A method to obtain sustained data quality at Distimo

Master Thesis

M. Niblett
m.niblett@alumnus.utwente.nl
June 2012

DISTIMO UNIVERSITY OF TWENTE.

A method to obtain sustained data quality at Distimo

Master Thesis

M. Niblett

June 2012

Department of Management
and Governance
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Graduation committee:

University of Twente:
Dr. ir. Maurice van Keulen
Dr. Roland M. Müller

Distimo:
Remco van den Elzen MSc
Ruben Heerdink MSc

Abstract

This thesis attempts to answer the following research question: *"how can we determine and improve data quality?"*.

A method is proposed to systematically analyse the demands and current state of data quality within an organisation. The mission statement and information systems architecture are used to characterise the organisation. A list of data quality characteristics based on literature is used to express the organisation in terms of data quality. Metrics are established to quantify the data quality characteristics. A risk analysis determines which are the most important areas to improve upon. After improvement, the metrics can be used to evaluate the success of the improvements.

Distimo is an innovative application store analytics company aiming to solve the challenges created by a widely fragmented application store marketplace filled with equally fragmented information and statistics. As Distimo's products are very data driven, data quality is very important. The method will be applied to Distimo as a case study.

The proposed method provides a way to determine the current state of data quality, and to determine what to improve, and how to evaluate if the improvements provide the desired outcome. The case study of Distimo resulted in an in-depth analysis of Distimo, which in turn yielded a number of data quality improvements that at this very moment are in production and have improved data quality.

Because of the generic nature of input data, the proposed method is applicable to any organisation looking to improve data quality. The iterative improvement process allow for fine grained control of changes to organisational processes and systems.

Preface

This document is the result of a Master Business Information Technology at the University of Twente. It has proven to be quite a challenge to finish it, as the initial work started over two years ago. It is finally done, the end of an era!

The ground work for the topic of research was laid by the founding of Distimo in May of the year 2009. As the start up was short staffed in the attempt to quickly launch a working prototype application, I was asked to step up to temporarily lend a hand, eventually temporary would become permanent. A few months later all but one of my courses were completed and I was in search of a suitable research assignment. As I now was employed by Distimo, and Distimo was closely linked to the University, having Mr Maurice van Keulen and Mr Roland Müller as supervisors, it was a very suitable environment for a Masters assignment. Both Maurice and Roland agreed to be my supervisors. Representing Distimo, Remco van den Elzen and Ruben Heerdink accepted a seat in the graduation committee.

Despite having started with a great deal of enthusiasm resulting from an inspiring environment as well as a challenging assignment, keeping up motivation and thus progress whilst also working a daytime job turned out to be quite a personal challenge. Lucky for me, my supervisors have shown a lot of patience.

Over time some changes occurred, Roland moved to Berlin, Distimo moved to the centre of Utrecht, I moved to Utrecht, but focus remained on completing my Masters.

After overcoming the final hurdle, the last course for which I had to write a paper, which took significantly longer than anticipated, the time has finally come for me to conclude my Master.

Of course this would not have been possible without the help, knowledge and support of a lot of people. I would like to use this opportunity to thank everyone who in some way has helped me to achieve this. But I would like to name a few people specifically:

My supervisors, many thanks for reading all of my writings, time and time again! Maurice and Roland, you have been patient, supportive and a great inspiration for new ideas. Remco and Ruben, you have been a great help during this long process!

Many thanks to all my friends, who in all these years have never failed to keep asking me if I was almost finished, which kept me motivated until the very end.

Wendy, my sister, who just beat me by a few months, but that was definitely the signal to start my final end sprint.

Many thanks to my parents, for their never ending belief and unconditional support over the years!

Finally, last but not least Manon; thank you very much for bearing with me over the past few years, enduring all the times I was frustrated, fed up, depressed, unreasonable, angry, annoyed, etc while writing my thesis. I could not have finished it without your support!

Contents

List of Figures

Chapter 1

Introduction

This chapter will give an overview of the background and structure of this thesis as well as a brief introduction into a number of key concepts.

1.1 Distimo

Distimo [1] is an innovative application store analytics company aiming to solve the challenges created by a widely fragmented application store marketplace filled with equally fragmented information and statistics.

Distimo Report produces custom in-depth analytical reports for parties interested in the mobile application ecosystem, giving valuable insight into important trends happening within and across application stores. For developers, Distimo offers a free analytics tool, Distimo Monitor, to monitor their applications and competitors' applications in all application stores.

Distimo was founded in May 2009 by four University of Twente alumni and is a privately held company based in The Netherlands.

1.2 Research questions

This paragraph describes the challenges at Distimo and how they lead to a research question.

1.2.1 Challenges

Distimo handles a lot of raw data. One of the challenges Distimo faces, is aggregating and presenting this data to their customers. How to combine revenue numbers in different currencies or how to combine download statistics gathered over different time intervals, to name a few. Because customers rely on this data being correct, it is very important for Distimo to be as sure as possible about

for instance the correctness, the reliability, etc of this data. Furthermore it is important to be very clear about the meaning of the data that is presented.

1.2.2 Research question

The main research question may be formulated as follows:

> *How can we determine and improve data quality?*

However, to make it easier to carry out the research to come to a suitable answer, it may be split into a number of sub-questions:

- What is data quality?
- How can we measure data quality?
- How can we systematically determine and improve data quality?
- Can we use Distimo as a case study?

In the following chapters, these questions will be answered.

1.3 Approach

This thesis attempts to classify certain characteristics of data quality and create a method to systematically assess these properties in order to measure data quality. First a literary background for the theory involved in data quality research will be discussed. Next, a method to determine and improve data quality will be proposed, followed by a more in-depth analysis of Distimo. The proposed method will then be applied to Distimo as a case study. After the results of the Distimo case study, conclusions about the case study as well as the method in general will be discussed.

1.4 Significance

The results and conclusions of this research project may be used by Distimo to influence the quality of the data as produced by the Distimo Monitor application. As stated earlier, the quality of the data is very important for Distimo and its customers. Furthermore the design of the method is aimed to be generic, i.e. it will be applicable to other similar cases where data quality is concerned.

Chapter 2

Literature study

This chapter will give a literature overview of the key concepts regarding this research project. For any research project it is important to have a starting point, and a study of relevant literature will provide this. Using the standard scientific literature search engines, an overview of the state of the art of various subjects was constructed. The articles found were used as a starting point, after which both forward and backward searches were done to establish the quality of the found articles. The following paragraphs will provide a more thorough detail of the findings.

2.1 Data quality

As data quality is a very important part of this research project, this section will explore the current status of data quality in terms of literature. Data quality and knowledge about the current data quality is very important for organisations in the current economy. Many organisations rely blindly on the data on their systems although the quality, or rather the lack of quality, of the data can have detrimental effects on the organisation.

2.1.1 Terminology

Before starting the discussion of the various views from the literature, it is important to note that often a distinction between data and information is made [2]. The distinction refers to data being raw and factual, whilst information is a subset, or transformed form of the original data, adapted to a meaningful form for a specific purpose. This difference however, often depends on the point of view taken. In this thesis the term data will be used, but may refer to either terms.

2.1.2 Data quality Problems

Poor data quality is not uncommon in many companies [3]. A surprisingly large number of errors is found in many databases. Many organisations are not aware of problems with data quality, or the effects poor data quality may have on the business as a whole [4]. This may be due to organisations overestimating the quality of their data, but also to unawareness of the presence of poor data. There is a big number of causes of poor data quality, to name a few:

- **Lack of input validation** - data entered into a system, without being checked. For example, a phone number should consist of numbers, and perhaps some spaced and a + sign. Using alphabetic characters would result in invalid data. Failing to validate input can lead to useless data. With more and more systems being directly coupled to the web, allowing users to directly input data, it is even more important to have good validation rules for input data.
- **Incorrect data** - even validated data does not necessarily have to be correct. Take the phone number for example, even though a sequence of numbers is entered, this does not mean that the number entered is the phone number of the user. It may be a fake number, or the user may have made a typing error. These kinds of errors are very hard to catch, although some integrity constraints may help. For instance, the net number (first part of a phone number in the Netherlands) may be verified using a city of residence, or perhaps a postal code.
- **Format or syntax mismatch** - when combining data from multiple sources, it is important to be sure they use the same format. For instance an American database may use the MM/DD/YYYY format, whilst a Dutch database may use the DD/MM/YYYY format. In most instances it is possible to detect this, but not always (if the day of the month is less than 13 for instance). Of course this is a fairly trivial example, which can be solved by using appropriate data types and/or localisation packages, but the principle also holds for less trivial cases.
- **Changes in interfaces** - if a systems gathers data from various sources, it depends on a certain interface being available at these sources. If an upstream source decides to change that interface, problems regarding the quality of the obtained data may occur. In the best situation a change in interface is detected and no more data is imported, and an administrator is notified of the problem. In the worst case, the change is not detected because for instance the change is only in the semantics of the data (i.e. the upstream source switches from USD to EUR as their base currency). This would cause incompatible data to be entered into the database, which would only manifest itself during retrieval, or worse.
- **Lack of integrity checks** - when dealing with very large databases, a way to improve performance is to temporarily disable certain integrity checks. When performing an insert, the database has to check if all constraints are met, update all indices, etc. If a very large number of inserts has to be performed, it may be an advantage to disable or postpone certain checks. However, most checks are there for a reason. Failing to re-enable them

introduces a huge risk of invalid data i.e. inconsistencies in the database.

- **Poor systems design** - sometimes when developers face strict deadlines, shortcuts are taken when designing or implementing a new system. This may result in broken data. For instance, if not enough attention is paid to using proper data types, character encoding, transactional properties, scalability etc, data integrity may be violated.
- **Conversion errors** - quite often data is obtained from various sources before it is stored in a local database. This is a result of ETL (extract, transform, load). It is very likely that the source data needs to be transformed to an appropriate form before it can be stored. Failure to perform this conversion adequately results in broken data. This may occur if insufficient attention is paid to the source data. Alternatively this may occur when a local database is converted, for instance when software is upgraded and the corresponding database as well. Proper attention should be paid to conversion to make sure no data is lost or corrupted across the conversion.
- **Unclear definitions or rules** - when integrating systems, even within an organisation, it is important to be aware of differences in definitions or rules. One department may measure profit growth in absolute numbers, while another department uses a relative percentage. Both departments refer to the same term, but use a different definition. This problem becomes even bigger when taking inter-organisational integration into account. This may occur when integrating systems across the value chain, or after mergers, acquisitions and the like.
- **Changing dimensions** - the dimensions of data may change over time, thus changing the semantics of these dimensions. For instance a supplier of an organisation may also become a buyer. This would result in the count of relations not being equal to the sum of buyers and suppliers, since this particular relation appears in both.

This list is by no means exhaustive, so it is quite clear that there can be a lot of ways data quality becomes poor. Since a lot of organisations rely (sometimes blindly) on their data being of high quality, it is an important issue to keep track of this data quality.

2.1.3 Characteristics

Although data quality may seem a trivial concept, a lot has been written about the subject in literature. Viewing data quality from various angles can result in subtle differences. However, comparing different authors, a lot of overlap in findings regarding data quality characteristics becomes evident. In this section a brief introduction into the relevant literature is given. In table 2.3 on page 9 the results are aggregated in a table as a list compiled of data quality properties from the various papers.

Wang et al [3] carried out a study into data quality properties amongst what they call data consumers. Data consumers are the actual users of data and may have their own perspective on what the important aspects of data quality are. A

Category	Dimensions
Accuracy of data	• Believability • Accuracy • Objectivity • Reputation
Relevancy of data	• Value-added • Completeness • Relevancy • Timeliness • Appropriate amount of data
Representation of data	• Interpretability • Ease of understanding • Representational consistency • Concise representation
Accessibility of data	• Accessibility • Access security

Table 2.1: Data quality categories by Wang [3]

two phase research method was used. In the first phase, the data consumers were asked to list all data quality properties they regard as important. The results are a list of 178 different data quality characteristics. In the second phase the items in this list are ranked. The most important ones are categorised into four categories. In table 2.1 the results are shown.

In their paper, Huh et al [5] present a list of four dimensions that define data quality:

1. **Accuracy** is measured by comparing data to an identified source. As data usually represents an abstract representation of reality, accuracy pertains to the level of similarity between the data and the reality it is aimed to represent (this only involves the relevant characteristics of reality that are specific to this representation).

2. **Completeness** refers to whether all relevant records are present in the data. It also concerns the completeness of individual records for instance the presence of blank fields within a record.

3. **Consistency** between two data sets is determined by checking they do not conflict with each other. There are basically two kinds of conflicts: logical and formatting. A logical conflict means that data from two sets have different logical implications. For instance: field C should be the sum of fields A and B. If this holds in data set 1, but not in data set 2, data set 1 and data set 2 are said to be inconsistent. A formatting conflict

occurs when the format of the data (for instance different data types to represent the same data) between two data sets is in conflict.

4. **Currency** determines if data is up to date. Depending on the application, it may be important for data to be highly current. The measurement may differ per case. For instance, in real-time applications data may be current if it is less than 1 seconds old, while in less real-time applications, data may still be current if it is a day old.

An important factor in data quality is the input. An analogy is made between data and a lake. In order for the lake to have clean water, the water flowing into the lake must be clean. That way it is easier to keep the lake clean, in contrast to continuously cleaning the lake. The same may be said for data, if the quality of the input data is high, it is not necessary to sanitise it later on.

Pipino et al [6] look for a way to assess a company's data quality. Experience suggests a "one size fits all" set of metrics is not a solution, rather a method is proposed that combines subjective and objective assessments of data quality. The data quality is defined using the dimensions in table 2.2.

The data quality characteristics as presented by the various papers are aggregated into a single list in table 2.3.

Dimension	**Description**
Accessibility	The extent to which data is available, or easily and quickly retrievable.
Appropriate amount of data	The extent to which the volume of data is appropriate for the task at hand.
Believability	The extent to which data is regarded as true and credible.
Completeness	The extent to which data is not missing and is of sufficient breadth and depth for the task at hand.
Concise representation	The extent to which data is compactly represented.
Consistent representation	The extent to which data is presented in the same format.
Ease of manipulation	The extent to which data is easy to manipulated.
Free-of-error	The extent to which data is correct and reliable.
Interpretability	The extent to which data is in appropriate languages, symbols and units, and the definitions are clear.
Objectivity	The extent to which data is unbiased, unprejudiced and impartial.
Relevancy	The extent to which data applicable and helpful for the task at hand.
Reputation	The extent to which data is highly regarded in terms of its source or content.
Security	the extent to which access to data is restricted appropriately to maintain its security.
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand.
Understandability	The extent to which data is easily comprehended.
Value-added	The extent to which data is beneficial and provides advantages from its use.

Table 2.2: List of quality dimensions by Pipino [6]

Characteristic	Description	References
Accessibility	Are the data accessible and easily retrievable?	[3], [6], [7]
Accuracy	How accurate are the data, i.e. is there a difference between the data and a verified source, and how big is that difference?	[5], [7], [3]
Appropriate amount of data	Is the amount of data proportional to its application?	[3], [6], [7]
Believability	The extent to which data are regarded as true and credible.	[6], [7]
Completeness	Are all data available to interpret it (context), does the data set contain all data (depth and breadth) for the specific application?	[3], [5], [6], [7]
Consistency	Do all data meet format requirements and are they without conflicts?	[5]
Currency/ timeliness	What is the age of the data set, is it current/new enough for its purpose?	[3], [5], [6], [7]
Free-of-error/ reliability	To what is extent is the data correct and reliable?	[6]
Interpretability	Are data represented using appropriate languages, symbols and units, and clearly defined?	[3], [6], [7]
Objectivity	To what extent are the data unbiased, unprejudiced and impartial?	[6], [7]
Relevancy	Are the data applicable (i.e. suitable for the intended application)?	[6], [7]
Representation	Are the data represented consistently? is it comparable/compatible with previous data?	[3], [6], [7]
Reputation	What is the reputation of the data source, is it reliable, well-known, etc?	[3], [6], [7]
Security	Data access is restricted to authorised persons only.	[3], [7]
Understandability	Is it easy for the intended audience to comprehend the data?	[6], [7]
Value-added	Do the data provide added value?	[3], [7]

Table 2.3: Aggregated data quality characteristics

Chapter 3

Method

In this chapter a method will be proposed which can be used to analyse the current state of the art of data quality within an organisation. The method proposes a systematic approach to analyse an organisation and identify what data quality properties are important based on the characteristics of the organisation. Subsequently these characteristics will be improved upon iteratively. First a brief overview of the method will be given, after which the various components will be discussed in greater detail.

3.1 Overview

Because an image is worth more than a thousand words, a graphical overview is given in figure 3.1 on the next page. As can be seen the method globally consists of two phases:

1. **Analysis** - during this phase the existing organisation and its processes are characterised in terms of data quality. Furthermore, goals in terms of that data quality are set.
2. **Measurement** - during this phase the results of the analysis phase are used to measure the actual data quality within the organisation. An improvement cycle is incorporated to improve the data quality iteratively in order to meet the proposed goals.

The philosophy behind the method is that the goals of the organisation and the processes of the organisation (can be both internal and/or external) can be used to determine the important data quality aspects. In other words, the organisational profile created in the analysis profile defines the organisation in terms of data quality properties. From this vantage point the method can be seen as characterising the organisation in terms of data quality and benchmark these using appropriate metrics. The outcome of that analysis can then be used to initiate a process to try to improve on certain areas or processes in the organisation where data quality is not on par with the desired levels. This is achieved with the improvement cycles.

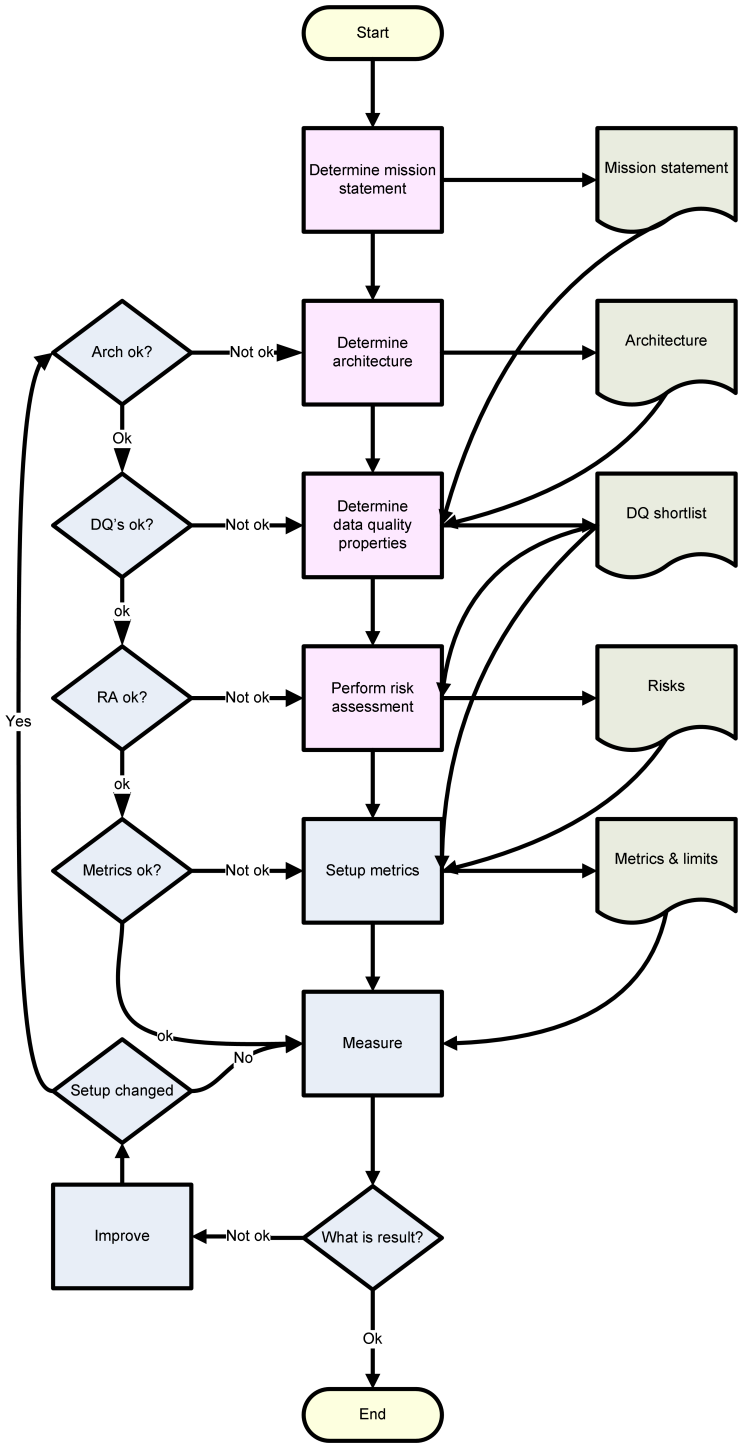

Figure 3.1: Method process flow

3.2 Analysis

In the analysis phase, a profile of the organisation is built. The profile attempts to describe the organisation in terms of its goals, its processes, and finally in terms of data quality aspects related to the aforementioned properties. The assumption is that there is a relation between the reason of existence of an organisation and possible data quality properties.

3.2.1 Data quality properties

As discussed in the literature overview, it is possible to come up with huge numbers of data quality properties, especially when looking at very specific applications. However, the goal of this method, is to provide a generic guideline. Therefore it is neither in the scope of this method, nor possible to provide the reader with a complete and exhaustive list of all data quality properties possible. Instead, a list of the most important ones will be provided. The proposed list can be found in table 2.3 on page 9.

3.2.2 Profile

A profile of the organisation is created in order to create a context to determine which data quality aspects are relevant. The profile basically consists of the following parts: a mission statement, a description of one or more processes and a risk analysis. The mission statement is used to determine high level quality characteristics. The processes are used to determine the various information flows and transformations i.e. they are used to find out possible data quality issues.

Mission statement

Generally speaking, a mission statement of any organisation is a brief description of the reason of being for that organisation [8]. A mission statement should contain the purpose or goal of the organisation. This can be anything from just generating revenue, to making a profit, or perhaps providing a particular service. Furthermore the most important stakeholders of the organisation and their relation to the organisation should be mentioned. These can be anything from customers, suppliers, clients, etc. Finally the most important products or services provided will be listed.

This means that a mission statement is a perfectly suitable starting point to assess the data quality requirements for an organisation. The goal of an organisation, its stakeholders and the products offered may all place specific demands on the data quality. Using the mission statement, it is possible to determine high level data quality properties that are important for that specific organisation. The mission statement creates a context.

Architecture

Whilst the goals and targets of an organisation as defined in the mission statement may help to classify the most relevant and important data quality characteristics for an organisation, this is still only part of the process. Once the most important data quality aspects have been defined, it is time to have a less abstract view at the organisation. Looking at organisational processes and architecture enables identifying possible points where data quality issues may be present.

To be able to pinpoint potential locations within an organisation where data quality problems (defined in terms of the earlier chose data quality properties) may arise, the information systems architecture of the organisation, or a specific process within can be used. The architecture is basically a breakdown into various components that make up the organisation [9]. A suitable technique should be applied that allows for the architecture to be modelled in such a way that all potential points where data quality issues may arise can be expressed. This means that the architecture should contain at least the following aspects:

1. components
2. relations
3. processes
4. information flows

Risks

Assess risks in terms of data quality. Risk management basically revolves around the following questions [10]:

1. What are potential threats (risks)
2. What is the likelihood that they will occur
3. What will be the impact of their consequences

The risk analysis will help in determining which of the found data quality characteristics are the most important to improve. For instance if a certain risk has a small chance of occurring or only minor consequences, it may have a lower priority.

The data quality characteristics found in the previous stages (mission statement and architecture) can be be combined with the risk assessment to determine which of of the characteristics pose the biggest risks and should therefore be a potential candidate for an improvement cycle.

3.3 Measurement

After defining the the data quality properties and the organisation itself, it is time to make the data quality properties a little more concrete. A metric is

basically a measure of a property [11]. Using appropriate metrics, it is possible to measure data quality properties. This means that using metrics it is possible to measure the data quality of the organisation.

After setting up metrics to measure the data quality characteristics, the measurements may be used to determine if the data quality is satisfactory.

Special care should be taken when interpreting results as provided by the employed metrics. Take into account the possibility of false positives (metrics signalling problems, which are not real problems at all) and false negatives (metrics signalling everything is ok, when in fact this is not the case).

3.3.1 Improvement

The results from the measurements should give a clear and objective image of the current state of data quality. The results can then be used as input for an improvement cycle in order to improve certain data quality aspects. After changes have been made to processes and/or information systems, the existing metrics can be reused to see how the changes affect these.

The improvement cycle basically has two routes that may be followed. The "short" route performs an improvement process and afterwards returns to the measurement process to directly evaluate the results of the improvements.

The "long" route also performs an improvement process, but also re-evaluates the data obtained in previous steps. This means that architecture, data quality characteristics, risk assessment and metrics are evaluated to see if they are still up to date and relevant. If this is the case, then the cycle returns to the measurement process. If this is not the case, the specific document is updated, and the method follows the standard flow from there on.

The improvement process may be used for a number of reasons:

1. An actual improvement to the process which is being examined. This could for instance be changes to the architecture.
2. An improvement or addition to monitoring systems. It may be the case that desired metric(s) (resulting from the analysis of data quality characteristics and the risk assessment) are not available or not of sufficient quality.

3.4 Using the method

The method basically consists of two parts. In the first part an analysis of the organisation is done in order to construct a profile of the organisation in terms of data quality. In principle this is done only once. In the second part, a so called feedback loop evaluates the current state of the data quality and benchmarks it using metrics. Using the results, changes can be made in order to improve the data quality, furthermore the metrics may need some adjustments. Once that is done, it starts over i.e. run benchmarks and see what the performance is.

A short overview of the steps to follow in order to use the method:

1. **Mission statement** - Determine mission statement to characterise the organisation in terms of data quality characteristics.
2. **Architecture** - The information systems architecture is used to determine where to look for data quality issues.
3. **Data quality properties** - Based on the previous steps, choose a limited number of relevant data quality characteristics.
4. **Risk assessment** - Perform a risk assessment based on the short list of data quality aspects. This will help to determine the most important data quality aspects.
5. **Metrics** - Determine metrics to measure the previously selected data quality characteristics.
6. **Measurement** - Carry out measurement of the metrics to see if the data quality is on par.
7. **Improvement** - Make adjustment to the information systems in order to improve weak areas with respect to data quality and reevaluate the earlier obtained results.

The method consists of two sections; a static section, this is the first part where the organisation is categorised, and a dynamic section. In the dynamic section performance is measured and adjustments can be made to compensate if performance is not up to standard. As can be seen from figure 3.1 on page 12, after completing the method, it is possible to reenter the method through an improvement cycle via various routes. However the first step to be carried out if the results are not satisfactory is to try and improve the processes.

1. **Adjust architecture** - In some cases the measurements uncover a problem that is a direct result of certain architectural choices. If the motivations for these choices are not valid (or sometimes non existent), changes have to be made at architectural level. Obviously, the greater part of the method has to be done again to get usable results.
2. **Data quality properties** - If the wrong data quality properties were selected, they need to be evaluated and changed where appropriate. Following this appropriate metrics and boundaries have to be set after which the measurements can be rerun.
3. **Risk assessment** - If any of the input parameters for the risk assessment have changed, risk assessment has to be reevaluated in order to be reliable.
4. **Setup metrics** - It is perfectly possible that after carrying out the measurement, it turns out that wrong metrics were chosen, or perhaps incorrect ranges were selected. In this case, the course of action is to correct the metrics, and rerun the measurements.
5. **Setup changed** - If the setup has not changed, i.e. if the documents created from the previous phases like “determine mission statement” and “determine architecture” have not changed as a result of improvements made, it will most probably suffice to carry out the measurements again

to see if data quality is affected by the changed made, and what these effects entail.

Chapter 4

Distimo in general

As Distimo is used as a case study, this chapter will give a more in-depth view of what Distimo does and its history.

4.1 Appstore Analytics

As stated on their website [1] *"Distimo is an innovative app store analytics company built to solve the challenges created by a widely fragmented app store marketplace filled with equally fragmented information and statistics"*.

But what does this mean exactly? It all starts with smart phones and the concept of applications and so called application stores, or in the lingo: *app stores*.

4.1.1 Smart phone

For a long time since the introduction of the first mobile phones, more and more functionality has been added to them over time. At first mobile phones were merely used to call other people. Then services like SMS were added. As technology development progressed, more and more possibilities arose (not only processing power memory, storage, interface of he phone itself, but also in terms of connectivity, for instance large scale broadband networks like GPRS/UMTS emerged), giving birth to the smart phone. Although it lies outside of the scope of this thesis to give a formal definition of the concept of smart phones, a smart phone may be defined as a mobile phone that offers advanced computing capabilities and connectivity over a traditional mobile phone. In other words, a smart phone could be viewed as a small but mobile networked computer.

4.1.2 Applications

Given the availability of smart phones, a lot of potential new uses for what used to be just a phone emerge. Phones are not just used for calling, but can be

used for any number of different purposes like navigation, email, photography, etc. Moreover, functionality of a phone is no longer bound to what a hardware manufacturer enables at assemble time, but can be expanded by installing additional applications, or apps. Being basically a small computer, installing additional apps on your smart phone is quite similar to installing software on a regular personal computer.

4.1.3 App store

Smart phones enable their users to install various applications on their phone, enabling them to add functionality to their device. Examples of applications are basic applications like email clients, calendaring apps, messengers, web browsers. But also more specific/ complex applications for example navigation apps (enabled by the presence of GPS chips in modern phones), games, media players, etc.

Traditionally, the process of installing an application on to your computer involves somehow obtaining the application, for instance by downloading it, or by buying a CD or DVD. The concept of an application store is to have a central place containing all available applications. From this central place, also called an application store, or app store for short, it is possible to search for, obtain and install applications. One of the most well known app stores is the Apple App Store, which provides applications for Apple's iPhone and the Apple iPad.

Because of the vast amount of apps available, app store providers have sought to find ways to structure the apps in such a way that users can find the right app for their particular purpose.

For instance, most app stores use some kind of categorising system, where apps are assigned to one or more relevant categories (usually based on the functionality of the app). That way, if a user is looking for an app for a specific goal, the search field is narrowed down to the relevant category, which in turn means less applications to search through and probably a higher percentage of relevant search results.

Although the Apple App Store is the best known (and biggest in terms of number of available applications) app store, Apple did not invent it. Companies like for instance GetJar [12] were around long before Apple even had a mobile phone for sale.

Most app stores offer an end user the possibility to search for applications. A user can then select an application from a list to download and install. Some applications are for free, meaning that downloading, installing and using them is free of charge. Other application may cost a small fee, although usually it is possible to try an application for 24 hours, and get a refund if the application is not satisfactory. A third form of application, is where the actual application is free, but the content of the application is not free. This means that a user can use the application for free, but has to pay for in-application content.

Other functionality of app stores often include the possibility to rate applications. Users can give their opinion on the usefulness, the quality, performance, etc. of an application. This feedback can be very useful to other potential users

of an application, because they can use it when searching for a particular application. Furthermore, the feedback may also prove useful the developer of the application for improvement purposes.

So what are the properties of an app store? A distinction between different actors involved has to be made:

- **Developers/publishers** - The parties responsible for the applications that are available from any app store are of course the developers. An app store provides them with the means to easily reach their potential end-users. It provides them with the infrastructure to distribute the application, in some cases monetising and, depending on the particular app store also with statistics regarding the usage of a particular application.

  There are some differences between the various app stores regarding the functionality offered to the developers. For instance, the Apple app store does not allow all applications. All applications are screened first, and after they are approved, they may be added to the Appstore. Apple has a very strict policy regarding the rules an app has to conform to in order to make it into the Appstore. Other app stores have different acceptance policies.

  Most app stores have a price model, where they collect a certain percentage of the revenue generated by paid apps. For example Apple takes about 30% of the revenue generated by paid apps.

- **End users** - The owners of smart phones are the people who end up using the applications from any app store. Usually the phone contains some kind of application which enables a user to browse the available apps in an app store. Using for instance charts of popular apps, or search functions a subset of all available apps may be obtained. After selecting an app, it may be downloaded and installed. Usually the app store takes care of this. In case of a paid application, payment has to be taken care of before downloading can commence.

- **Distributor** - The distributor is responsible for providing an infrastructure to facilitate the distribution of the apps. This means it provides an environment to developers to upload their applications, usually some kind monetising scheme and an interface to monitor the performance of uploaded apps.

  There are basically three kinds of distributors; the hardware and/or operating system vendors that run an app store for their own platform, network operators who have an app store targeted specifically at their customers and the so-called independent vendors. The first category obviously only has apps supported by its own platform in the application store. The second group may have apps for multiple platforms, at least the platforms they sell. The last group may support all platforms. In table 4.1 on the following page an overview of some of the stores is given.

Vendor	App store	OS	# apps
Apple	Appstore	iOS	590.000
Apple	Macstore	OSX	10.000
Google	Play	Android	400.000
Nokia	Ovi Store	Symbian OS	116.000
Microsoft	Windows Marketplace 7	Windows Phone 7	100.000
RIM	Appworld	Blackberry OS / QNX	58.000

Table 4.1: Short list of app stores (as of June 2012) [1]

4.1.4 Statistics

The mobile app market is a not very transparent market. Because of the many parties involved, it is hard to get a clear view of what the trends are. What are the popular applications, which app stores are popular, which countries are big, and which are not, etc.

Of course some statistics are (publicly) available. Many of the app store applications on the mobile devices allow for browsing by charts so a user can see which applications are currently popular. Most app stores have these charts available either through their app store application and in some cases also through a website. However, how is popularity measured? It could be based on the total number of downloads, or number of downloads over the past week, of perhaps based on user ratings, etc.

Apple for instance enables its users to use iTunes to install applications on their phones. This means the user can use iTunes to select and install applications from the App Store. This means that applications installed on an iPhone or iPad are not always installed directly from the phone, and as such are less easy to track.

Other interesting information about app stores is of course the number of downloads. While rankings give a hierarchy between the apps within a category, and often within a country, this does not allow comparison between categories, countries or even app stores.

Of course most app stores keep track of the downloads of applications. However, these figures are usually not available publicly. Usually the developer has to log in to some kind of back-end system which provides the performance statistics of his application(s).

4.2 Distimo products

Distimo basically has two core products:

1. **Report** - Distimo provides reports containing in-detail analysis of the app store market. Each month a report on the US market is released for free downloading. For more specific and tailored reports Distimo provides a customised solution.

2. **Monitor** - Monitor is a portal aimed at developers, providing them with analytical tools regarding the performance of their apps.

4.2.1 Application Statistics

In order to be able to have anything to report on the mobile application market, Distimo obviously needs data regarding the app store market. The various application stores provide statistics on the usage i.e. applications, downloads, purchases, rankings, pricing, etc. of their application store. These statistics form the core of the business model of Distimo.

The data gathering system is able to connect to a number of mobile application stores, this is achieved by emulating the behaviour of the client(s) for that particular platform. Since each of these application stores is different (different protocols, data structures, functionality etc.), a custom system has been developed to be able to communicate with them. Currently (June 2012) the following application stores are supported:

- Apple Appstore [13]
- Apple Macstore [14]
- Google Play [15]
- RIM Blackberry Appworld [16]
- Nokia OVI store [17]
- Microsoft Windows Marketplace 7 [18]
- GetJar [12]
- Samsung Apps [19]

Development is ongoing to support even more application stores. The goal is to have support for all major platforms.

The data basically consists of three types of data: downloads, rankings and pricing info. All of these data are of course centred around the actual applications and their publishers.

Downloads

All application stores keep track of the number of downloads of a certain application. Furthermore, additional information is stored, for instance the country where the application was downloaded, the date, possibly the type of the smart phone used. Download statistics are however not available publicly. However, a number of applications stores offer its customers (the application developers), to log in and view their personal download statistics.

Monitor uses log in credentials of registered developers to connect to the application stores and retrieve the download statistics for the applications registered by that developer. This data is referred to as private data, as it is only available to the owner of the application.

Rankings

As the application stores want to give some kind of overview of the popularity in their stores, charts are published. Applications are divided into various categories. Within these categories, the applications are ranked according to their popularity (may be based on the number of downloads, or any other suitable metric). These charts are available publicly and are continuously updated.

Revenue

Since Monitor needs to be able to convert the currencies over and forth, the exchange rates are stored daily. This way it is possible to convert historic data with a certain amount of accuracy, since exchange rates change over time.

A problem with the exchange rates is that the internal implementation of application stores is unknown, that is, it is not known which exchange rates are used by the application stores to convert the currencies of sold applications into the currency of the application developer. This means that small differences may occur when converting currencies.

4.2.2 Monitor

Distimo provides the Distimo Monitor application to subscribed developers. Distimo Monitor is an application that allows developers to view statistics of their applications in terms of downloads and rankings in the various application stores. Furthermore it is possible to compare the application to competing applications.

When a developer subscribes to the monitor service; his credentials are used by the back end systems to fetch the data concerning his apps. This means that public and private data are retrieved daily.

4.2.3 Report

Distimo provides aggregated transactional reports, tailored to the specific needs of their customers. These aggregated reports are based on the gathered public data. The eventually delivered reports mainly consist of Excel spreadsheets containing the actual data in tables/worksheets. Generation of these reports can take a lot of time and result in a lot of data.

4.2.4 System architecture

A back end system is used to accumulate and store statistical information regarding the mobile applications in a database. An in-house developed framework is used to gather data from various sources.

The framework gathers the data from each individual application store and converts it into a more uniform format. It is then stored into the data warehouse.

The data warehouse is currently implemented using a database on a MySQL [20] server.

A subset of the data in the date warehouse is pushed to the front end database server. This is used as data source by the web front end, which is Monitor.

The data stored in the data warehouse is a combination of publicly available data, and data gathered using authentication details of developers registered at for instance the Apple Appstore.

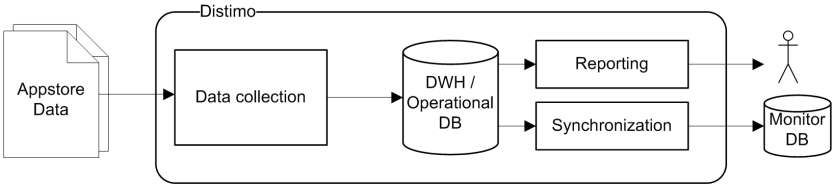

Figure 4.1: Top level architecture overview

In Figure 4.1 a high-level overview is given of the process of gathering data from external sources, storing the data in the data warehouse, and pushing data to the front end database server.

A web based front end allows developers to log in and generate graphs. Currently Monitor is implemented in a framework that follows the MVC design pattern and provides a querying interface, template loading, caching, etc.

4.3 Problem exploration

The topics described in this paragraph are part of the proposed research. During the research phase we will gain more insight in these subjects.

4.3.1 Data quality

The data as stored in the data warehouse (DWH) is considered factual data. However, the data that is eventually presented to the end user is a result of pushing data to the front end database server, various queries and possibly some aggregation (for example cumulation, interpolation, averaging, etc.) across one or more dimensions. Because data is the actual product of Distimo, it is important that all data that is presented is correct. Therefore it is important to maintain a high level of data quality. But what is data quality? It may be defined by various characteristics depending on the organisation concerned.

4.3.2 Data aggregation

Distimo gathers a lot of statistical data on application stores. To be able to present this data or subsets of this data in a useful way, it is necessary to

aggregate data. This presents a lot of issues, for instance how to combine various data, like download statistics collected over different time intervals, revenue statistics in different currencies. Conversion from one form to another may not be trivial, or at least not transparent to the end user.
Another hot topic is that the user needs to be aware of which data is shown (semantics). It needs to be transparent what data is shown, if it is factual, or aggregated in some way, to prevent wrong conclusions from being drawn. i.e. the application should have a clear description of what data is shown, and how the data was generated.

Chapter 5

Case Study

confidential

Chapter 6

Reflection

This chapter will discuss the results achieved at Distimo at trying to maintain and or improving data quality by employing the proposed method.

In the following paragraphs, the application of the steps of the proposed method as applied to the Distimo case study will be discussed.

6.1 Analysis

The first part of the analysis is determining the mission, which is quite easy for Distimo. Determining the architecture is a lot harder. Distimo is a relatively small and young company, which means that there are a lot of ever evolving dynamic processes. That makes it hard to make a clear snapshot at any given time. On the other hand, because of the relatively small number of employees, it is easy to tap into the wealth of knowledge present in all employees.

The mission and architecture have proven to be a very solid basis to identify the important data quality characteristics. Because the mission basically defines the goals of any organisation, combined with the architecture, it became clear what the most important data quality characteristics within Distimo are. And how they can be defined within the Distimo context i.e. what they mean specifically for Distimo.

The risk analysis is a handy tool, to identify the data quality characteristics which are most important, based on the impact they have on Distimo.

6.2 Measurement

In principle, the measurement phase should couple metrics to the identified data quality characteristics. An although this is perfectly possible for Distimo, it is not always very easy to determine proper metrics for all identified data quality characteristics (for example, how do you measure believability of a product).

Furthermore, even if it is possible to define metrics, chances are they are not in place, or are nearly impossible to implement.

6.3 Improvement

As the reporting process is one of the most important processes within Distimo, since it is the primary source of revenue, it was used for the improvement cycles. two iterations were carried out, to improve in multiple areas.

The following data quality characteristics were addressed in improvement cycle 1:

- Accessibility
- Amount of data
- Completeness
- Consistency

and in improvement cycle 2:

- Accessibility
- mount of data
- Completeness
- Consistency
- Currency

Although there seems to be very much overlap in the two lists given, the aspects of these data quality characteristics that were tackled in the two improvement cycles were quite different, as described in the previous chapter.

6.4 Overview

The steps from the proposed method are a good guide to a structured approach of describing and improving data quality through a systematic analysis of an organisation, and characterising it based on a predefined list of data quality characteristics. The risk analysis forces the motivated prioritisation of the determined quality characteristics, which helps to select the process to improve upon. It is also very useful in aligning the data quality improvement process with the organisation in terms of strategy.

The steps taken in the improvement cycles at Distimo are already in place and the effects on the process and thus also data quality are quite apparent and positive.

Distimo is used as a case study, but as the method uses standard information like mission statement and organisation information architecture, which are available within every organisations as its inputs, it should prove easy to apply it to

all kinds of organisations. Distimo is a very data-centric organisation, but although in other organisations the stress might not be so evidently on data itself, almost every organisation relies on information systems for daily operation as well as strategic planning. For example organisations may have systems to keep stock (e.g. a supermarket), keep track of financial transactions (e.g. a bank), or register person information (e.g. insurance companies), etc. Therefore the proposed method should provide such an organisation with a powerful tool to analyse and improve its data quality.

It is however advisable to implement a scoping step early on on the process. It is very difficult to accurately analyse all processes within an organisation. Even within a relatively young and small start up like Distimo, trying to chart the entire organisation at a detailed level proved to be quite a challenge. Furthermore, by taking on such a huge task, there is a relatively high risk of losing overview of the process and missing small details. Therefore, it would be easier to scope in the beginning by determining a subset of the organisation to research.

Chapter 7

Conclusion

This thesis answers the following research question: *"How can we determine and improve data quality?"*. The question is addressed by answering a number of derived sub-questions.

Using literature, data quality is characterised by identifying 16 data quality characteristics (see table 2.3 on page 9). These data quality characteristics can then be used to describe data quality within a certain context.

To actually measure data quality, metrics have to be defined. Once data quality characteristics have been determined, metrics may be set to measure the data quality characteristics. Combined, these give a representation of the state of the art of data quality.

A method is proposed to systematically assess the requirements for data quality and the current state of data quality within an organisation. The concept of data quality is broken down into a number of characteristics which can be used to describe the specific requirements of an organisation, or a process within an organisation. Using the analysis provided by the method, important data quality characteristics are identified. These data quality characteristics can then be made concrete by defining metrics to measure them. The results of these measurements will give an indication of where improvements can be made.

By applying the proposed method to Distimo as an organisation, a detailed analysis was made of its goals and architecture. Derived from there is a list of important data quality characteristics: accessibility, accuracy, amount of data, believability, completeness, consistency, currency, reliability and security. Risk analyses determined that reliability and security are probably the most important ones.

Because the organisations and its processes are quite big and complex, the reporting process was chosen for the improvement cycle. By making changes to the process and the underlying architecture, improvements were made in the areas of the identified data quality characteristics. A number of very labour- and data intensive tasks was automated, reducing the risk of data quality issues. The end result being a thorough understanding of Distimo's processes and systems, as well as clearly visible improvements to data quality in key processes. Given

the end result, it is safe to conclude that Distimo provides a suitable case study for the proposed method and that the method produced a satisfactory result.

As the proposed method makes use of generic organisational information like a mission statement and organisation information architecture, it can be easily applied to any organisation. Moreover, since virtually all organisation rely on information systems and thus high quality data in some way, it provides a very useful tool to monitor and enhance data quality within these organisation in a structured and systematic fashion.

Furthermore, because of the iterative character of the improvement process, the method can met used to continuously monitor and improve data quality. This is extremely useful, as organisations may change over time, and as such data quality characteristics and demands as well. Moreover, by employing an iterative process, an organisation may choose to improve data quality in small steps, which allows for close control and monitoring of proposed changes.

7.1 Recommendations

One problem observed in the application of the proposed framework is perhaps the lack of scoping. This could result in a lot of data quality characteristics being important, and even more metrics, resulting in a very complex and hard to manage risk analysis and subsequent improvement cycle. A solution to this issue may be to insert an extra step in the method to select only a part of an organisation or process.

Another interesting field to look into maybe the integration of the measurement phase as well as the improvement cycle with the development methodology in use. Methodologies like for instance continuous integration [11] would allow for automatic testing of new code when it is committed by the developer. Of course there are many different methodologies currently in practice [21].

Appendix A

Glossary

This glossary contains a list of terms and their definitions as used within Distimo and this thesis.

Term	**Description**
application (app)	A mobile application which is used on mobile devices like smart phones or tablets
app store	an application store, such as the Apple Appstore, but refers to the concept instead of the brand.
csv	comma separated values, text-based format to store row based data, with individual fields separated by a separator, this is often a comma, of semi-colon.
(data) metric	a specific data set, as sold as part of the Distimo Report portfolio. For instance the Top Applications metric contains a list of popular applications defined by the specified parameters
download data	Every time an application is downloaded from an app store, it is counted as one download.
in-app (sub-app)	An application that can be purchased to expand the functionality of an existing, already installed application. For example, additional levels to a game.
ranking	refers to the position of an application within an ordered list, for instance ordered by generated downloads, popularity or revenue
report	A report supplied to a customer. This consists of any number of metrics from a standard set, in the form of Excel spreadsheets.

Table A.1: List of used terms

Bibliography

[1] "Distimo website." http://www.distimo.com, June 2012.

[2] K. Laudon and J. Laudon, *Essentials of management information systems: managing the digital firm*. Prentice Hall, 2005.

[3] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[4] W. Eckerson, "Application development trends," *Data Warehousing Special Report: Data Quality and the Bottom Line*, 2002. cited By (since 1996) 2.

[5] Y. Huh, F. Keller, T. Redman, and A. Watkins, "Data quality," *Information and Software Technology*, vol. 32, no. 8, pp. 559–565, 1990.

[6] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, pp. 211–218, April 2002.

[7] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," *Commun. ACM*, vol. 40, no. 5, pp. 103–110, 1997.

[8] R. Daft, *Management*. The Dryden Press, 4 ed., 1998.

[9] J. Zachman, "Framework for information systems architecture," *IBM Systems Journal*, vol. 38, no. 2, pp. 454–470, 1999. cited By (since 1996) 22.

[10] R. S. Pressman, *Software Engineering: A Practitioner's Approach*. McGraw-Hill Higher Education, 5th ed., 2001.

[11] N. E. Fenton and S. L. Pfleeger, *Software Metrics: A Rigorous and Practical Approach*. Boston, MA, USA: PWS Publishing Co., 2 ed., 1998.

[12] Getjar, "Getjar website." http://www.getjar.com, June 2012.

[13] "Apple appstore." http://itunes.apple.com/us/app/id375380948?mt=8, June 2012.

[14] Apple, "Apple macstire." http://www.apple.com/macosx/whats-new/app-store.html, June 2012.

[15] "Android market." http://play.google.com, June 2012.

[16] "Blackberry appworld." http://appworld.blackberry.com/webstore, June 2012.

[17] "Nokia ovi store." `https://store.ovi.com`, June 2012.

[18] Microsoft, "Windowsphone marketplace." `http://www.windowsphone.com/en-US/marketplace`, June 2012.

[19] Samsung, "Samsung apps." `http://www.samsungapps.com/`, June 2012.

[20] "Mysql website." `http://www.mysql.com`, June 2012.

[21] D. Aveson and G. Fitzgerald, "Methodologies for developing information systems: A historical perspective," in *The Past and Future of Information Systems: 19762006 and Beyond* (D. Avison, S. Elliot, J. Krogstie, and J. Pries-Heje, eds.), vol. 214 of *IFIP International Federation for Information Processing*, pp. 27–38, Springer Boston, 2006.