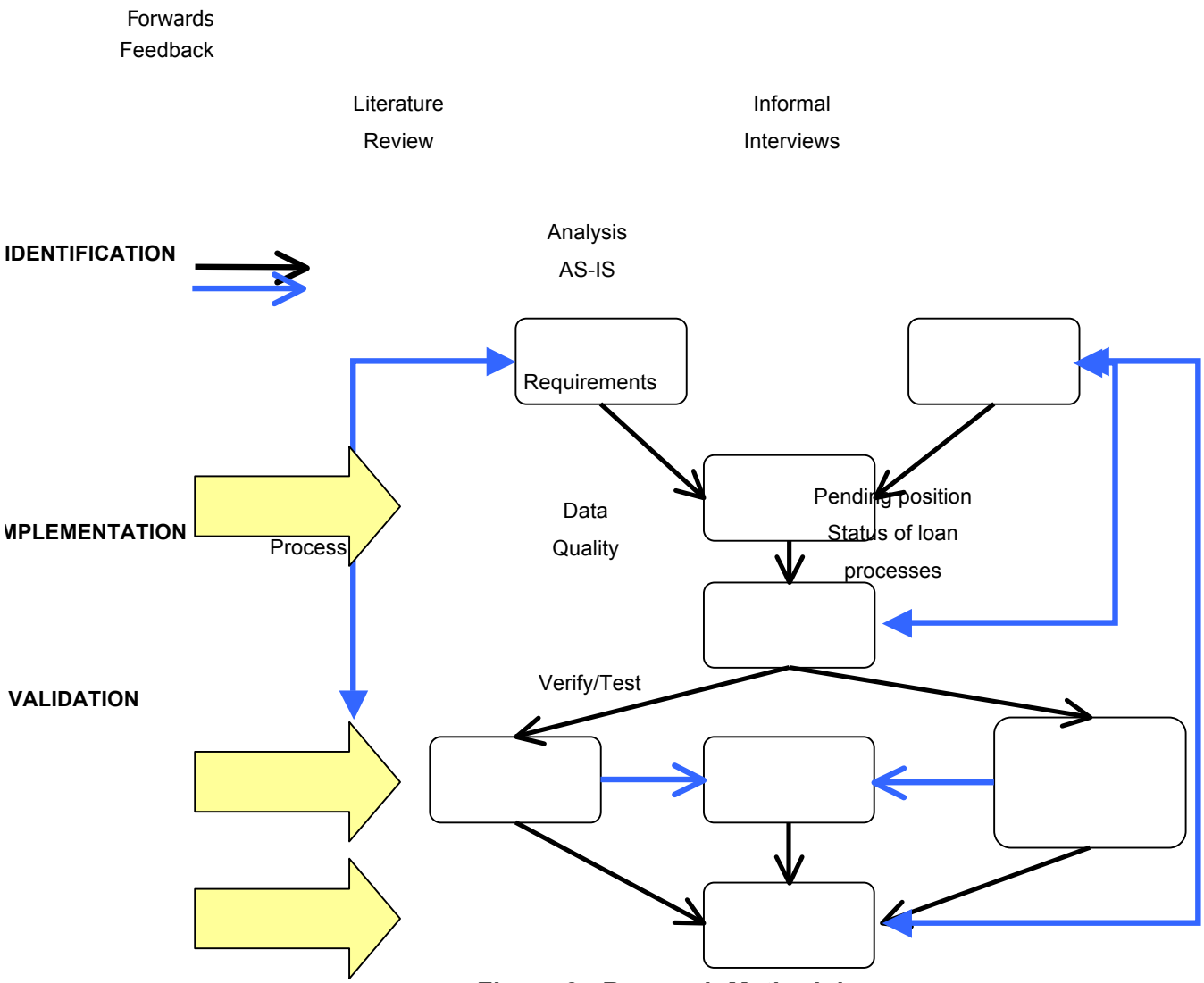

Figure 2 - Research Methodology

There are three main phases in the methodology: Identification, implementation and validation. In the identification stage, the main goal is to identify the possible requirements of the existing process and the main improvements to incorporate. This was accomplished through a literature review, internal research at the bank and informal meetings with business experts, end-users and the project manager. The author was able at the end of this stage to define the key performance indicators (KPI) that would be analysed on the validation phase. In the second stage, the implementation phase, the existing process was described with a flow chart tool and improvements were suggested through this visual tool. The business experts were consulted about these improvements by way of informal meetings and the testing phase was conducted by evaluating the continuous feedback, primarily from the end-user and modifying the program accordingly. The third phase was the analysis of results and validation of the program. The validation was carried out in the last phase of the testing stage. A group of business experts, end-users and the project manager was

selected to run the tool with the user guide and analyse it and validate it against the KPI defined in 2.1.4.

Usefulness of these methods has been proved by using the defined methods in the first section through a successful development of the applications which helped to achieve the goals of this project. A detailed summary of the followed steps is provided below to meet the objectives:

1) Understanding of current processes using re-engineering methodologies and set up functional and data flow documents to identify existing tasks of the controls process tool.

2) Continuous detection of requirements and improvements identified together with business experts and end-users collaboration.

3) Analyse the current method of IC for lending products and develop an automatic tool based on software Re-engineering. The application was initially developed to reduce the time consumed running the process everyday. But the other important objective was to monitor the critical data and track the possible errors that can occur within the core platform. According to the meetings with business experts, there were changes that were not identified during the requirement phase and they were incorporated to the process to meet specific needs. Data quality was tested and verified via Agile methodology.

4) Assurance of data quality by implementing the automation of control process and software code developed decreasing errors on data and reducing time consumption via elimination of manual process.

5) Understanding the current method of calculation of the status in mortgages and loans and develop an automatic tool and mechanical algorithm calculation based on software reverse engineering according to the literature. The application was mainly developed to minimise the time employed on this calculation with every account that appears as an anomaly detected by the IC process run daily.

2.1 Set up the goals of the control process improvement

2.1.1 Analyse the current situation of the control process of lending products through face-to-face informal interviews, database research and literature reviews

After an extensive literature review to understand the context of the subject a wide learning process was developed from the internal documents of the bank to learn about the different banking products, processes and data storage. In the following, the products this research is referring to are mortgages and loans. There are many different types of mortgages and loans that banks are currently offering. Therefore, an extensive research in these specific products of this bank was carried out in order to get a better understanding of the whole field. Information was also extracted from internal documents available at the bank. This phase was also conducted along with several face-to-face discussions and meetings with the current end-users of the control process and the manager of the project. This helped to understand the existing flow process that was being run to control the data about the banking products as well as how the internal Core Banking Platform works to reproduce the path of the movements that the accounts follow within the core database. The information captured was revised and confirmed by the business experts. In addition to this, some of the gaps in the information were filled that previously could not be extracted.

Once the existing process was analysed and understood, setting up the possible improvements of the control process and the most efficient way to get it automated was identified as a necessity. This automation was one of the largest requirements in terms of key performance indicators (explained in Section 2.1.4). Based on the Agile methodology, the analysis of the requirements also implicated comprehensive end-user involvement. During this stage, potential improvements and new functionalities were identified to incorporate. These were analysed and consulted with the business experts in terms of feasibility, functionality and relevance so the key performance indicators that the author identified could be defined.

2.1.2 Precise mapping of the control process for

2.1.2.1 Automation of the bank daily IC and data quality assessment

A flow chart tool was used to get a clear idea of the functions in each stage of the control flow. The IC process is used to supervise the core platform in processing the information correctly and to investigate the possible irregular movements that can appear within the database.

According to the literature, a software reverse engineering methodology was used to get a thorough understanding of the entire analysis because the importance of *learning by doing* how the existing control process worked was identified. Also the questions that appeared during the learning process to ensure the correct use of the control process and its interpretation were checked and confirmed by the relevant staff.

With regards to the process, the movements within the customer accounts are tracked in the core platform as a method of accounting. In concordance with the nature of the banking product, there are certain movements that cannot occur. Occasionally, due to inappropriate working standards of the core platform those movements may happen; this is the reason controls are run daily. According to the literature, IC can ensure the quality and correctness of data that represents the movements within the database as well as keeping track of any past issues to prevent them from reoccurring.

Currently due to the large number of lending products sold by banks, the amount of data generated is extremely big, so in order to check that every account is functioning correctly, there are some reports generated automatically overnight. These reports are used as a source to produce the daily summary report and the individual tracking reports by the Data Quality and Structural Systems department (DQSS).

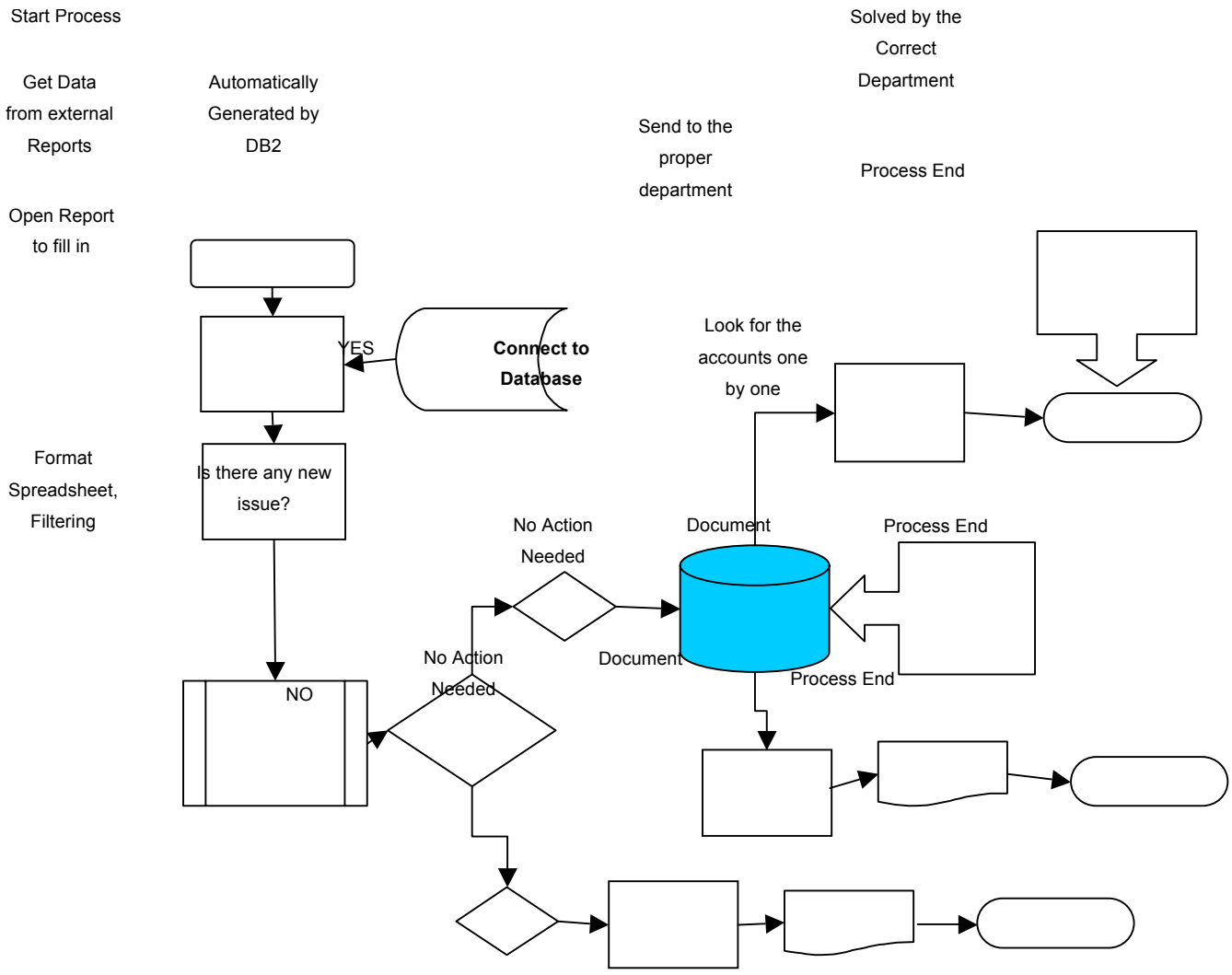

Figure 3 - Data monitoring and data quality assessment (AS-IS)

2.1.2.2 Control of pending position status of a loan processes using Flow charts

The manual process that was being executed was represented in figure 4. This practice helps to allocate the requirements needed in each stage and the already existing bottlenecks in some of these stages. In order to be able to understand this process, the knowledge background about loans and mortgages mentioned in the introduction chapter, was necessary. Also to get a better understanding of the process, reading documents provided by the bank and having several interviews with business experts in this field was also a requirement.

One of the controls mentioned previously was designed to examine the quality of the data processed through a tedious manual process. This was one of the main issues that the process had to face according to the information extracted

from the interviews with the business experts, specified as the time used to finish the task. Moreover in the bank environment, one of the main concerns is the customer and the impact that any mistake has on it. If any issue has customer impact, it has to be resolved with higher priority than if not. This is why they were chosen as some of the Key Performance Indicators (2.1.4).

Pending positions of capital and interest of a loan or mortgage were analysed. This is an accounting field within the core banking system that is populated when a loan requires a payment. When the settlement date of a loan appears, the pending positions are automatically populated until the collection day arrives. The settlement date is the date when the customer has to make the monthly repayment; this is the same day each month since the loan was first opened. These pending positions can be occupied up to 30 days. When this position exceeds the tolerance of this time period, an investigation must be initiated by the appropriate department.

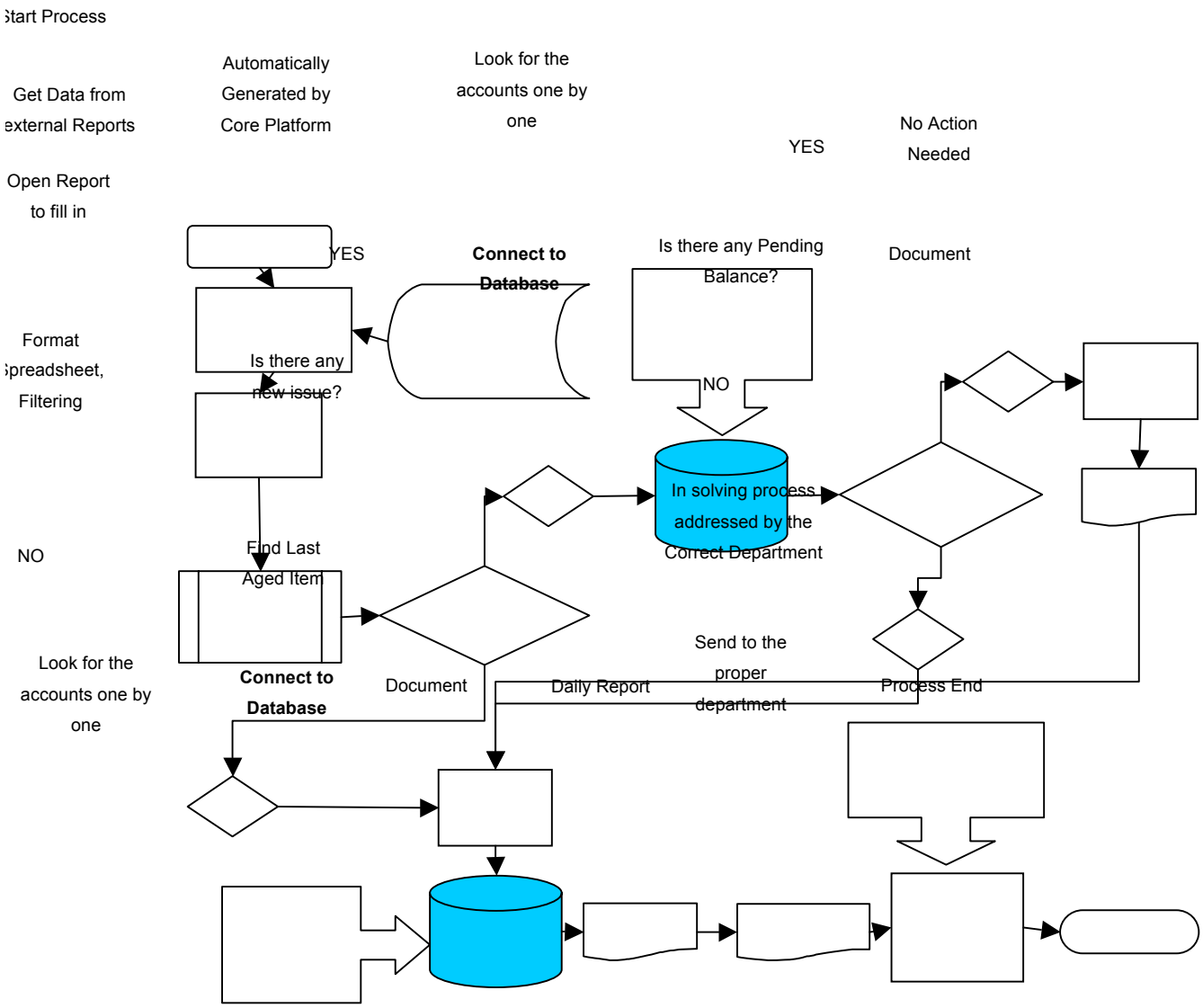

Figure 4 - Pending position status of a loan processes (AS-IS)

2.1.3 Establish a plan of improvement for the process based on the mapped processes

The main issue addressed by the current users of the control process was the time employed in running the controls. This concerned the accumulation of multiple issues with the customer accounts to deal with and the tight deadline to deliver the daily report was relevant to consider the time as the main performance indicator. The starting point of the plan was to speed up the process to control the possible errors on the customer accounts. It was also to reduce the intervention of the user and therefore minimise the number of manual errors. According to the literature it has been pointed out that manual data analysis and transformations of data are the key reasons for bad data quality whilst using spreadsheets [7]. So the solution was thought by using an

automated tool with repetitive tasks and manual adjustments to avoid human errors hard to detect.

The project was designed to be a continuous testing process with as many feedback sessions as possible. That means a quick delivery between feedbacks and daily cooperation with the end-users. Based on the information found through the literature review, Agile methodology was used in this stage of the process. Due to the characteristics of the project development the collaboration with the business experts, delivering software and dynamic plans adapted to the testing phases and corresponding feedback could be emphasised.

	Metric	Current value	Target value
Cost	Customer Impact	Loss of confidence	
	Time consuming (Process 1)	5 hours	1:30 h
	Time consuming (Process 2)	2 hours	2 min
Quality	Data quality	Approx 10 new accounts with issues per day	No issues
	Data quality	Hazard non detected accounts	Complete Detection
Performance	Implementation	(June to July 2013 – based on weekly average)	
	KPI	Time	
	KPI	Accessibility	
	KPI	Flexibility	
	KPI	Reliability	

Table 1 - Plan of improvement

2.1.4 Identify the key performance indicators of the improvement by

2.1.4.1 Measure the relevance of the improvement

In order to measure the relevance of the improvement, several indicators have been extracted from the interviews and the literature and have been identified as the key ones.

- Time employed
- Accessibility for the end-user
- Reliability of the tool
- Flexibility of the software

2.1.4.2 Defining measures for assessing the target performance

Every indicator has its own measure unit to be able to track the results of every test in the testing phase.

- Time measured in hours, comparing the time employed in the previous control process with the new one, to check the improvement in terms of time and percentage of improvement.
- Accessibility measured in terms of how easy can understand and run the control process correctly the new user. The learning process speed measured in number of days that takes to run the controls without any question or doubt.
- Reliability is calculated with the number of errors produced by the control process and compared it with the previous process to get relative results.
- Flexibility is measured in terms of adaptability to improvements. Once designed, tested and implemented, future work will be necessary to improve the control process. With flexibility and adaptability, the control process is thought and built to be able to adapt the possible changes that could occur in this dynamic environment.

2.2 Develop and implement tools

2.2.1 Develop a theoretical tool for the automation bank's daily IC and data quality assessment based on interviews and literature reviews and its implementation

Code	Name	Responsible	Description
00			Start Process
01	Select Report actual date	DQSS	The external reports are automatically generated every night and send by email. The information from these reports is used for the daily controls.
02	Open the previous day report	DQSS	Every day is produced a report that combines a summary of all the controls.
03	Run Macro to Format spreadsheet	System	Select information that is needed, formula, and other formatting. Filter by criteria by the conditions allowed for each control
04	Is there any new account with an issue?	DQSS	If yes, go to step 05. If no, go to step 09.
05	Run Macro calling a query to a DB2	System	It pulls the data from DB2 into the spreadsheet straightaway
06	Is there any issue/no pending balance?	DQSS	If yes, there is not an issue, go to step 08. If no, go to step 07.
07	There is no issue on the account	DQSS	No action needed. Fill the indicated field columns
08	There is an issue on the account	DQSS	Fill the indicated fields columns
09	Fill Report	DQSS	Fill the report with the date and action to take with each issue
10	Address Issues	DQSS	Send the issue to the department that can solve it
11			End process.

Table 2 - Data monitoring and data quality assessment procedure description

The next stage focused on the analysis and implementation of new functionality. The base tool is the spreadsheet. In every control there are certain rules to check if the information extracted from the report is correct. If any issue appears, it must be checked in the core platform to track the cause of the error. Once the root of the problem is found, the appropriate department must be addressed in order to resolve the issue. Checking accounts one by one is highly time consuming. The solution would be to make a query through database management software. However, this lacks efficiency when it has to be included in the daily report with a deadline. The solution was to call a query that could bring the data and paste it directly into the spreadsheet. It was also identified that formatting the report manually could cause errors on data and its results and was also highly time consuming. So the automation of formatting was considered as an opportunity of improvement as well.

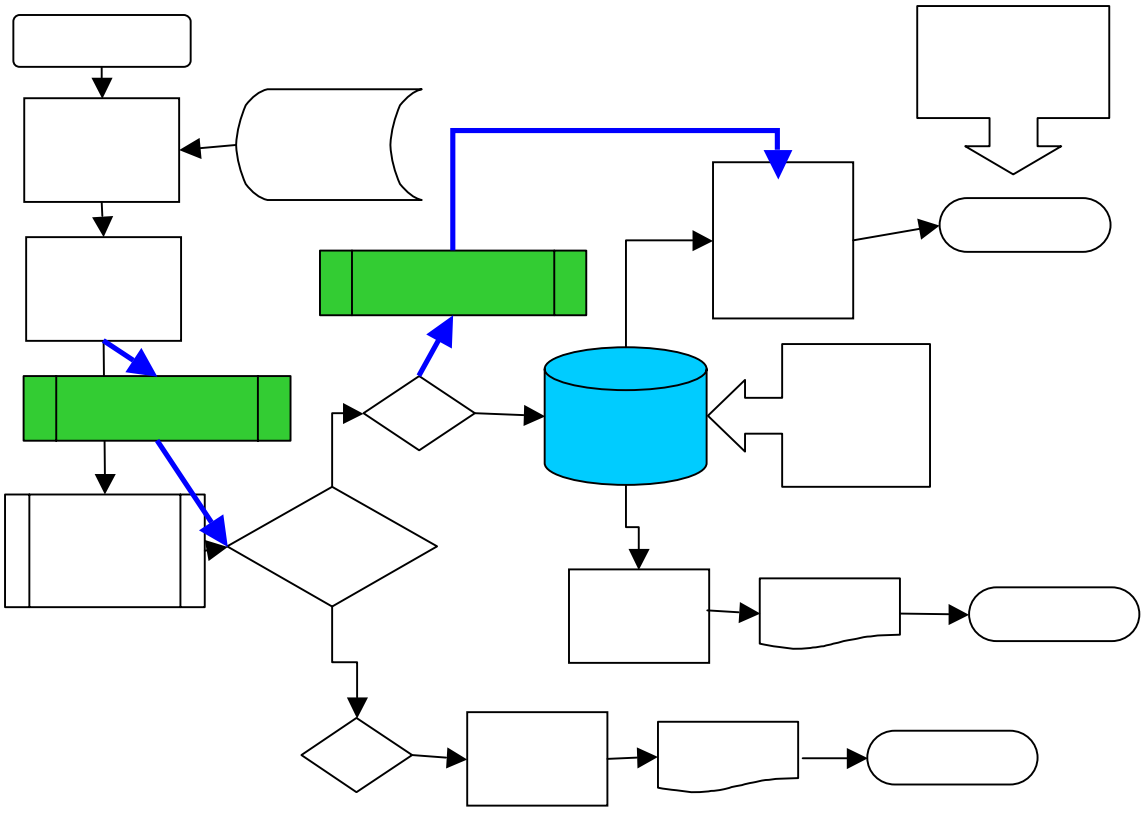

Figure 5 - Data monitoring and data quality assessment (TO-BE)

During the implementation stage of the tool, the improvements were identified as requirements to optimise the validation process and suggested to the end-

users. The implementation of the tool was carried out straightaway. According to the nature of the tool, once the first version was developed, it was automatically implemented and tested to check the possible errors and to receive some feedback from the end-user. However during the testing phase, the lending controls were run by both methods at the same time in order to get results to compare and to make sure that the new tool worked correctly.

2.2.2 Develop a theoretical tool for an automatic retrieval of pending position status of a loan based on interviews and literature reviews and its implementation

Code	Name	Responsible	Description
00			Start Process
01	Select Report actual date	DQSS	The external reports are automatically generated every night and send by email. The information from these reports is used for the daily controls.
02	Open the previous day report	DQSS	Every day is produced a report that combines a summary of all the controls.
03	Run Macro to Format spreadsheet	System	Select information that is needed, formula, and other formatting. Filter by criteria by the conditions allowed for each control
04	Is there any new account with an issue?	DQSS	If yes, go to step 05. If no, go to step 09.
05	Run Macro calling a query to a DB2	System	It pulls the data from DB2 into the spreadsheet straightaway
06	Is there any pending balance?	DQSS	If yes, there is not an issue, go to step 09. If no, go to step 08.
07	There is no issue on the account	DQSS	No action needed. Fill the indicated field columns
08	There is an issue on the account	DQSS	Fill the indicated fields columns

Code	Name	Responsible	Description
09	Filter by criteria	DQSS	Filter by criteria by the conditions allowed for each control
10	Find Last Aged Item	DQSS	It has to be found the Last Aged Item
11	Run Macro calling a query to a DB2	System	It pulls the data from DB2 into the spreadsheet straightaway
12	Run algorithm to find the last aged item	System	It makes calculations automatically to find the last aged item and its date.
13	Fill Report	DQSS	Fill the report with the date and action to take with each issue
14	Address Issues	DQSS	Send the issue to the department that can solve it
			End process.

Table 3 - Pending position status of a loan processes procedure description

The next stage is the analysis and implementation of the new functionality. At DQSS an Excel based tool is used to validate positions status of a loan. Because this validation is highly important in terms of customer impact, it is necessary to ensure that the repayments are allocated correctly in each customer account. Therefore it was also vital to ensure that the process of bringing the correct data from the database worked as expected.

Based on the flow chart of the as-is process, the best way to automate the current process and get the same result obtained with the manual calculation was consulted with the business experts. It was decided that to investigate and clarify the status of aged items in the pending positions of a loan account, the solution was to create a query through database management software to call the data from the core database platform and paste directly into the spreadsheet and then write an algorithm that reproduces the same function that was being run manually into the core platform for each account at a time.

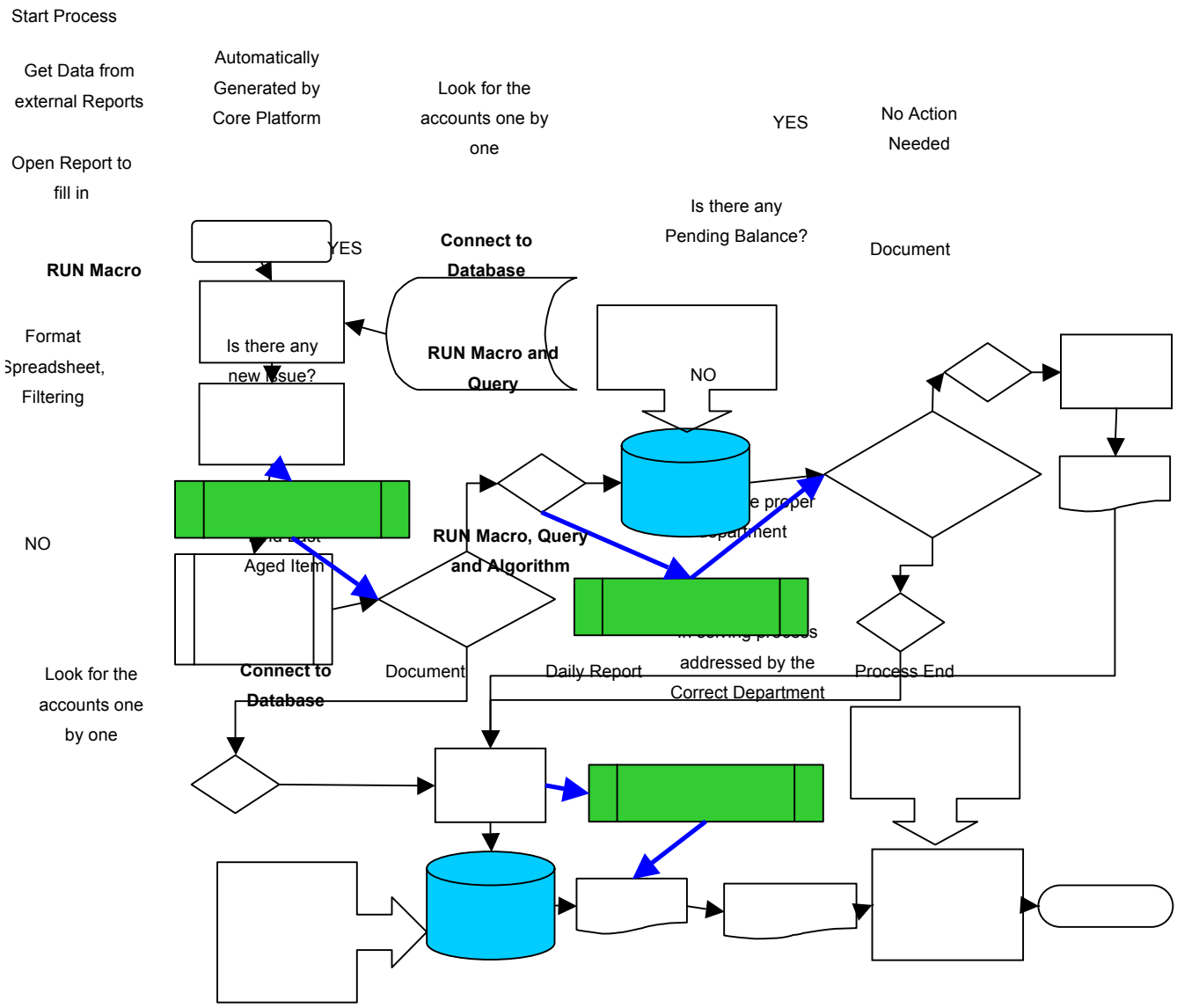

Figure 6 - Pending position status of a loan processes (TO-BE)

Before the implementation stage started, data from the core database was collected and a dummy database was built to compare the results. Once the tool was developed, it was directly implemented to run and store the data in the parallel database. The data generated by the tool, was compared against the data stored in the real database in order to check that the tool was correctly implemented. During the implementation stage, modifications required to optimise the verification process were identified and suggested to the business experts.

2.3 Test and validate the process through case studies and expert judgement

The validation of the process was carried out according to Agile methodology over several sessions along with the project manager and the business experts presenting the beta phase of the tools. The presentation included a brief demonstration and a hard copy of the user guide to enable full understanding of the platform at the start of each session. After the first session, feedback was solicited and the improvements were incorporated into the tools again. Once ready to be presented, a second session was arranged with the final tools to get final feedback and prepare the tool to be running as the definitive one.

3 RESULTS

The results of the process were produced during the test phase. It is the last phase according to Extreme programming [8]. As mentioned earlier, the testing lasted a month. The results were recorded and analysed so as to extract the conclusions and check if the objectives were achieved.

In order to verify that the tools were working as intended, a mirror database was created, with which the results were compared. Data quality was checked and verified using Agile methodology. By checking the errors against the duplicated database the cause of these could be analysed and tracked.

3.1 Automation tool for bank daily IC and data quality assessment

The test phase was planned to last one month having daily meetings with the end-user as the tester and weekly meetings with the project manager and business experts to get a brief summary of the situation, new requirements and possible improvements. On a daily basis, the tester asked any enquiry or any debug as soon as it appeared to be resolved.

The automation process was measured in terms of the key performance indicators defined previously. About the time, it was observed that with the new

tool, it took around 1.5 hours to run the entire process and produce correct output whereas with the old process was 5 hours. This means a reduction of 70% of the time employed.

With regard to accessibility performance, the learning process of the end-user to use the tool without any problem, lasted 15 days. Due to the particular conditions in the department, this indicator was possible to be tested with a new user. With a brief guideline document and the feedback meetings, it was observed that the learning process got faster as the time went by. With the old model, the learning process was tested by the user and took 30 days to learn how to run and format it correctly. So the improvement in terms of accessibility was a reduction of 50% of the learning phase.

In terms of reliability, it increased up to 100%. The old control process and the new one were run simultaneously to check for any differences that could appear. The results were compared, and the new model produced 0 errors whereas the old control process generated an average of 5 issues per day. The results of both methods were verified against the backup data used for testing from the database and also verified by the end-user. Also any discrepancies that appeared were investigated, recorded and the tool was modified in concordance.

In regards to flexibility, it was checked if the tools are built to be easily adapted to new changes introduced. This was measured in terms of time employed upgrading the application, concretely in number of hours that were employed to modify the tool. After meetings with the end-user and business experts some of new features were required to be incorporated to the developed application. The average of the time employed to introduce the changes into the tool was 1 hour. This could be adapted with easy due to the building structure of the application. There is no comparison with the previous control model because of the inexistence of automation on it.

KPI			
	Old Model	**New Model**	**% Improvement**
Time (hours)	5	1.5	70%
Accessibility (days)	31	15	52%
Flexibility (hours)	0.00	1.00	-
Reliability (errors)	5	0	100%

Table 4 - KPI Data monitoring and data quality assessment

3.2 Software tool for an automatic retrieval of pending position status of a loan

The development of the process was divided in different phases as mentioned previously. According to Agile methodology, it was combined with continuous meetings to request feedback and implementing the requirements identified. However the test phase was also planned to be one month.

The results were measured and tested against the key performance indicators as well. In terms of time, the results were tracked during the entire testing phase. In the manual process, the time could vary depending on the number of new issues found, and it could increase exponentially with the amount of customer accounts with issues. Nevertheless with the automation tool the number of issues does not affect the time employed. The process before the automation with a daily average of ten new accounts with anomalies took two hours whereas with the automated tool this was reduced to two minutes. This means 97% of reduction of time, which is a significant advancement in terms of time and economic cost.

With regard to reliability of the software tool was measured in number of errors generated by running the process. It was observed that the number of errors was reduced to nil when the corrections of the algorithm were made. One of the main benefits of this tool is that there is an evidence of the control being checked whereas with the manual test, the checking was not recorded anywhere. Furthermore, the potential for error is removed by eliminating the manual process.

In terms of accessibility for the end-user is measured by teaching hours that a new user needed to understand and be able to run the process. The teaching phase for this software tool took two hours in total whereas with the old control process took five days to be able to interpret the results right.

Flexibility of the system was measured in terms of hours employed to adapt the software to changes occurred in the production of reports. Due to the change of the layout of the data and the reports source from where the data used was extracted, some changes had to be incorporated to the first draft of the automation control process. The tool was tested against all the feedback sessions with the end-users and the improvements could be incorporated to the tool in two hours on average. This indicator could not be compared with the old process due to the lack of automation.

KPI			
	Old Model	**New Model**	**% Improvement**
Time (hours)	2	0.07	97%
Accessibility (days)	5	0.08	98%
Flexibility (hours)	0.00	2.00	-
Reliability (errors)	10	0	100%

Table 5 - KPI Pending position status of a loan processes

3.3 Validation

The validation phase was performed during the testing stage. The validation of the software tools was carried out by the end-user, who was assigned the responsibility of testing the control process and feedback all issues and potential improvements to the designer. However the control process also had to be validated by the business experts and the project manager. It was concluded from this process, that the software tools would enable the control process to be more efficient and would result in the current control model being more effective. These improvements will result in immediate and accurate identification of any issues, while ensuring that any potential for “human” error is removed. The validation of the data quality is achieved through the automation

of the process. By eliminating the manual process the potential for error is removed.

4 DISCUSSION

This thesis provides insight into the existing IC and reporting of Data Quality in a bank, and the link with automation requirements and software process improvement. One of the objectives of the thesis was to develop code using existing client software to automate the bank's daily internal controls to analyse data quality from the database. This would allow for the use of an effective and efficient tool during the control process in order to detect the potential flaws on data with the shortest possible time.

Firstly, to achieve this objective, the current control process had to be analysed and represented in a flowchart in order to identify the key points of improvement with the automation. Subsequently, an application was developed with a user-friendly interface using VBA to automate the attachment and combination of several reports and the formatting of the final report. This application was run and tested successfully. The tool is currently used by the end-users to generate the daily reports. The automation of Data Quality Internal Controls process and Reporting has proven to be a suitable process to improve the quality of data, minimising the errors and also reducing the time employed according to the results and the conclusions from the validation. The application can also be modified and adapted to compare data from other sources such as financial data among other uses.

The next objective was with regard to the Data Quality topic to ensure the quality of data by implementing the automated tools via software code developed. The attributes of data quality were extracted from the research done in Chapter 1. This objective has been met with the application and software tool developed during the testing phase. Therefore, the accuracy and quality of data is ensured through the implementation of automation of the process. Data quality is supported by the reduction of the number of errors which is consistent

with the conclusion found on the research stage regarding the conclusion about manual data analysis and transformations of data that are the key reasons for bad data quality using spreadsheets.

The final goal was to develop a software tool to retrieve the pending position status of a loan. This allows for the data to be reliable and robust which also helps to accomplish the previous objective. In order to understand the process, a flowchart was mapped with the help of business experts and project manager knowledge, which enables any necessary changes to be made with ease in future. This objective was met with the development of user-friendly application and software code (VBA) that enables the user to integrate the two main software clients MS Access and MS Excel. An algorithm was also written with code to reproduce the correct calculations, based on the core platform algorithms, so the movements can be tracked easily. The tool increases the efficiency of the process exponentially and therefore the time taken for the number of issues to be resolved is no longer relevant. Moreover, the automation of the process via the implementation of the new software tool has led to a vast reduction in customer impact through early identification of the issue.

There are several limitations in this study that must be accounted for. The main issue was the short time frame to carry out the whole process. With this in consideration, the scope could not be as broad as desired and all the stages of the plan (case analysis, software design, software testing, User Acceptance Testing) had to be planned out meticulously for all of these elements to be completed. The objectives were also conditioned to this constraint because scope creep could not to be tolerated.

The project has to be completed with existing software and hardware. The resources available at the bank set the boundaries of proposed solutions. Due to this, the procedure and the final output had to be adapted to this constraint.

The final issue to be considered is the coding used to develop the tools. This could not be overly complex, as it would need regular maintenance and modification, which would be difficult to resource externally.

The constructive findings obtained from the study, have potential implications on the scope beyond that of the project. There is a new depth to the problem that was not initially considered at the start of the project because the field of study has been expanded. The control process needs to be constantly improved so that the new potential issues to be detected. Therefore, due to the necessity of maintaining the tracking of possible errors on data generated by the core platform, the control process is a priority in any environment. Furthermore, there is opportunity to create new controls and automate them in order to improve the quality of the data stored thanks to the reduction of time consuming with the current control process.

5 CONCLUSIONS

There are several conclusions that have been gained from this project. As mentioned in Chapter 1, the focus of the study was the identification and analysis of existing internal controls and reporting of Data Quality in a bank environment, and the relationship with automation requirements and software process improvement. The gap found in the literature was with regard to IC over data quality and status reporting and its automation. There are already IC in financial reporting in place but there is no literature on using internal Data Quality IC and Reporting to help ensure quality of data and implemented as an automated process. For this reason, there is a necessity to control the data stored and to ensure that it is accurate and reliable because it could provoke significant cost for a firm and even more when errors on data may affect customer experience. The aim and the objectives of the project were set up accordingly in order to cover this requirement.

The main objectives were to develop an automated process through software tools that help improve and control internal banking processes as well as managing the data quality assessment, and monitoring the retrieval of the pending position status of loan processes in banks. Due to the characteristics of the project and the environment, Agile methodologies were followed during the design, development and testing stages.

The findings have various conclusions. Firstly, the automation of the software tools, the time of the process has been reduced in 70% and 97%, which has created a significant improvement in terms of efficiency. Secondly, the number of errors has been reduced to nil in terms of reliability. Data accuracy is ensured through the automation of the process eliminating the potential “human” errors caused by manual check and manipulation. The inexistence of literature regarding this necessity is outstanding. Also the understanding of the current process was vital in order to meet the expectations of the new software tools developed. Nevertheless, the developed software tools still require manual operation in some calculations and tolerance testing, where errors may still occur. As a future work this can also be automated and integrated into the tools.

The findings have implications for future studies that can further contribute to knowledge. Due to the requirement for the control that stores the data to have a high degree of reliability, automation was identified as the main contribution to data quality in terms of errors minimisation and time consuming. Decision-making is an additional reason why data quality becomes important as well as customer impact. Therefore the project aimed to develop an automated process and software tool that help improve and control internal banking processes as well as managing the data quality assessment and monitoring the retrieval of the pending position status of loan processes in banks. The work focused on analysing and understanding existing processes in Santander and identifying improvements for them. One of the outcomes of the project was to create a data flowchart of both applications developed using reverse engineering methods that can be used as a base for future improvements to be incorporated into the applications. This research work provided a method to implement the automation of lending controls via software developed tool that can be used to control the correctness of the data stored in the core database. The work also contributed to the development of an automatic retrieval of the pending position status of loans, enabled by a software tool. This tool can be used to track critical data movements within the core system applied to multiple applications.

Both applications have been successfully implemented within DQSS department in the case study bank and can also be incorporated by other departments and areas. These are generic tools that can be adapted by other banks and companies.

In the case that any additional studies are to be carried out, the author has identified several suggestions for further work. The positive results extracted from this thesis are valuable and can form a basis to undertake new challenges. New controls can be built in order to be more precise and improve the control model tool. The automation of the control process allows the business experts to be more proactive than reactive. With the awareness and knowledge acquired from this project, new ideas for control process and consequently an improvement in detection the potential errors on data can be innovated. Future research may also address the possible changes that could occur on the source reports used to get the information. If this were to happen, the code written for the tools would have to be modified. However this could be easily adapted due to the flexibility feature with which the tool was built. Thus, the model becomes more universally adaptable for any change within the source or output demand.

REFERENCES

[1] Abrahamsson, P., Warsta, J., Siponen, M. T. and Ronkainen, J. (2003), "New directions on agile methods: a comparative analysis ", *Conference on Software Engineering*, 2003. Proceedings. 25th International pp. 244.

[2] Andersson, F., et al. (2001) “Credit risk optimization with Conditional Value-at-Risk criterion”, *Mathematical Programming*, vol. 89, no. 2, pp. 273–291.

[3] Ashbaugh-Skaife, H. et al, (2007) ”The discovery and reporting of internal control deficiencies prior to SOX-mandated audits” *Journal of Accounting and Economics* Vol. 44, pp. 166–192.

[4] Aubrey, C.A. II (1985), *Quality Management in Financial Services,* Hitchcock Publishing Co., Wheaton, Illinois.

[5] Banco Santander, (2012), *Anual Report 2012*, Banco Santander website, http://www.santander.com/csgs/Satellite/CFWCSancomQP01/es_ES/Corporativo/Relacion-con-Inversores/ Informe-anual.html

[6] Banco Santander, (2012), *Risk Report 2012,* Banco Santander website, http://www.santander.com/csgs/Satellite/CFWCSancomQP01/es_ES/Corporativo/Relacion-con-Inversores/ Informe-anual.html

[7] Baskarada, S., (2011) “How Spreadsheet applications affect information quality”, *Journal of Computer Information Systems,* Vol. Spring 2011, pp. 77.84.

[8] Beck, K. (1999), "Embracing change with extreme programming", *Computer,* vol. 32, no. 10, pp. 70-77.

[9] Beck, K. et al., (2001) "Manifesto for Agile Software Development", *Agile Allianc,.* Retrieved 14 June 2010.

[10] Campbell, A., (2007) “Bank insolvency and the problem of nonperforming loans”, *Journal of Banking Regulation* vol. 9, no. 1, pp. 25–45.

[11] Financial Risk, *BusinessDictionary.com*, WebFinance, Inc. http http://www.businessdictionary.com/definition/financial-risk.html

(accessed: June, 2013).

[12] Fisher, C.W., (2001) “Criticality of data quality as exemplified in two disasters”, *Information& Management*, Vol. 39, pp. 109-116.

[13] Grant, G.H., Miller, K.C., Alali, F.,(2008) “The effect of IT”, *Managerial Auditing Journal* Vol. 23, no. 8, pp. 803-82.

[14] Gorschek, T., Davis, A.M., (2008) “Requirements engineering” *In search of the dependent variables Information and Software Technology*, Vol. 50, no. 1-2, pp. 67-75.

[15] Gosh, M., (2011) *Banking Settlement Pricing Analysis and Automation*, (MSc thesis), Cranfield University, Cranfield, Bedford.

[16] Kremin, L.A., (1986), "Improving the Quality of Banking Services", National Productivity Review, Vol. 5, pp. 32.

[17] Koronios A., (2006) “Validating the data quality framework in engineering asset management.” In: *Proceedings of the 17th Australasian conference on information systems*, Adelaide, South Australia, December pp. 6–8.

[18] Lambert, R., C. Leuz, and R. Verrecchia, (2007),“Accounting Information, Disclosure, and the Cost of Capital”, *Journal of Accounting Research* Vol. 45, pp. 385-420.

[19] Leone, M. (2006), “Section 404, year three: ready for it?”, CFO.com, May 10.

[20] Loan, *Investopedia.com*, Investopedia US, A Division of ValueClick, Inc. http://www .investopedia.com/terms/l/loan.asp (accessed: June, 2013).

[21] Mahmoud M. Yasin, Ronald F. Green and Marwan Wafa, (1991) “Statistical Quality Control in Retail Banking”, *International Journal of Bank Marketing*, Vol. 9, no. 2, 1991, pp. 12-16.

[22] Min, H., Xin-tian, Z., (2006) “Control Model of Bank Credit Risks Based on Commercial Age of Enterprise” *Management Science and Engineering, ICMSE '06. 2006 International Conference* 5-7 Oct. 2006, pp.1546-1551.

[23] Mortgage, *Investopedia.com*, Investopedia US, A Division of ValueClick, Inc. http://www .investopedia.com/terms/m/mortgage.asp (accessed: June, 2013).

[24] Murphy, G.D., (2009), “Improving the quality of manually acquired data: Applying the theory of planned behaviour to data quality”, *Reliability Engineering and System Safety*, Vol. 94, pp. 1881–1886.

[25] Nolan, R.,(1994),“Note on estimating the value of the IT asset”, *Harvard Business School* Note Vol. 9, pp. 195-197, Cambridge, MA.

[26] Orr K. (1998) “Data quality and systems theory.” *Communications of the ACM*, Vol. 41, no. 2 pp.66–71.

[27] Peltier, J.W., et al, (2013) ”Organizational Learning and CRM Success: A Model for Linking Organizational Practices, Customer Data Quality, and Performance”, *Journal of Interactive Marketing* Vol. 27, pp. 1–13.

[28] Pemberton, J.D., Robson, A.J., (2000), “Spreadsheets in business”, *Industrial Management & Data Systems* Vol. 100, no. 8, pp. 379-388.

[29] Public Company Accounting Oversight Board (PCAOB), (2004), *An Audit of Internal Control over Financial Reporting Performed in Conjunction with an Audit of Financial Statements*, Auditing Standard no. 2, Washington, D.C.: PCAOB.

[30] Risk, *Oxford Dictionaries Online*, http://www.oup.com/Oxford University Press, http://oxforddictionaries.com/definition/english/risk (accessed: June, 2013).

[31] Rosenberg, L. H., (1997), "Software Re-engineering", SATC-TR-95-1001, October 1996.

[32] Schmenner, R.W. (1990), *Production/Operations Management Concepts and Situations* (4th edition), Macmillan Publishing Company, New York.

[33] Stephanou, C., Mendoza, J.C., (2005) “Credit Risk Measurement Under Basel II: An Overview and Implementation Issues for Developing Countries” *World Bank, Washington, DC*. https://openknowledge.worldbank.org/handle/10986 /8557.

[34] Stoel, M.D., Muhanna, W.A., (2011) “IT internal control weaknesses and firm performance: An organizational liability lens”, *International Journal of Accounting Information Systems*, Vol. 12, pp. 280–304.

[35] Stults G., (2004) “An overview of Sarbanes–Oxley for the information security professionals” *White Paper, Bethesda, MD: The SANS (Sys Admin, Audit, Network, Security) Institute*.2004. (available at http://www.sans.org/reading_room /whitepapers/legal/1426.php).

[36] Unterkalmsteiner, M., (2012)“Evaluation and Measurement of Software Process Improvement— A Systematic Literature Review”, *Ieee Transactions On Software Engineering*, Vol. 38, no. 2, pp. 398-424.

[37] Benefits of Agile Software Development, *VERSIONONE,* http://www.versionone.com/Agile101/Agile-Software-Development-Benefits/ (accessed: June 2013)

[38] Wagner, S., Dittmar, L., (2006) “Unexpected Benefits of Sarbanes-Oxley” *Harvard BusinessReview,* Vol. 84, no. 4, pp.133-40.

[39] Wang, R.Y., Strong, D., (1996)“Beyond accuracy: what data quality means to data consumers”, *Journal of Management Information Systems,* Vol. 12, no. 4, pp. 5-34.