Cut the Crap

A method to determine the business value of electronic documents

Michiel Bax

1/2/2010

Master of Science Graduation Thesis

Industrial Engineering & Management

Cut the Crap

A method to determine the business value of electronic documents

Author
Michiel Bax
0004723
Industrial Engineering & Management
Track *Information Technology & Management*
michielbax@gmail.com

Thesis Committee

UNIVERSITY OF TWENTE.

School of Management and Governance
Dr. A.B.J.M. (Fons) Wijnhoven
Dr. C. (Chintan) Amrit

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Financial Services Sector NL
Alexander Bijl
Hans van Rijs

SUMMARY

Research Background

We are living in the *information age*. Never before information was so widely available, the Internet and the use of ICT in organizations have helped to make this possible. The number of information systems in organizations is growing and the systems are becoming more complex. As a result, the global accumulation of data was 5 Exabyte (10^{18}) in 2002. The accumulation of data is becoming less and less controlled. The uncontrolled accumulation of data results in a number of problems for organizations:

- People are spending on average 25% of their time looking for information.
- More storage is needed with higher hardware, software and service costs.
- The risk of losing important information increases.

Information Lifecycle Management (ILM) has been developed with the goal to store data on the appropriate medium that provides the service level that is required in the phase of the lifecycle the data is in. At the end of the lifecycle the data is either archived or deleted, this way ILM aims to reduce the proliferation of data. The business value of data forms the basis to guide the data though its lifecycles.

Determining the business value of data is complex because data value is resistant to quantitative measurement. The goal of this research is to find a method that can be used to determine the business value of data in a practical way. This research focuses on the business value of electronic documents.

Research Approach

This research starts with a structured literature review to search for suitable data valuation methods. Based on the outcomes of the structured literature review the ACE framework is selected as data valuation method. The ACE framework uses policies to determine the business value to documents. Earlier research showed that it is difficult for business people to understand how these policies have to be specified. Earlier research also showed that specifying the policies is a complex and time consuming task. The next section of the research therefore focuses on developing a method that can be used to specify the policies in the ACE framework.

The goal of the policy specification method is to determine the business value of electronic documents by measuring the behavior of electronic documents and the characteristics of the users of documents. The behavior of documents is measured with '*file system metadata*'. The metadata provides information about document type, document age, last modification time of the document, document size, amount of use and document location. The position that the user of a document has in the organization is included as user characteristic.

To test if it possible to determine a business value based on document behavior, a field test is conducted. In the field test the business value of documents is determined by the users using an *'information value questionnaire'*. The '*file system metadata'* of the documents is collected to measure document behavior.

To evaluate the usefulness and practicality of the designed method, experts are interviewed. In these interviews the designed method and the results from the field test are discussed.

Research Results

The field test showed the following causal relations between document behavior and the business value of documents (see also figure 0.1):

- The business value is higher when:
  - The perceived amount of use is higher.
  - The last modification time is more recent.
- The business value lowers as the document becomes older.
- A higher grade of the user that completes the questionnaire results in a higher business value.

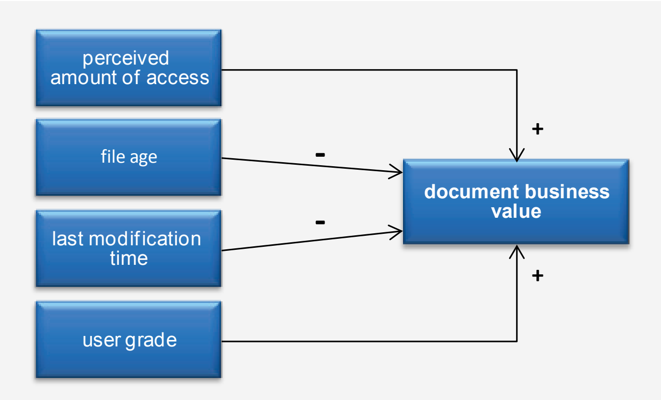


Figure 0.1: Research Findings

The interviews show many possible applications for the designed method in practice. Some examples are:

- To select valuable documents to publish on a knowledge portal.
- To help business people that have to decide which documents should be migrated to a new storage environment and which documents should be deleted or archived.
- To reduce the gap between the work of archivists and the business environment.

The experts support the use of a questionnaire because it allows business people to easily quantify the business value of a document. Furthermore, the questionnaire makes people aware the differences in business value of documents. They find the use of the questionnaire useful and practical.

Using the causal relations between file behavior and business value to specify policies and to determine the business value of documents for which the business value is unknown, sounds promising to the experts. The experts also indicate that this approach needs more testing before if it is reliable enough to use in practice.

PREFACE

This thesis is the '*piece the resistance'* of my educational career (so far). My study started with moving to Enschede, this also meant leaving the safety of my childhood home. With this change a time of exploration started, an exploration of, a new environment, new freedoms and of oneself. Because of my desired thoroughness in exploration, some explorations took a bit longer than average. I am proud that these explorations in the end also led me to this research which has proven to be the last but definitely not the least exploration so far.

This document is the result of years of work, study, interesting courses, not so interesting courses, extracurricular activities, sports, research, parties, hangovers, reading, vacations, discussions, friendships, love and above all fun! Completing my master thesis and the rest of my study is something that I could not have done alone. I would therefore like to use this opportunity to thank everyone who helped me along the way.

Capgemini thankfully provided me with all the freedom that can be needed to do a master research. I thank Alexander and Hans for the inspirational discussions, interesting conversations and good advice that I have received from them. I also thank all my fellow graduation students at Capgemini who helped me along the way with coffee breaks and other welcome distractions.

During the research I have enjoyed the visits to Fons and Chintan in Enschede. Fons, I still wonder how you managed to read through all my work every time. Thanks for the fruitful discussions, for challenging me to give my best and of course for all the laughter. Your enthusiasm is a welcome source of energy that keeps me focused on my work. Chintan, thank you for the flexibility to join in my research project when it was already so close to the finish line.

I also thank all my fellow students and friends in Enschede who made my time in Enschede more than worthwhile. A very special thanks goes out to my parents. I know that I caused some worries over the years. And I am very grateful for your support with everything that I did. I could not have done it without your help and patience, thank you. Last but definitely not least I would love to thank the girl of my life. Iris, I cannot describe how much your patience, advice and inspiration helped me to complete this research. I look forward to spending my free time together with you.

I dedicate my thesis to a person which always has been a great inspiration for me; my grandfather Dr. W.A. Bax who unfortunately passed away before I finished my study.

Michiel Bax

Utrecht, January 2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF USED ABBREVIATIONS

DBV	Document Business Value
DCE	Data Classification Engine
ECM	Enterprise Content Management
ECMS	Enterprise Content Management System
ICT	Information Communication Technology
ILM	Information Lifecycle Management
IS	Information System
IVQ	Information Value Questionnaire
MOSS2007	Microsoft Office SharePoint Server 2007
SNIA	Storage Network Industry Association
WPF	Windows Presentation Foundation

Chapter 1

1 RESEARCH INTRODUCTION

This research is conducted to find a practical way to determine the business value of data that is stored in the information systems of organizations. To explain why this is an important issue, first the problem of data proliferation will be introduced. This is followed by an introduction of the concepts of Enterprise Content Management and Information Lifecycle Management and the valuation of data. Based on this introduction the research will be initialized. In the initialization a problem statement and research question for the research is formulated.

1.1 PROBLEM BACKGROUND

We are living in the *information age*. Never before information was so widely available, the Internet and the use of ICT in organizations have helped to make this possible. The number of information systems in organizations is growing and the systems are becoming more complex. Also, people have become used to collecting and storing large amounts of data in their personal archives. As a result, the global accumulation of data was 5 Exabyte (10^{18}) in 2002 (Lynman & Varian, 2003). Every year, the accumulation of data is growing with an average rate of more than 30% (Lynman & Varian, 2003).

We are producing 5 Exabyte of data a year. How much is 5 Exabyte of data? The Library of Congress in the US is the biggest library in the world. It contained seventeen million books in 2002. 5 Exabyte; the total amount of data created in 2002, is equal to 37,000 copies of this collection; 629 billion books. Assuming an average book is about 3 cm. thick, all these books together form a stack that covers the distance from the earth to the moon more than 49 times.

Rapid growing accumulation of data in itself does not have to be problematic. The growth of data is however becoming less and less controlled. The data is proliferating. Proliferation is defined as; *“to grow by rapid production of new parts, cells, buds, or offspring”* (Merriam-Webster, 2009a). The proliferation of data results in a exponential growth of unstructured data, such as e-mail archives, intranet pages and archives which are growing at a rate of 25-30 percent a year (Govil, Kaur, Kaur, & Govil, 2008; IBM, 2006).

The proliferation of data results in a number of practical problems for organizations. It becomes harder to retrieve data promptly, more people are needed to manage all the stored data and required networks and application performance suffers because of the excess traffic that is generated by users searching again and again for data (IBM, 2006).

Besides the practical problems, the proliferation of data also results in more costs for organizations. Even though storage on a cost-per-gigabyte basis keeps declining at a steady rate, the consumption of storage is growing much faster. This results in ever increasing expenses for data storage. While vendors continue to market hardware as a way to reduce the total costs of ownership (TCO) of storage, hardware costs account for not more than 30% of the TCO. In reality, service or labor costs are the primary factors of the TCO of storage (Tallon & Scannell, 2007). Even though hardware only accounts for a small part of IT costs, storage is already consuming about 10-15% of IT budgets at the moment. CIOs fear that any further increase could seriously erode strategic IT spending (Tallon & Scannell, 2007).

The ability to effectively use information is limited for human beings. If too much information is available there is a risk for *'information overload*'. Information overload results in inefficiencies, ineffectiveness and a lower level of decision accuracy and decision quality (Edmunds & Morris, 2000; Eppler & Mengis, 2004). One of the major costs that is incurred by organizations is the large amount of time that people spent on the search for specific data, let alone the accuracy and quality of decision making which is also affected. Research shows that the average knowledge worker is spending almost a quarter of the day looking for data either internally or externally (Smith & McKeen, 2003). This means that someone who is working a full working-week spends more than one day every week looking for information.

The management and security of unstructured data is problematic (Brocke, Simons, & Schenk, 2008; Govil, et al., 2008; Moore & Karel, 2008). An example of this are five US banks who have been fined US $1.25 million each because they failed to retrieve e-mails that were demanded from them (IBM, 2006). With the proliferation of unstructured data, the security risks for important data such as critical e-mail messages, contracts, and data under privacy regulations is increasing. When the data is not stored on secure locations and access control is not properly managed the data can be lost or even fall into the hands of competitors. Unfortunately, data proliferation is still regarded an IT related issue, to be solved by IT specialists instead of demanding strategic decisions at the highest level (IBM, 2006; Munkvold, Päivärinta, Hodne, & Stangeland, 2006; Nordheim & Paivarinta, 2006; Scott, Globe, & Schiffner, 2004; Short, 2006).

ENTERPRISE CONTENT MANAGEMENT

Enterprise Content Management (ECM) is an integrated approach to managing all of an organization's information including paper documents, data, reports, web pages, and digital assets. ECM is used to create structures in the information of an organization. This structure is used to stop the proliferation of data. ECM is defined as;

> *"The strategies, tools, processes and skills an organization needs to manage all its information assets (regardless of type) over their lifecycle"* (Smith & McKeen, 2003).

Enterprise Content Management Systems (ECMS) are systems that are used to support Enterprise Content Management (ECM) activities. Establishing principles and standards for the retention, preservation and disposal of data in an ECMS is an important issue (Smith & McKeen, 2003). As ECM grows to become a corporate strategy for managing all forms of content, this issue will become an increasingly complex challenge (Smith & McKeen, 2003). To increase the control on the proliferation of data in organizations and their ECMSs, information lifecycle management is developed (Govil, et al., 2008; Middleton & Smith, 2002; Peterson & Pierre St., 2004; Reiner, Press, Lenaghan, Barta, & Urmston, 2004; Tallon & Scannell, 2007; Wrozek, 2001). This concept is therefore introduced next.

INFORMATION LIFECYCLE MANAGEMENT

Information Lifecycle Management (ILM) is an information management standard developed by the Storage Network Industry Association (SNIA). SNIA defines ILM as;

> *"Information Lifecycle Management is comprised of the policies, processes, practices, and tools used to align the business value of information with the most appropriate and cost effective IT infrastructure from the time information is conceived through its final disposition. Information is aligned with business requirements through management policies and service levels associated with applications, metadata, and data"* (Peterson & Pierre St., 2004).

The goal of ILM is to store data on the appropriate medium that provides the service level that is required in the phase of the lifecycle the data is in (Tanaka, et al., 2005). A typical ILM solution includes tiered storage hardware, a software stack that consists of storage software and middle ware such as an ECMS and databases (Chen, 2005). The tiered storage consists of expensive, fast and reliable high-end storage, less expensive, less reliable SATA-based mid-range storage, and low cost, low speed tape based storage (Chen, 2005). ILM data placement and migration policies define the conditions that determine the alignment of data to storage devices throughout the lifetime of the data.

According to the SNIA, the lifecycle of data consists of four stages. The first stage is the creation of new data or the modification of existing data. In the second stage the data is transferred to others using for instance digital, written or verbal communication. When transferred, the data is accessed and used, this is the third stage. After a period of usage the data is either archived or deleted. The final stage is called retention. The ILM lifecycle is illustrated in Figure 1.1.

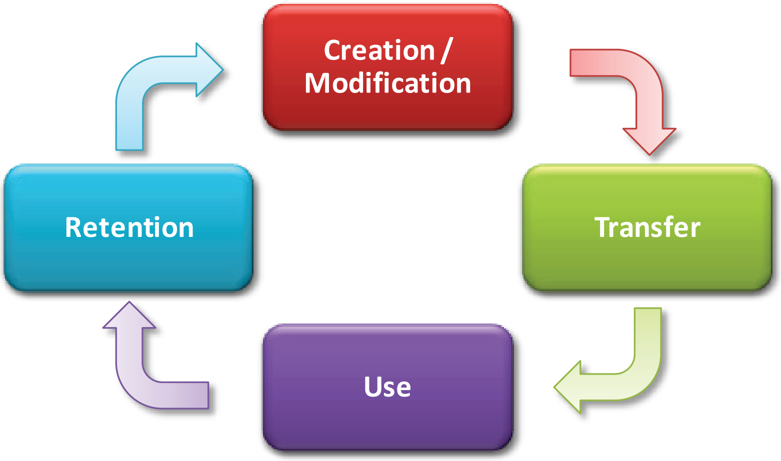
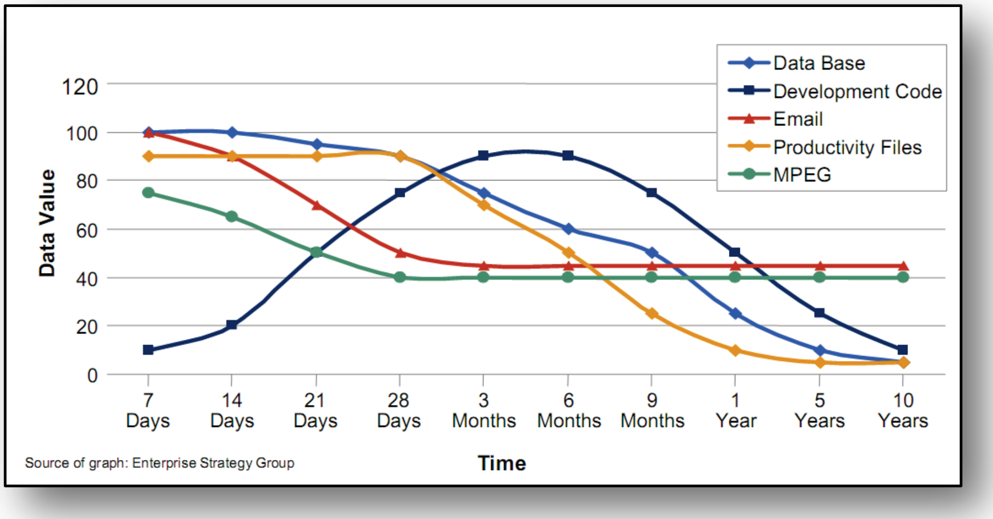


Figure 1.1: ILM Lifecycle

Because data is eventually replaced by new data or becomes less relevant as a result of new developments, the value of data follows a trend (Tallon & Scannell, 2007). Throughout the lifecycle the value of data in general grows after the first stage and declines again in the final stage. With the decreasing value of data, the intensity of usage decreases and the accessibility of data becomes less important. However, not all types of data have the same value and the way the value evolves over time can also depend on the type of data. Figure 1.2 shows some examples of how the value of different types of data can change over time.



Source of graph: Enterprise Strategy Group

Figure 1.2: Data Value Changes over Time (Haeusser, Osuna, Bosman, Jahn, & Tarella, 2007)

In order to use ILM in data warehouses the storage infrastructure has to be structured around the changing business value of data. This is a difficult and time consuming process (Shah, Voruganti, Shivam, & Alvarez, 2006). One of the most important steps towards a successful ILM implementation is the ability to differentiate data by values in an unbiased manner and understand how the value changes over time (Chen, 2005). To do so, the value of data needs to be determined. This is therefore introduced in the next section.

DETERMINING THE VALUE OF DATA

Determining the value of data has proven to be difficult. According to research, the valuation of data is a complex problem because data value is resistant to quantitative measurement (Moody & Walsh, 1999; Reiner, et al., 2004; Tallon & Scannell, 2007). Determining the value of data is essential for ILM. Proper data classification is considered as the corner-stone of ILM (EMC, 2003; L. Turczyk, 2009; L; Turczyk, Frei, Liebau, & Steinmetz, 2008). And the core of data classification lies in appropriate data valuation (Chen, 2005). The valuation is used to determine the alignment of data and the appropriate IT infrastructure (Chen, 2005; Matthesius & Stelzer, 2008; Middleton & Smith, 2002; Peterson & Pierre St., 2004; Reiner, et al., 2004; Shah, et al., 2006; Tallon & Scannell, 2007; L; Turczyk, et al., 2008; L; Turczyk, Groepl, Liebau, & Steinmetz, 2007; Wrozek, 2001).

The ability to accurately determine the business value of data is required to effectively use ILM in practice. This research is therefore conducted to find a way to determine the business value of data stored in an ECMS in a practical way.

1.2 RESEARCH INITALIZATION

The initialization begins with scoping the problem environment for the research. Based on the research scope and the problem background above, a problem statement and research question is formulated.

SCOPING THE RESEARCH

Microsoft Office SharePoint Server 2007 (MOSS2007) is an ECMS. Capgemini expects that MOSS2007 will be flourishing in years to come. A problem in the use of MOSS2007 is the proliferation of data in the system. To reduce this problem, Capgemini is looking for ways to extend the capabilities of MOSS2007. More specifically, the ability of MOSS2007 in supporting the ILM concept. The scope regarding ECMS's for this research is therefore limited to MOSS2007. A more detailed description of MOSS2007 is available in Appendix II: Microsoft Office SharePoint Server 2007.

MOSS2007 is used to support people that are collaborating in projects. MOSS2007 supplies them with a platform. This platform can be used to share data, search for available data, provide a portal to the existing knowledge base and help people to structure the processes in their projects. The data that is stored in MOSS2007 consists of documents such as Microsoft Office or PDF documents. In this research, electronic documents are the collections of data for which the business value has to be determined.

RESEARCH PROBLEM

In the previous section the background of this research is discussed. A couple of problems are identified. To summarize, the most important problems are;

- The proliferation of data results in inefficiencies, higher costs, increased security risks and a lower quality of decision making
- Information Lifecycle Management is ineffective without a method to determine the business value of data
- Determining the business value of data is a complex task

These problems are related in a sense that resolving them starts with resolving the last problem first. This research therefore focuses on resolving the last problem; determining of the business value of data. The goal is to determine the business value that is stored in an ECMS in a practical way. The problem statement for this research is therefore;

> ***Problem Statement***
>
> *"It is not yet possible to determine the business value of electronic documents in a practical way."*

RESEARCH QUESTION

The problem statement that is formulated in the previous section is a design problem. To solve the problem of determining the business value of electronic documents a method has to be found or designed. Therefore the following research question has to be answered in order to solve the research problem;

> ***Research Question***
>
> *"How can the business value of electronic documents be determined in a practical way?"*

1.3 DESIGN SCIENCE GUIDELINES

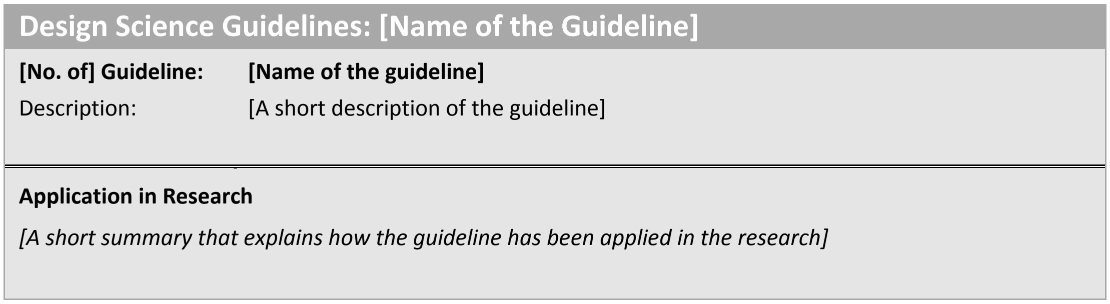
As stated in the previous section, the goal of this research is to find or design a method for determining the business value of documents in a practical way. To structure the search and design process, design science guidelines for information systems research are used. These guidelines are introduced in this section.

In their article from 2004, Hevner et. al. provide a general framework to guide information system researchers and practitioners in how to conduct, evaluate, and present design science research (Hevner, March, Park, & Ram, 2004). The work that is presented here aims to exemplify this design research approach by applying the seven guidelines as they are introduced by Hevner et. al. (2004). These seven guidelines are described in table 1.1.

Guideline	Description
[1] Design as an Artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
[2] Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
[3] Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
[4] Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
[5] Research Rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
[6] Design as a Search Process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
[7] Communication of the Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Table 1.1: Seven Guidelines for Design Research (Hevner, et al., 2004)

Throughout the rest of the thesis, references to these guidelines are made to show how the guidelines are applied in the research process. Textboxes are used to show how the guidelines are applied. The textboxes reflect on the research that is done, the design of the textbox and a description of the elements in the textbox can be found in figure 1.3.



Design Science Guidelines: [Name of the Guideline]

[No. of] Guideline: **[Name of the guideline]**

Description: [A short description of the guideline]

Application in Research

[A short summary that explains how the guideline has been applied in the research]

Figure 1.3: Design Science Guidelines Textbox Template

When a design science guideline is successfully applied, a similar textbox is presented that summarizes the application of the guideline.

The second guideline; 'Problem Relevance' is applied in this chapter. The textbox below shows how this guideline is applied.

Design Science Guidelines: Problem Relevance

Second Guideline:	**Problem Relevance**
Description:	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies. (Hevner, et al., 2004).

Application in Research

In chapter 1, the impacts that the proliferation of data has on IT budgets, efficiency of personnel and the quality and accuracy of decision making is introduced. Reducing the proliferation of data and/or the effects of data proliferation in organizations is an extremely relevant undertaking. The introduction also showed why the ability to determine the business value of data helps to solve business related issues. This demonstrates the importance and relevance of this research.

Chapter 2

2 LITERATURE REVIEW

The goal of this research is to find a method for measuring the business value of documents in a practical way. The systematic literature review is conducted to find whether such a method is developed. If a suitable method is found, its applicability in practice is evaluated. If no suitable method can be found, the findings from this literature review act as an input for the design of a new, practical method for the valuation of electronic documents.

2.1 LITERATURE REVIEW METHODOLOGY

A literature review is used to find and discuss relevant scientific publications to date. To ensure that all relevant publications are included in the review it is important to use a structured methodology for the literature review. A transparent presentation of the methodology shows the readers that the literature is not selected by random sampling, biased sampling or convenience sampling. Instead a clear, step by step approach is presented, allowing researchers to replicate and validate the search process and its outcomes. A structured approach also saves time and improves the completeness and quality of the outcomes of the literature review.

In this section the different parts of the literature review process are explained. First, the top 25 Management Information Systems (MIS) journals are selected, the journals are used to conduct an initial search for literature (§2.1.1). To perform the search a search engine is selected that covers most of the top 25 MIS journals (§2.1.2). Selection criteria are used to select the articles included in the review. These can be found in §2.1.3 in this section the used keywords are also listed. The results of search process are presented in 2.1.4.

2.1.1 TOP MIS JOURNALS

The literature search is initially limited to the top MIS journals. To select these journals, the ranking as published by the AIS has is used (AIS, 2009). In this overview, the results of 9 different published journal rankings have been averaged. There is however a risk when referring to this overview. The average rank point that determines the final ranking, has been calculated by totaling the different ranks of a certain journal and then dividing the total score by the number of rankings. Therefore, a journal that has been ranked a 6th place in one ranking will score 6th place overall. This is for instance the case with the 'Artificial Intelligence' journal (AIS, 2009). To reduce this risk, a journal is only included in the top25 if it appears in at least three different rankings.

In the top 25 as published by the AIS there are seven journals which are ranked less than three times. These journals are therefore left out. To come to a total of 25 journals, the first seven journals ranked higher 25th place and ranked more than three times are added to the selection. This results in the selection presented in Table 2.1: Top 25 MIS Journals.

Top 25 MIS Journals	
1. MIS Quarterly Management Information Systems	14. ACM Transactions
2. Information Systems Research	15. Sloan Management Review
3. Communications of the ACM	16. ACM Computing Surveys
4. Management Science	17. Academy of Management Journal
5. Journal of Management Information Systems	18. Organization Science
6. Decision Science	19. IEEE Transactions on Computers
7. Harvard Business Review	20. Information Systems Journal
8. IEEE Transactions on Computers	21. Administrative Science Quarterly
9. European Journal of Information Systems	22. Data Base for Advances in Information Systems
10. Decision Support Systems	23. Communications of the AIS
11. Information and Management	24. Journal of the AIS
12. ACM Transactions on Database Systems	25. Journal of Management Systems
13. IEEE Transactions on Software Engineering	

Table 2.1: Top 25 MIS Journals (based on AIS Journal rankings (AIS, 2009))

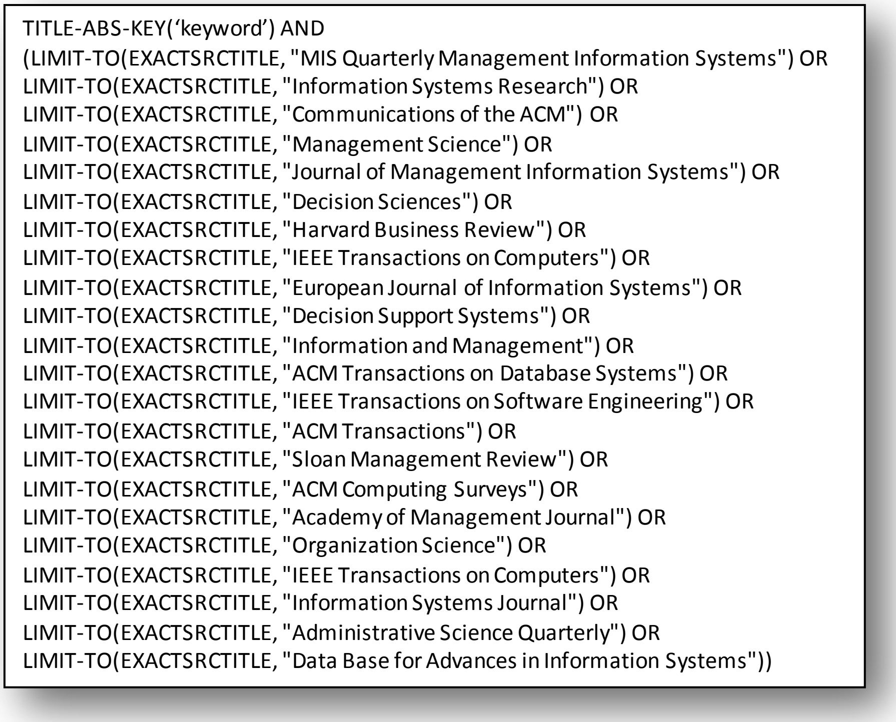
2.1.2 SELECTING A SEARCH ENGINE

To be able to search in a structured way in the journals, a search engine is used. Two examples of search engines are, Web of Science and Scopus. Both of these search engines feature a user friendly user interface and advanced search functionalities. The researcher can specify a set of journals to search and a specific time frame in which articles are published. Furthermore, forward and backward citation analysis is easy because references and citations are indexed.
Scopus is excellent because it covers 22 of the journals in the top 25. Therefore, only three journals are not covered;

- Journal of the AIS
- Journal of Management Systems
- Communications of the AIS

Because of the coverage and ease of use, Scopus is used as primary search engine and the three journals not covered by the Scopus search engine are searched manually. To be able to only search in the top 25

MIS journals from Table 1, the advanced search option of Scopus is used. The query used for searching is presented in Figure 2.1.



```
TITLE-ABS-KEY(‘keyword’) AND
(LIMIT-TO(EXACTSRCTITLE, "MIS Quarterly Management Information Systems") OR
LIMIT-TO(EXACTSRCTITLE, "Information Systems Research") OR
LIMIT-TO(EXACTSRCTITLE, "Communications of the ACM") OR
LIMIT-TO(EXACTSRCTITLE, "Management Science") OR
LIMIT-TO(EXACTSRCTITLE, "Journal of Management Information Systems") OR
LIMIT-TO(EXACTSRCTITLE, "Decision Sciences") OR
LIMIT-TO(EXACTSRCTITLE, "Harvard Business Review") OR
LIMIT-TO(EXACTSRCTITLE, "IEEE Transactions on Computers") OR
LIMIT-TO(EXACTSRCTITLE, "European Journal of Information Systems") OR
LIMIT-TO(EXACTSRCTITLE, "Decision Support Systems") OR
LIMIT-TO(EXACTSRCTITLE, "Information and Management") OR
LIMIT-TO(EXACTSRCTITLE, "ACM Transactions on Database Systems") OR
LIMIT-TO(EXACTSRCTITLE, "IEEE Transactions on Software Engineering") OR
LIMIT-TO(EXACTSRCTITLE, "ACM Transactions") OR
LIMIT-TO(EXACTSRCTITLE, "Sloan Management Review") OR
LIMIT-TO(EXACTSRCTITLE, "ACM Computing Surveys") OR
LIMIT-TO(EXACTSRCTITLE, "Academy of Management Journal") OR
LIMIT-TO(EXACTSRCTITLE, "Organization Science") OR
LIMIT-TO(EXACTSRCTITLE, "IEEE Transactions on Computers") OR
LIMIT-TO(EXACTSRCTITLE, "Information Systems Journal") OR
LIMIT-TO(EXACTSRCTITLE, "Administrative Science Quarterly") OR
LIMIT-TO(EXACTSRCTITLE, "Data Base for Advances in Information Systems"))
```

Figure 2.1: Used Scopus Query

2.1.3 SELECTION CRITERIA AND KEYWORDS

To ensure the quality and relevance of the literature found in the initial search, a number of selection criteria are specified:

- Articles have to be published in the top 25 MIS journals as presented in Table 2.1
- Articles have to be peer reviewed
- Articles are published in the year 2000 or later
- Articles have to be written in English, German or Dutch

Based on the articles that are included in the initial selection, forward and backward citation analysis is used to find more related publications. Rather than making use of the selection criteria defined above, the articles that are found by applying the forward and backward citation analysis, are included based on the relevance of their contents for this research.

The keywords or combinations of keywords that are used for the initial literature search are presented in Table 2.2. The keywords are used to search in the Titles, Abstracts and Keywords of articles.

Keywords	
information valuation	valuing information
information life cycle management	data valuation
information lifecycle management	valuing data

Table 2.2: Used Keywords

2.1.4 SEARCHING FOR ARTICLE

In this section the search method is introduced. The results of the search method for the different keywords are also presented in this section.

SEARCH METHOD

The literature search is conducted in May 2009. The query shown in Figure 2.1 is used together with the keywords from Table 2.2 to perform the initial search. Next the titles and abstracts of the initial search results are scanned and the relevant articles are added to the initial selection. Then forward and backward citation analysis is used to extend the set of articles. In forward citation analysis the researcher looks for articles that have referred to the article that he is currently reviewing. This way, more recent relevant articles in the same field of research can be found. Backward citation analysis refers to the process of evaluating the references used in the article that the researcher is currently reviewing. This allows the researcher to identify older relevant articles in the same field of research.

SEARCH RESULTS

Using the keywords from Table 2.2 and the research method described above, the literature is conducted. The number of articles found in the different stages of the literature search are presented in Figure 2.2.

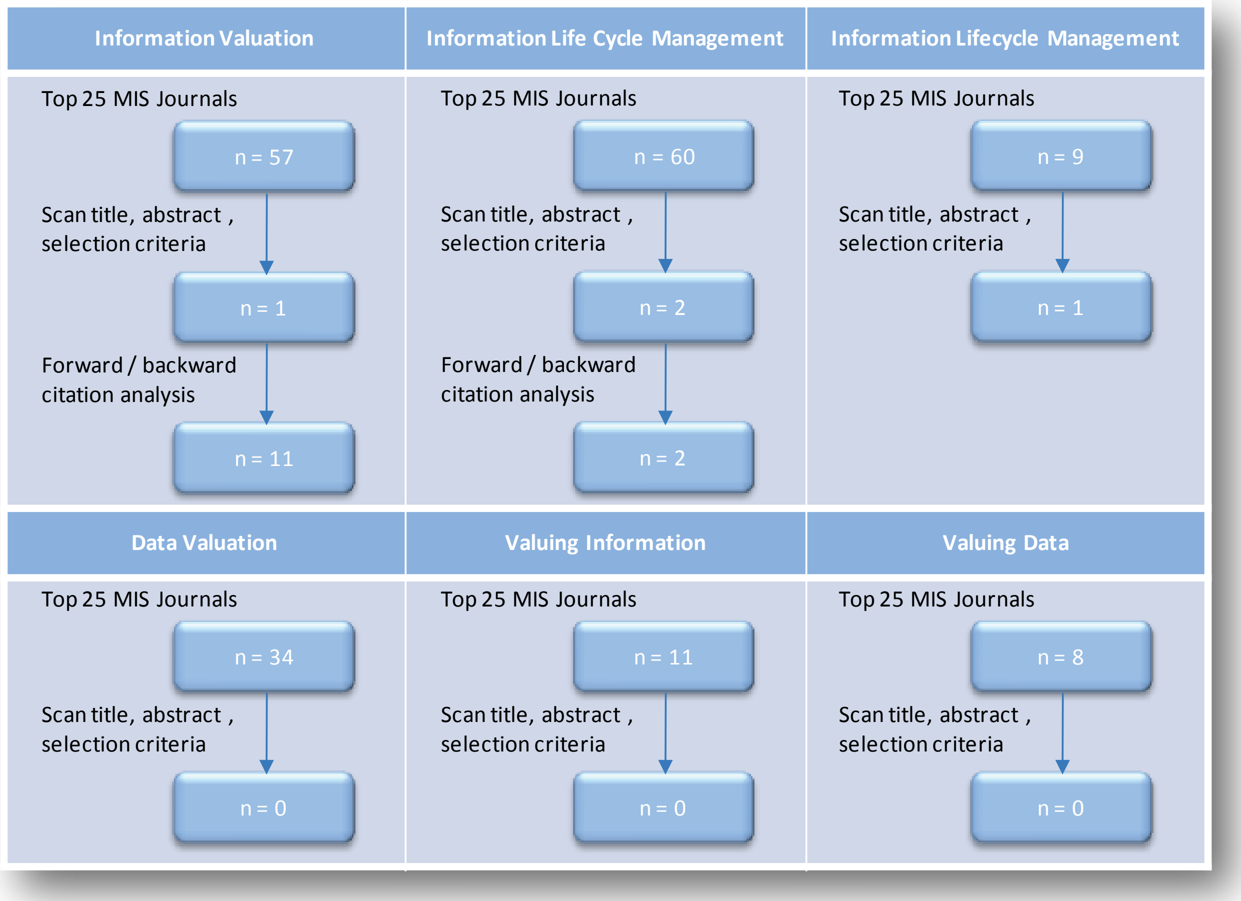

Information Valuation
Top 25 MIS Journals
n = 57
Scan title, abstract , selection criteria
n = 1
Forward / backward citation analysis
n = 11
Information Life Cycle Management
Top 25 MIS Journals
n = 60
Scan title, abstract , selection criteria
n = 2
Forward / backward citation analysis
n = 2
Information Lifecycle Management
Top 25 MIS Journals
n = 9
Scan title, abstract , selection criteria
n = 1
Data Valuation
Top 25 MIS Journals
n = 34
Scan title, abstract , selection criteria
n = 0
Valuing Information
Top 25 MIS Journals
n = 11
Scan title, abstract , selection criteria
n = 0
Valuing Data
Top 25 MIS Journals
n = 8
Scan title, abstract , selection criteria
n = 0

Figure 2.2: Search Results

As can be seen from the search results in Figure 2.2, the literature on the valuation of information is scarce. Applying the selection criteria from §2.1.3 resulted in only 3 articles for all the keywords in Table 2.2. One of the articles in the initial selection; “Information Lifecycle Management” by Tallon and Scannell (2007), referred to an article of Glazer (1993). Using forward citation analysis in combination with the article of Glazer (1993), 10 more usable articles on the valuation of data are found. These articles are also added to the selection.

2.2 LITERATURE REVIEW RESULTS

In this section the results from the literature review are presented. The goal of this literature review is to find the methods that can be used to determine the business value of data for ILM purposes. In total nine different methods have been found in the literature. In table 2.3 an overview of these nine methods is presented. The goal of the method as presented by the author(s), the measures used by the method to perform the valuation are included in the table.

Method	Goal of Method	Measures
[1] (Chen, 2005)	Captures the changing nature of file value throughout the lifecycles and presents the value differences among different files	Amount of use Recency of use
[2] (L; Turczyk, et al., 2008)	Determine the probability of future use of data to store data in a cost effective location	Time since last access Age of file Number of access File type
[3] (Bhagwan, Douglis, Hildrum, Kephart, & Walsh, 2005)	Laying out storage system mechanisms that can ensure high performance and availability	Amount of use
[4] (Verma, et al., 2005)	Optimize storage allocation based on policies	Amount of use File type
[5] (Mesnier, Thereska, Ganger, & Ellard, 2004)	Automatically classify the properties of files and predict the properties of files as they are created	Amount of use File type Access mode
[6] (Zadok, 2004)	Selecting files that can be compressed to reduce the rate of storage consumption as much as possible	Directory File name User Application
[7] (Strange, 1992)	Optimize storage in a hierarchal storage management (HSM) solution	Least recently used
[8] (Gibson & Miller, 1999)	Reduce storage consumption on primary storage location	Time since last Access
[9] (Shah, et al., 2006)	Design a data placement plan that provides cost benefits while allowing efficient access to all important data	Metadata User input Policies

Table 2.3: Data Valuation Methods

The most potent data valuation methods in table 2.3 are selected using selection criteria. These criteria are described and applied in section 2.2.1. The selected data valuation methods are described in detail in section 2.2.2.

2.2.1 SELECTION CRITERIA

Other authors mention a number of criteria for a data valuation method for ILM. These criteria are used as selection criteria for the nine valuation methods in table 2.3. The methods that fulfill all of these criteria are discussed in detail in the next section. The selection criteria are;

1. Automatic; the valuation method has to function with little to no human intervention (Chen, 2005; L; Turczyk, et al., 2008),
2. Value over time; the value of data has to be measured over time in the different life stages (Chen, 2005; L; Turczyk, et al., 2008),
3. Multiple criteria; the method has to use multiple criteria for the valuation process (L; Turczyk, et al., 2008),
4. Documents; MOSS2007 is used to store electronic documents; the selected method has to be suitable for the valuation of electronic documents.

All nine data valuation methods on table 2.3 can be automated. They therefore all fulfill the first criterion. In the valuation method of Mesnier [5] the files are only valued at the moment of creation. The value is not measured over time, this method is therefore excluded (criterion 2). The method of Verma is excluded for the same reason (criterion 3). The valuation methods of Strange [7], Bhagwan [3] et. al. and Gibson & Miller [8], are excluded because only one measure is used for the valuation of the data (criterion 3). The valuation method of Verma [4] creates storage pools to manage data storage. Policies are used to transfer the storage pools between different storage locations. This method is not suitable for the valuation of electronic documents because the valuation is focused on the pools in which the data is stored rather than the files or documents belonging to the storage pool (criterion 4).

2.2.2 SELECTED VALUATION METHODS

After applying the selection criteria in the previous section, the initial collection of nine valuation methods is reduced to four methods;

- Usage-over-Time Method (Chen, 2005)
- Probability of Further Use (L; Turczyk, et al., 2008)
- Elastic File Quota System (Zadok, 2004)
- The ACE Framework (Shah, et al., 2006)

These methods are described in detail in the next section.

USAGE-OVER-TIME METHOD

Chen has developed a *usage-over-time* approach to indirectly determine the value of a file or electronic document (Chen, 2005). The approach is based on the two fundamental conjectures;

1. Information value is realized and reflected through its usage
2. Information value changes over time

The second conjecture implicates that it has no use to refer to value of a piece of information without a reference to a specific point in time. Usage may cover multiple aspects of information usage such as usage count, the usage time, the source of usage and the purpose of usage, rather than one specific aspect of usage. The approach that has been developed is based on the usage count and the *recency* of usage.

The goal of the method is to calculate the value of information in *the present time.* This value is called *the present value.* It is therefore assumed that the history of usage serves as an indication of the importance of the information for the present time *t.* A piece of information is therefore more valuable if it is used more recently and/or it is used more heavily than others.

Another important factor is the length of the period of time which is used to calculate the value of information at time *t.* This period is called *the valuation period.* An effective valuation period can be determined by repeating the information valuation with increasing valuation period values and checking of the outcomes change significantly. If this is not the case anymore, the valuation period can be set according to that value.

To incorporate the *recency* factor, *the valuation period* is divided into fixed length *lifestages.* Different weights are assigned to the stages, the more recent the higher the weight.
The overall valuation method is defined as follows (Chen, 2005);

$$\mathrm{V_t}(d) = \sum_{i=1}^{N_t} \left(\mathrm{w(i)} \times f\left(U_i(d)\right)\right), \qquad 0 <= f\left(U_i(d)\right) <= 1,$$

$$\mathrm{w(i)} = \frac{\frac{1}{\chi}^{\mathrm{i}}}{\sum_{\mathrm{j=1}}^{\mathrm{N\,t}} \frac{1}{\chi}^{\mathrm{j}}}, \sum_{\mathrm{i=1}}^{\mathrm{N_t}} \mathrm{w(i)} = 1,\ \chi >= 1 \qquad (1)$$

$$\upsilon = [\, t - (N_t \times s\,), t],\ N_t = \frac{\upsilon}{s}$$

Where $\mathrm{V_t}(d)$ is the value of a piece of information d, at time t. *The valuation period* is denoted by υ, the length of the *lifestages* is s and N_t is the number of *lifestages*. $f\left(U_i(d)\right)$ represents the normalized usage of information d in its *lifestage* i. Its value is between 0 and 1. The normalization function is presented in formula 2. $\mathrm{w(i)}$ is the normalized recency weight for *lifestage* i. A smaller i represents a more recent *lifestage*.

$f\big(U_i(d)\big)$ has to be normalized to values between 0 and 1. In order to do this a *file access count scaling factor; c,* is introduced. Selecting must be done with care since sometimes there may be a few outlier files that have much higher access counts than others. If c is too high, only those outliers may be assigned with high values, while all others are assigned with low values even if there are still significant differences among the remaining files in reality. If c is too low, the model will generate high values for most of the files. $f\big(U_i(d)\big)$ is therefore defined as;

$$\big(U_i(d)\big) = \#\ \text{access of information } d \text{ in lifestage i.}$$

$$f\big(U_i(d)\big) = \begin{cases} 1, & U_i(d) > c; \\ \frac{U_i(d)}{c} & otherwise \end{cases} \quad c = assigned\ value \qquad (2)$$

Given the same N_t , the larger the χ is, the steeper the weight distribution is. Similarly, given a fixed χ, the larger the N_t , is, the steeper the weight distribution is. In general, significantlyflat or steep weight distributions should be avoided. Flat weight distribution essentially ignores the effect of the usage *recency* while the steep distribution considers primarily only the most recent usage and ignore all the past implications.

The model of Chen is very promising because it is a fully automatic method that provides comprehensible results. A drawback of this method is that it does not incorporate the knowledge of administrators and users about the systems and information. This could strongly enhance the capabilities of the model (Chen, 2005; Matthesius & Stelzer, 2008). Furthermore the method does not take into account that the value of information does not necessarily has to reflect in the usage of the information. For instance, a trade agreement or contract is of critical value for the business, the usage count for these type of documents does not necessarily has to be high. Developing and adding a classification scheme based on the contents of files or documents could further increase the effectiveness of this method.

PROBABILITY OF FURTHER USE

Turczyk developed a method which indirectly determines the value of a file based on usage information and expresses it as *a probability of further use (L; Turczyk, et al., 2008; L; Turczyk, et al., 2007).* Instead of using algebra such as Chen (2005), statistical distribution methods are used in the method of Turczyk et. al.

As measures for the value of a file the following variables have been reviewed;

- number of access;
- size of the files;
- size of the access;
- age of the files;
- file types (extensions);
- access types (version fetched, view, version added, move, reserve, unreserve, permission changed or miscellaneous).

These measures are analyzed to see whether they can be used to predict the access behavior of files. To do so, the correlation between these measures and the *number of days since last access* of a file is evaluated using Q-Q plots. If the hypothesis is accepted, the found distribution can be used to predict *the probability of further use* of that specific class of files (a subgroup within one of the measures).

Using the distributions that are accepted and given the days until last access of a file, *the probability of further use* of that file can be determined. The found distribution can also be used to define migration rules. To do so, a threshold value for a certain class of files can be set, for example;

> *"Migrate the file to the next tier of storage, if the probability of further access is below 5%"*

When the valid distribution has been found, the number of days since last access belonging to this rule can be calculated.

The method of Turczyk is the first known method that uses probabilistic methods to predict the future value of a certain file. No metadata is required to perform calculations, according to Turczyk this is a major benefit because of the effort that is required to collect and update the metadata over the lifecycles of files (L; Turczyk, et al., 2007).

Before this method can be used in a database, the files in a database first have to be examined and a classification of files has to be made in order to find the distributions which are suitable to predict the future value of files. Further research is therefore required to see how this method copes with very dynamic databases. A drawback of this method is that all calculations are based on the characteristics and use of files, the content and context of a file is not considered in the calculations.

ELASTIC QUOTA FILE SYSTEM

The EQFS method developed by Zadok et. al. aims to reduce the growth of data with an intelligent set of policies (Zadok, 2004). They have identified three ways to reduce growth. First, data can be compressed with lossless compression file systems. This method has very little risk since no data is destroyed.

Second, multimedia files such as MP3 or JPEG are re-encoded with lower quality, this carries some risk, but the data is still available and useful. Third, reproducible files (e.g. '~' files in windows file systems) can be removed, this method carries more risk since the file must be regenerated before it can be used again. These three methods have been tested in five data centers. The results were encouraging, between 16% and 73.2% of total disk space was saved, with an average of 48% over the five data centers.

After this first test, a new functionality was added to the method. To files that are compressed (lossless or lossy) are called 'elastic files'. Next to elastic files, there are persistent files, these files are never compressed. Users have a limited space for persistent files stimulating the use of elastic files as much as possible. To be able to select files that have to be become elastic, five methods were created. First, users can toggle the elasticity of a file on a per file basis. Secondly this can be done on a directory basis. Third, users can determine whether new files should be created elastic or not. Fourth, users can inform the system of files that should be elastic based on their file name or file type. In file systems there are a lot of temporary files, these files are created by applications, not by users. The last method therefore allows application developers to determine which temporary files should be created elastic.

The EQFS method is interesting because it shows how the experience of administrators and users can be used to identify the elastic files in a system. Based on the classification, policies for file handling are applied. When defining the policies, three considerations have to be made; convenience, fairness and gaming (Zadok, 2004).

Convenience The system should be easy to use and simple to understand. Users have to be able to see how much quota they have left and which of their elastic files will be deleted first.
Fairness It is important to provide a number of policies that can be tailored to the specific needs of a user. There are largest-file-first removal policies which might be considered unfair by users because recently created files may be reclaimed after a short period of time. Or an oldest-creation-time removal policy which is unfair because it does not take into account the amount and recency of use.
Gaming Users can find ways to circumvent the system and prevent their files from being deleted first. For instance when largest-file-first removal policies are used, users might find ways to split up files preventing them from being deleted. Or when old files are removed first owners simply access or touch files to circumvent the system. Good policies should be resistant to gaming. Especially in large group file systems where users are anonymous to each other, they will try to get as much out of the file system as possible, this increases the risk of gaming.

THE ACE FRAMEWORK

ACE is a framework of tools for ILM, that classifies data and storage resources, and generates a data placement plan for informed utilization of the available storage resources in the system (Shah, et al., 2006). ACE uses a policy-based approach to classify data based on metadata attributes. A classification of storage locations is made based on the technical capabilities of the storage hardware.

ACE consists of a data classification engine, a storage classification engine and a data placement engine that maps the data to the appropriate storage. The ACE architecture is as follows;

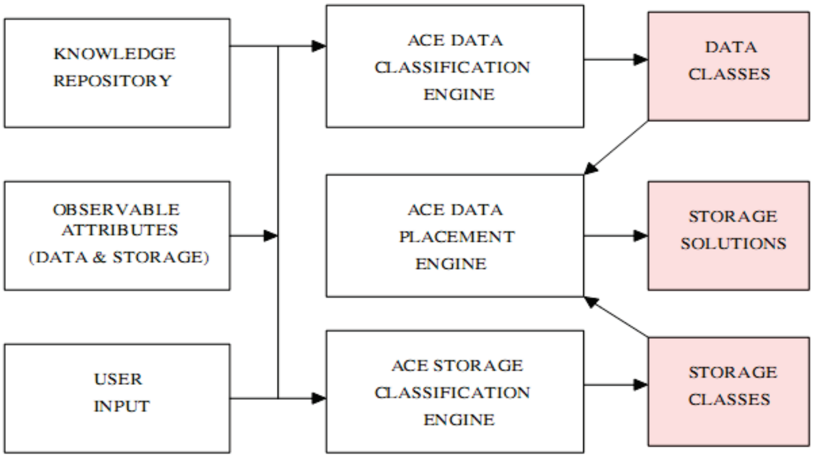
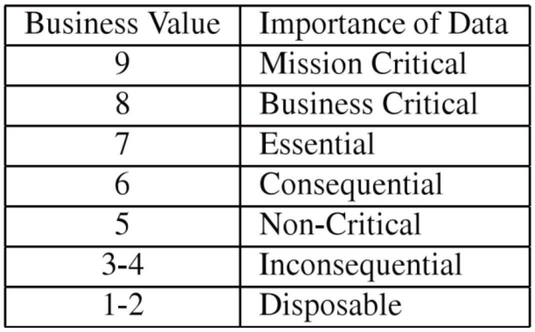

KNOWLEDGE REPOSITORY
OBSERVABLE ATTRIBUTES (DATA & STORAGE)
USER INPUT
ACE DATA CLASSIFICATION ENGINE
ACE DATA PLACEMENT ENGINE
ACE STORAGE CLASSIFICATION ENGINE
DATA CLASSES
STORAGE SOLUTIONS
STORAGE CLASSES

Figure 2.3: ACE Architecture (Shah, et al., 2006)

For this research the primary interest is the data classification engine of the ACE framework. The ACE framework has a number of key features which are as follows;
Classification ACE semi-automates the process of determining the business value of data, identifying the different classes of data based on their business values
Policy-driven Business Valuation To aid administrators to specify the business value of data, a policy-driven valuation mechanism can be used. The policies determine how the data gets mapped to different business values.
Temporal Business Value By monitoring the system and changing metadata characteristics the temporal nature of business value can be handled.

The classification engine in ACE mines the metadata attributes of data and provides an appropriate business value of data based on the available policies. The output of this engine consists of a collection of data objects with similar characteristics and the same business value. A range from 1-9 is used to assign business value to data according to the following mapping;

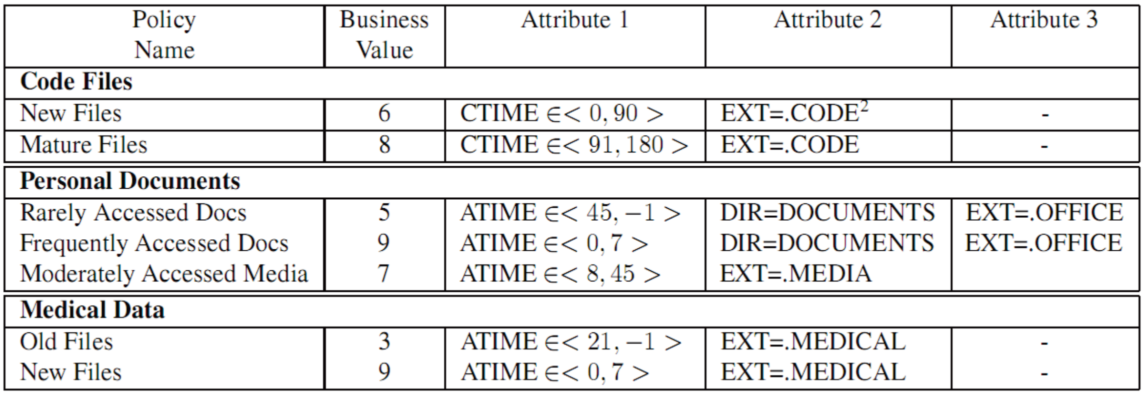


Business Value	Importance of Data
9	Mission Critical
8	Business Critical
7	Essential
6	Consequential
5	Non-Critical
3-4	Inconsequential
1-2	Disposable

Figure 2.4: Mapping of business value to the importance of data (Shah, et al., 2006)

As can be seen in Figure 2.3, the data classification engine uses three sources of input, a knowledge repository, observable attributes and user input.

Knowledge Repository The repository is a collection of policies that encapsulate domain knowledge for data classification. A policy consists of a set of observable attributes of the data, the corresponding attribute values and business value for the data that matches these attribute values. Each data object is compared with these policies to determine which one suits the most. For some examples of the policies in the ACE tool, see figure 2.5.



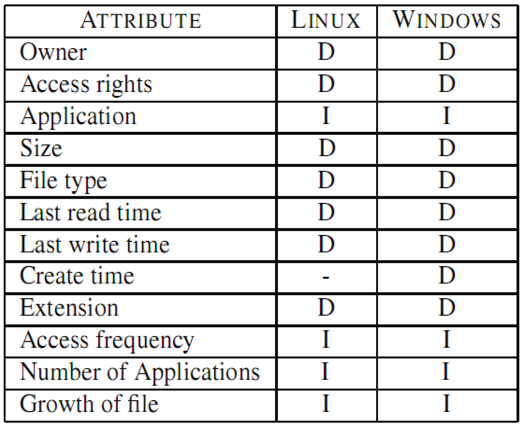
Policy Name	Business Value	Attribute 1	Attribute 2	Attribute 3
Code Files				
New Files	6	CTIME $\in< 0, 90 >$	EXT=.CODE^2	-
Mature Files	8	CTIME $\in< 91, 180 >$	EXT=.CODE	-
Personal Documents				
Rarely Accessed Docs	5	ATIME $\in< 45, -1 >$	DIR=DOCUMENTS	EXT=.OFFICE
Frequently Accessed Docs	9	ATIME $\in< 0, 7 >$	DIR=DOCUMENTS	EXT=.OFFICE
Moderately Accessed Media	7	ATIME $\in< 8, 45 >$	EXT=.MEDIA	
Medical Data				
Old Files	3	ATIME $\in< 21, -1 >$	EXT=.MEDICAL	-
New Files	9	ATIME $\in< 0, 7 >$	EXT=.MEDICAL	-

Figure 2.5 : Table showing some sample data classification policies for different domains. CTIME = Creation time, ATIME = Last Access Time, EXT = Extension. Some of the values such as .CODE and .OFFICE actually represent an array of values (Shah, et al., 2006)

In the policies above for example, a code file that is created between zero and ninety days ago is assigned a business value of '6'. And an office document that is accesses more than 45 days ago is considered 'rarely accessed data' and therefore is assigned a business value of '5'.

Observable Attributes These attributes come from the mining of metadata of the different data objects. The type and amount of attributes that are available depend on the data files and the file system that is used. The observable attributes that have been included in the ACE research are presented in figure 2.6.

User Input The administrator can provide additional input on how to classify the data, sample files can be used or customized policies can be added to the knowledge repository. The suggestions made by ACE can also be overridden.



ATTRIBUTE	LINUX	WINDOWS
Owner	D	D
Access rights	D	D
Application	I	I
Size	D	D
File type	D	D
Last read time	D	D
Last write time	D	D
Create time	-	D
Extension	D	D
Access frequency	I	I
Number of Applications	I	I
Growth of file	I	I

Figure 2.6: Attributes that have been mined for data classification in the ACE research. D = directly available, I = inferred using internal ACE mechanisms, or a combination of native system APIs (Shah, et al., 2006)

For the actual classification of data, two sorts of policies can be used; knowledge-based policies and expert-based policies.

Knowledge-based policies come prepackaged in the ACE framework based on the experience of experts. It is possible that a data object does not satisfy all the attribute values that have been defined in a knowledge-base policy, hence the data object cannot be classified directly. To be able to still classify the data the ACE framework applies the policy with the largest fraction of matching attributes. Let the matching policies be $P_1, P_2, P_3, \dots P_n$ with the number of attributes in each policy $T(P_i)$. Let the number of matching attributes be $M(P_i)$. Then the data object is classified with policy P_i, argmax $(\frac{M(P_i)}{T(P_i)})$, providing the matching attributes is greater than 50%. If two policies have the same fraction of matching attributes then by default the policy with the highest value is assigned to the data object. When none of the policies have a maximum matching ratio of 50% ACE assigns a default business value .

Expert-based policies allow the administrator to rank attributes relatively to define a custom policy. Suppose an administrator selects two attributes; $A_1 =$ owner and $A_2 =$ last access time. She ranks owner as being most important and access time as less important. ACE will map these relative ranks to actual ranks R_1 and R_2. The administrator will also provide three values for the owner attributes $a_{11,}a_{12}, a_{13}$ these values are also ordered based on importance. ACE will also generate internal scores for these values; $s_{11,}s_{12}, s_{13}$. Similar there are attribute values for last access time $a_{21,}a_{22}$ with scores $s_{21,}s_{22}$. Based on these values ACE generates an internal policy function for a data object d as follows;

$$BV(d) = R_1 \cdot ((s_{11} \cdot v(a_{11}) + s_{12} \cdot v(a_{12}) + s_{13} \cdot v(a_{13}) + s_{21} \cdot v(a_{21}) + s_{22} \cdot v(a_{22}))$$

Where $BV(d)$ is the business value of data object d. $v(a_{ij})$ = 1 if $A_i = a_{ij}$, and 0 otherwise. Generalizing for any policy:

$$BV(d) = \sum_{n=1}^{n} R_i \cdot \sum_{j=1}^{m_i} s_{ij} \cdot v(a_{ij})$$

Where n is the total number of attributes in the expert policy and m_i is the number of values for attribute i.

The ACE framework is an interesting example of a complete approach of data handling for ILM. It presents tools and methods for the classification of data and storage locations as well tools for data placement. The data classification method of ACE is based on metadata (data attributes) these attributes are compared with predefined policies. These policies are very important because they determine the business value of a data object. In the article of Shah is stated that these policies are included in the framework and are based on the consultation with experts and experience (Shah, et al., 2006). How these policies have been collected is not discussed. This makes the use of ACE framework based on this article problematic. Without proper guidelines the policies are difficult to formulate, without the correct policies, the ACE framework will not function.

2.2.3 ASSESMENT OF METHODS

To compare the methods introduced in the previous section, different assessment criteria discussed in ILM related publications are used.

Chen formulated five assessment criteria for a data valuation method for ILM automation, these criteria are (Chen, 2005):

- Require little or no human intervention;
- Rely on tangible and measurable metrics;
- Be simple and comprehensible to allow users to easily interpret the valuation outputs and gain insight;
- Capture key trends in information values: the differences among information and value changes over time;
- Adapt to changing environments.

Turczyk introduced three critical characteristics for data valuation methods (L; Turczyk, et al., 2008):

- Valuation methods have to be automatic, manual valuation is not feasible because of the labor intensiveness and the low speed of valuation.
- The valuation of data has to be dynamic , preferably on a daily basis.
- Multiple criteria for valuation have to be used to perform a realistic valuation of files.

Matthesius and Stelzer used the following assessment criteria to compare data valuation methods (Matthesius & Stelzer, 2008):

- The method incorporates 'amount of use' as a valuation criterion;
- Divides the data into classes;
- The method can be automated;
- Uses the knowledge and experience of data managers and users;
- Takes system performance into account;
- Aims for cost reductions.

Peterson&Pierre defined ILM in 2004. This is a part of the definition; "…*to align the business value of information with the most appropriate and cost effective IT infrastructure…"* (Peterson & Pierre St., 2004). A data valuation method should therefore try to determine the *business value* of the data. The business value of data is not necessarily reflected in the use of the data (Chen, 2005). Valuating data purely on amount of use is not sufficient (Jin, Xiong, & Wu, 2008). The data valuation methods are therefore also assessed on the following criterion:

- The data valuation method determines the *business value* of data.

Some of the assessment criteria mentioned by the different authors, are redundant. These criteria are therefore combined. Also, some of the criteria have already been used as a selection criterion in section 2.1.3. The methods that have been described in section 2.2.2 all fulfilled the selection criteria. The selection criteria are therefore not used again as assessment criteria. The following criteria remain to assess the different valuation methods; to provide proper valuation for ILM purposes, the valuation method must fulfill the following key criteria:

1. Require little to no human intervention;
2. Incorporates 'amount of use' as a valuation criterion;
3. Rely on tangible and measurable metrics;
4. Divides the valuated data into classes;
5. Uses the knowledge and experience of data managers and users;
6. Aims for cost reductions;
7. Takes system performance into account;
8. Determines the *business value* of data.

The assessment criteria have been evaluated for the four methods that are discussed in section 2.2.2. In table 2.4 an overview of this assessment is provided.

Criterion / Method	(Chen, 2005)	(L; Turczyk, et al., 2008)	(Zadok, 2004)	(Shah, et al., 2006)
[1] Little human intervention	X	X	X	X
[2] Amount of use	X	X		X
[3] Tangible and measurable metrics	X	X	X	X
[4] Classification of data	X	X	X	X
[5] Knowledge and experience of data managers and users			X	X
[6] Cost reductions	X	X	X	X
[7] System performance				X
[8] Business value of data				X

Table 2.4: Assessment of Methods

The ACE framework is the only method which fulfills al the assessment criteria. The ACE framework is therefore chosen to use in this research. However, before ACE is applied as a valuation method, difficulties in the application of the method that were found in the literature review are discussed.

DIFFICULTIES OF THE SELECTED METHOD: THE ACE FRAMEWORK

The ACE framework uses a policy based approach for ILM. According to literature there are two main problems with the use of a policy based approach for ILM. These problems are both related to the specification of the policies. These problems are discussed next.

The first problem concerns the person who is responsible for specifying the policies. Policies should be specified by an information administrator not by a system administrator (Tanaka, et al., 2005). An information administrator is a business person. A business person often has difficulties with understanding metadata attributes that are important inputs for policies. This makes it difficult for a business person to specify policies (Tanaka, et al., 2005).

The second difficulty concerns the required effort for the specification of policies. Ohta et. al. conclude that developing ILM related policies is a time consuming and complex task (Ohta, Dai, Kobayashi, Taguchi, & Yokota, 2006) and building a complete set of policies is a problem (Jin, et al., 2008). Administrators use rules-of-thumb for policy selection, often in anticipation of a certain workload. There are two problems with this. First, setting global policies often results in few files getting managed optimally, as a generic choice is made. Second, workloads are complex and variable, often preventing effective human configurat ion (Mesnier, et al., 2004). Without policy based migration ILM is not powerful and building those policies is a huge undertaking both from a definition and a data ownership standpoint (Short, 2006).

This chapter shows that a method for the specification of policies for ILM data valuation methods is an important research contribution. Thus, the fourth design science guideline has been applied. The textbox below shows how this is done.

Design Science Guidelines: Research Contributions

Fourth Guideline:	**Research Contributions**
Description:	The objective of design-science research is to develop technology-based solutions to important and relevant business problems (Hevner, et al., 2004).

Application in Research

The literature review showed the problems that are related to the specification of policies for ILM data valuation methods. A method to specify these policies is an important contribution that extends and improves the capabilities and usefulness of the ACE framework and the existing knowledge base.

2.3 SUMMARY

In this chapter a literature review is used to find valuation methods for data. Nine methods are identified in literature and summarized in Table 2.3. Based on criteria from literature a selection of the valuation methods is made. Four methods are discussed in detail. These four methods are evaluated using assessment criteria from literature. The assessment showed that the ACE framework is the valuation method with the most capabilities. The ACE framework uses policies to manage the data and guide its lifecycle. Research shows that specifying the policies is problematic.

In order to use the ACE framework effectively, a method for specifying policies is required. In the next phase of this research such a method will therefore be designed and evaluated.

Chapter 3

3 CONCEPTUAL POLICY SPECIFICATION METHOD

In this chapter a conceptual method which can be used to specify policies in the ACE framework is developed. First the relevant parts of the ACE framework are briefly discussed. Second, it is showed how the required inputs for a policy in ACE can be collected. After collecting the inputs new policies can be specified with the use of statistical techniques. The last section describes how this can be done.

3.1 RELEVANT PARTS OF THE ACE FRAMEWORK

For this research the main interest of the ACE framework is its data classification engine (DCE). The DCE compares the values of observable attributes from metadata with (pre-) defined policies in a knowledge repository to make a classification of data. The users of the system can modify, add or delete policies in the knowledge repository. A detailed description of the DCE can be found in section 2.2.2. The different parts of the DCE are displayed in figure 3.1.

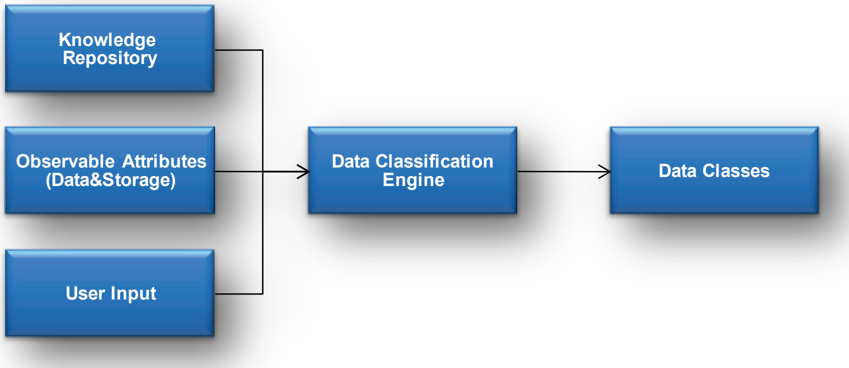

Knowledge Repository
Observable Attributes (Data&Storage)
User Input
Data Classification Engine
Data Classes

Figure 3.1: Data Classification Engine Architecture

In the description of the ACE framework is mentioned that knowledge-based policies come prepackaged with the ACE framework. These policies have been developed by consulting experts and making use of the experience of the developers of ACE. This description is rather vague, no additional guidelines on the developments of policies is provided. Because policies play a key role in the classification of data with

ACE, a structured method that guides in the development of policies is needed to improve the effectiveness of the use of ACE.

3.2 POLICIES IN THE ACE FRAMEWORK

The policies in the ACE framework consist of a set of observable attributes of the data, the corresponding attribute values, and a business value for the data that matches these attribute values. The observable attributes can be found in Table 3.1 and some examples of policies in Figure 2.5.

As discussed in section 1.1, because of the non-triviality of data valuation, most administrators need guidance in the data valuation process. According to Shah (2006) it is required that administrators are aware that different metadata attributes are relevant for different types of data classification. When administrators are not capable of doing so, developing effective policies for the knowledge repository is impossible. Without these policies the ACE framework is of no use whatsoever. It is therefore critical for the effectiveness of ACE to have a method that assists in the development of these policies.

Three inputs are required for a policy in ACE:

- Business value
- A set of observable attributes
- A specific value or value ranges for these attributes

How the appropriate information for these three inputs can be collected is discussed in the next section.

3.2.1 BUSINESS VALUE

In the ACE framework policies are used to automatically assign a business value to data. In this research, ACE will be used to assign a certain business value to a document. What is business value and how can the business value of a certain document be assessed? In the conducted literature review, two information valuation approaches have been identified which can be used to assess the business value of a piece of information. The methods are the asset valuation approach and the information value questionnaire. These methods are introduced next.

ASSET VALUATION APPROACH

Moody and Walsh (1999) have developed a method that allows organizations to value information consistent with accepted accounting principles. To do so, the major asset valuation paradigms from accounting theory are (Godfrey, Hodgson, Holmes, & Kirsch, 1997);

- Cost (historical cots)
- Market (current cash equivalent)
- Utility (present value)

Moody and Walsh conclude that utility value; *"the benefits that can be derived from information, in terms of future cash flows"*, is theoretically the best indicator for the value of information. It takes into account how the information is used. However, estimating the future benefits is very time consuming and very subjective. It is very difficult to isolate the influence of a piece of information on the organization's future revenue and thus its value. Information acts as a catalyst rather than a direct source of revenue (Moody & Walsh, 1999).

Using market value; the amount that other firms are willing to pay for information, is only possible in very few cases where information is sold as a product in its own right. And if information is sold as a product over and over again, the utility value model is more suitable to use.

The historical cost; how much was originally paid to acquire the asset, seems to be the most workable method for the valuation of information. The method is reliable, it is easy to collect necessary data and most importantly, it can be applied to all information. Moody and Walsh add that the traditional cost method used in accounting is not suitable as-is, because the concept of *use* is not incorporated (Moody & Walsh, 1999). They therefore propose a number of modifications to the traditional cost method:

- The costs for the collection of information should be standardized for all data items.
- Management information should be valued based on the costs resulting from the processes that are used to abstract the information from operational systems .
- Redundant information has zero value.
- Unused information has zero value.
- The number of users and the amount of access to data should be used to multiply the value of information.
- The value of information should be depreciated based on length of its lifecycle.
- The value should be discounted based on the relative accuracy of information compared to what is *acceptable,* if information is more accurate than acceptable, the value of the information increases, if the accuracy is lower than acceptable, the value decreases.

INFORMATION VALUE QUESTIONNAIRE

Sajko et. al. (2006) developed a questionnaire that allows information workers to value the information they use. The answers of this questionnaire result in a total score for a certain piece of information.

According to Sajko the value of information has five dimensions:

- Lost
- (Re)building
- Market Value
- Legislative
- Time

For each of these dimensions five possible ratings have been formulated, the information worker has to choose the most suitable rating. The dimensions and the range of the ratings are therefore summarized next: The lost dimension tries to measure what the consequences for the operations for the business will be if the information is lost. This can be anything from *nothing special* to *making wrong decisions with major consequences.* (Re)building measures the cost that are incurred when the information has to be replaced or has to be produced again (from *negligibly small* to *intolerably high costs*). The market value measures the consequences if the information comes into the hands of competitor (from *nothing* to *competitor gets competitive advantage*). Legislative identifies if there is an obligation to keep the information and if so, what the consequences are if the information is lost (from *no obligation* to *keeping information is obligatory and sanctions are strict*). Time measures the rate with which the information depreciates in value (from *very quickly* to *does not depreciate at all*).

The questionnaire for the assessment of information value is presented in Appendix III: information value Questionnaire. The rating that is chosen by the information worker relates to a score between 0-4. These scores are summed, this results in a total score for the business value of the reviewed piece of information between 0-20.

This method is easy to use for information workers. By answering a few questions about the information they possess, it allows them to rank the different pieces of information. Combining the rankings of more people of a certain piece of information, can reduce the variation in rankings, making the method more reliable. This questionnaire method cannot be automated and is therefore not directly suitable for use in ECMSs. The measures that are used are subjective, the rankings of different persons are therefore required to create intersubjective reliability. Performing a valuation of information with this method can be a labor intensive task, especially if lots of pieces of information have to be valued (Kaiser, Smolnik, & Riempp, 2008). The amount of labor can however be distributed over a large number of people minimizing the amount of labor per person as much as possible.

SELECTING A METHOD

For the development of policies for the ACE framework a number for the business value of a document is required. As discussed in section 2.2.3, policies should not be specified by system administrators, but information administrators from the business should do the specification (Tanaka, et al., 2005). The valuation paradigms as introduced by Moody and Walsh are too complex to use in practice (Matthesius & Stelzer, 2008; Shah, et al., 2006; L. Turczyk, 2009). And the information that is needed to do a valuation based on one of these paradigms is difficult to collect (Matthesius & Stelzer, 2008; L. Turczyk, 2009).

Using the information value questionnaire of Sajko is straightforward. The questions are simple and easy to understand. A drawback of this method is the intersubjectivity of the valuation. Because people answer the questionnaire according to their perceptions, the business value that is determined is 'perceived business value'. By combining multiple valuations, the outcomes of the information value questionnaire (IVQ) become more reliable. Because of its ease of use, the IVQ will be used to guide information workers in assessing the business value of documents. Besides the business value of documents, there are two more inputs required to specify a policy in the ACE framework. The outcome of the IVQ is a score between 0-20. In this research this score is called; *'document business value'*.

To complete a policy in the ACE framework a set of observable attributes and a certain range or value of these attributes is also needed. How these attributes and their values are collected is discussed in the following section.

In this section the first design science research guideline has been applied. The policy specification method is the artifact that is produced in this research. The textbox below shows how the guideline is applied.

Design Science Guidelines: Design as an Artifact

First Guideline:	**Design as an Artifact**
Description:	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation. (Hevner, et al., 2004).

Application in Research

The policy specification method that is introduced in this chapter uses a combination of existing methods and approaches. The ACE framework is extended with the information value questionnaire to provide the ACE framework with the inputs that are required to develop policies for ILM. This creates a policy specification method which is uses inputs that are provided by the owners and users of documents.

3.2.2 OBSERVABLE ATTRIBUTES AND THEIR VALUE

To complete a policy for the ACE framework, a set of observable attributes and a range or a specific value of these attributes is required. The attributes that are required are available in the metadata of the file system in which the documents are stored. The attributes found in table 3.1 are directly available in a Windows file system. The values of these attributes are also directly available from the file system.

Metadata Attributes		
File Type (extension)	File last Access Time	File Name
File Size (KB)	File Last Modification Time	User
File Creation Time	File Path	

Table 3.1: Metadata Attributes in a Windows File System

Chapter 4

4 METHOD EVALUATION

The conceptual method to specify policies in the ACE framework that is developed in the previous chapter has to be evaluated. This can be done by developing propositions about the method and testing them. These propositions are introduced in section 4.1. To test the propositions a dataset is required. To collect this dataset a field test is conducted. The field test is conducted in the Capgemini Financial Services NL sector using custom made software. In this chapter the goal of the field test is explained (4.2). Next, the design and test process of the custom software, and the way the field test is conducted is discussed in sections 4.3 and 4.4.

4.1 PROPOSITIONS

The goal of the conceptual method developed in chapter 4, is to define policies in the ACE framework. There are three different inputs required to define a policy in the ACE framework. The first two inputs are the attributes from metadata and their values. The attributes from metadata and the values of these attributes are directly available from the file system in which the documents are stored. The attributes are presented in Table 3.1.

The third input for a policy is the business value of the document. The business value is assessed by business people using the Information Value Questionnaire (IVQ). This questionnaire assesses the *'document business value'* (DBV) of a respondent by asking him five questions related to the value of the document.

It is expected that the behavior of a document has causal relations with the business value of the document. Based on these causal relations it is possible to explain the business value of a document, using the behavior of the document. Based on these causal relations, policies for the ACE framework can be defined. This is the case if the following proposition is corroborated;

Proposition 1:
"The behavior of a document predicts the business value of a document."

To be able to corroborate proposition 1, some expected causal relations will be tested using the dataset that is collected with the field test. These expected causal relations are based on the assumptions and findings of the authors of other data valuation methods.

According to Chen a piece of information is more valuable if it is used more heavily than others (Chen, 2005). Unfortunately, the amount of use is not logged in a Windows file system. To get information on the amount of use of a document, the respondents are asked to give an indication of the number of times they access the document a day, week, month or year. This measure is called, 'the perceived amount of access'.

Proposition 1a

"A higher perceived amount of access results in a higher document business value."

Gibson and Miller developed a 'file-aging' algorithm based on the assumption that older files are used less and therefore less valuable (Gibson & Miller, 1999). The leads to the following proposition;

Proposition 1b

"The older the document the lower the business value of the document."

The last modification time of a document refers the number of days since the document was last updated. If the documents are updated recently, people are actively working with these documents and therefore the business value of these documents is higher.

Proposition 1c

"A more recent last modification time results in a higher document business value."

Turczyk examined the characteristics of different documents to find probability distributions that can be used to determine the probability of further use. He found that the probability distribution depends on the file type of a document (L; Turczyk, et al., 2008). To test if the document business value (DBV) also depends on the file type of a document the following proposition is made;

Proposition 1d

"The file type of a document can be used to predict the business value of a document."

The propositions above are abstracted from the literature on the valuation of data. The position that a person has in the organization may also influence the DBV of the documents. The reason for this is that the type of documents that are used by people in an organization depends on the line of work they are involved in. In the Capgemini organization 'grades' are used to define the function level of the personnel. To see whether this influences the business value of the documents the following proposition is tested;

Proposition 2

> *"A higher grade of the user results in a higher business value for the documents they use."*

The propositions made above are all summarized in an empirical model. This model displays the observed variables of the documents (file age, last modification time, file type), the behavioral construct of the respondent (user grade) and the perceptional constructs of the respondent (perceived amount of access and document business value). The document business value is determined by the questions of the IVQ, see Appendix III: information value Questionnaire. The different constructs are numbered C.1 to C.6, these numbers are used in the next section to describe the constructs in detail. Figure 4.1 shows the empirical model.

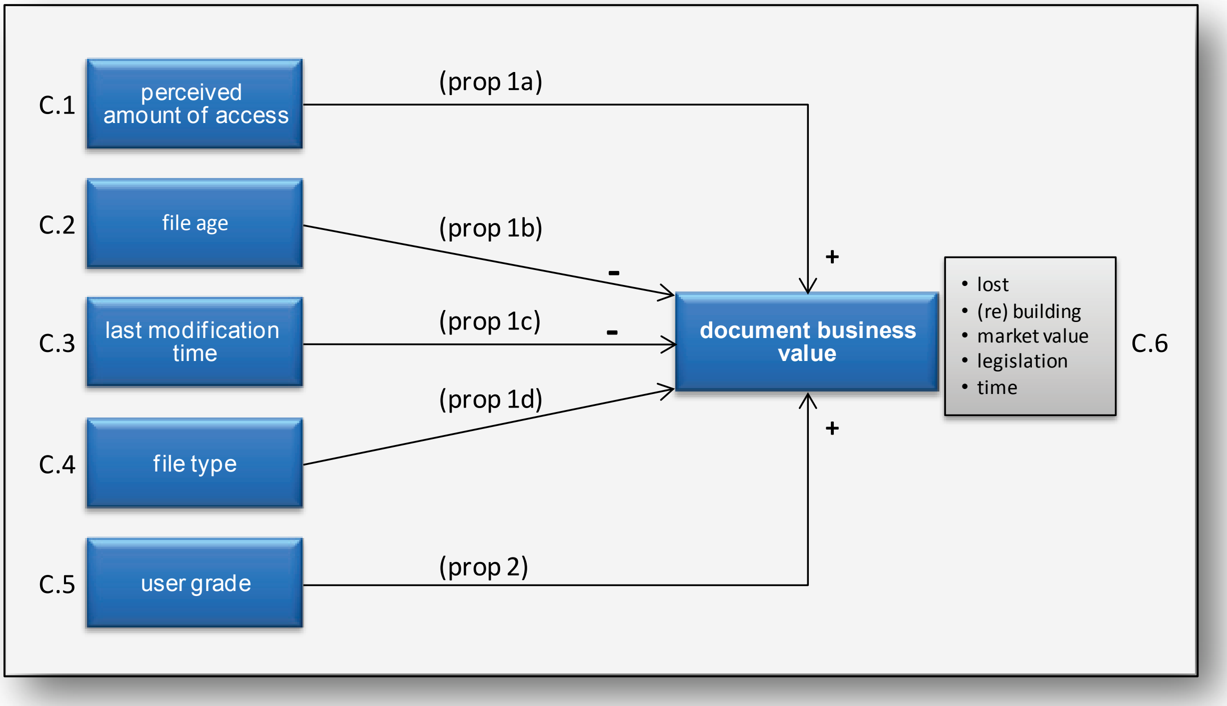


Figure 4.1: Empirical Model

4.2 DATA COLLECTION

To validate the conceptual policy specification method a field test is conducted. In the field test, a dataset is collected. The dataset is used to validate the conceptual policy specification method by testing the propositions from the previous section. In the dataset the following elements are collected;

- The metadata attributes of a document (table 4.1);
- A completed IVQ for this document;
- The grade of the respondent (table 4.2).

Metadata Attributes		
File Type (extension)	File last Access Time	File Name
File Size (KB)	File Last Modification Time	User
File Creation Time	File Path	

Table 4.1: Collected Attributes in the Metadata

The attributes in the metadata of documents stored in a Windows based file system are listed in table 4.1. The user that completes the questionnaire is asked to indicate his or her current grade at Capgemini. Capgemini uses these grades to indicate the function level of their personnel, these grades are listed in table 4.2. To increase the effectiveness of the questionnaires two measures are used:

- Each respondent is asked to complete the IVQ for at least 5 documents.
- Only documents of the following file types can be selected; .doc, .xls, .ppt and .pdf.

Grades at Capgemini		
Consultant	Managing Consultant	Vice President
Senior Consultant	Principal Consultant	Business Support (Secretary)

Table 4.2: Grades at Capgemini

The constructs C.1 to C.6 are abstracted from the data that is available in the dataset, using the metadata attributes, questions answered in the IVQ and additional questions asked to the respondent. An overview of the constructs the data their based on and the scale that is used for the different constructs is presented in table 4.3.

Constructs			
No.	**Name**	**Based on**	**Scale**
C.1	Perceived Amount of Use	Question in questionnaire, asking the respondent the indicate the amount of access per day, week, month or year	The answers are normalized to 'number of access moments per year'
C.2	File Age	'File creation date' in metadata	No. of days since creation date
C.3	Last Modification Time	'File last modification date' in metadata	No. of days since last modification
C.4	File Type	'File type' in metadata	Extension of file (.doc / .ppt / .xls / .pdf)
C.5	User Grade	Question in questionnaire	Grade at Capgemini (see table 5.2)
C.6	Document Business Value	Scores in 'information value questionnaire'	Total score of the five questions in the IVQ, ranging from 0 to 20

Table 4.3: Constructs in Detail

4.3 DESIGN AND TESTING

For the field test an electronic application is developed. To develop a application that is feasible from a security, user and technical point of view, there are three important considerations that have to be made. First, the application has to work within the possibilities that come from security restrictions. Second, in order to keep respondents motivated while completing the questionnaire, user friendliness is important. Third, it is important that the technical design of the application is of high quality. If the application crashes, a valuable response is lost. The application is tested extensively, to ensure that all these requirements are fulfilled before the application is used.

SECURITY RESTRICTIONS

The metadata attributes that have to be collected are stored in the file system of the computer along with the files. To collect these attributes, access to the file system is needed. This poses a challenge for the design of the application. Applications that are published over the web in applications using PHP (php.net, 2009), Flash (Wikipedia, 2009a) or XBAP (MSDN, 2009), run in a 'Sandbox'. The Sandbox is used to run 'untrusted' applications from third parties. The Sandbox provides a tightly controlled set of resources of the local machine. Network access, inspecting the host system or reading input from devices is not allowed (Wikipedia, 2009b). Because of these restrictions, a survey running in a Sandbox environment is unable to read the attributes from the metadata that are required for the study.

To run the application without the restrictions of the Sandbox, the application needs to run in the normal application environment of the local machine. The easiest way to achieve this is to distribute an executable file which is opened by the user. To do so, an application is developed in Windows Presentation Foundation (WPF) (Microsoft, 2009b) using Microsoft Visual Studio 2008 as the development environment (Microsoft, 2009a).

USER INTERFACE DESIGN

The application is used to collect the valuation of electronic documents and the metadata of these documents. The respondent will manually select five documents that he or she wants to value. To guide the respondent in the selection of documents, the respondent is asked to select documents which relate to the questions from the 'Information Value Questionnaire' (IVQ) (see Appendix III: information value Questionnaire for details). After selecting five different documents, the respondent can progress to the next page. On this page the IVQ is displayed for the first document. The IVQ has five multiple choice questions with five possible answers which represent a score in a range from 0 to 4. A sixth question has been added, asking the respondent to give an indication of the number of times he or she uses the information in the document.

When the IVQ has been completed for all selected documents the respondent is asked to select the current grade at Capgemini. The respondent can receive a summary of the research results by leaving an e-mail address on this page. Now the results of the survey are submitted to a webserver by pressing a 'submit' button in the application. Once the results have been sent successfully, the application will show the final page. On this page the respondent can leave comments about the questionnaire or the research in general. The questionnaire is completed and the comments, if any, are sent to the researcher. A more detailed description of the application with screenshots, can be found in Appendix IV: The Application.

TECHNOLOGICAL DESIGN

The application is developed in two parts. A respondent application and a server application. The respondent application consists of an user interface and a file reader that collects the metadata attributes of the different selected documents. The metadata is added to an instance of a document class. The answers of the different questions are added to this instance. When the questions of all documents have been answered and the user has selected the current grade, the results are sent to a web server using a webservice.

The webservice runs on a Windows Server 2003 machine which is located within the Capgemini network. Once connected to the webservice the results of the five files are transmitted. The results are saved in txt files on the server. It is important to use a txt file because it is possible that two results are transmitted to the server at the same time. And a txt file accepts two simultaneous write actions. For each of the files a large string is added to the txt file. Appendix VI: Results txt File shows a snapshot of this txt file. This txt file it is transformed to an XML file. Excel 2007 is used for data analysis.

To ensure that the results are always received by the researcher, two backups for the transmission of the survey results are built into the application. If the webservice is unavailable the application tries to send the results via a SMTP server that is available within the Capgemini network. This results in a mail message that is sent to the researcher and the tester of the survey. If this backup fails because the SMTP server cannot be found, the application prompts an error message which informs the respondent. The application automatically creates a new mail in Outlook and puts the results in the body of the message. The respondent is asked to send this mail to the researcher.

TESTING

A application that does not work is one of the biggest fears of a researcher. Once the invitation to participate in the research has been sent and the application is not working, valuable respondents are lost. To make sure that the developed questionnaire application is technically robust and fool proof a senior test consultant of Capgemini was asked to test the application. He advised to add an additional backup for the data transfer and rules to allow only 'correct' inputs in the textfields. With these modifications the application passed the tests.

To test the interpretation of the questions in IVQ and the user friendliness of the application, four people from Financial Services NL were asked to test the survey. This test delivered valuable feedback. As a result important textfields were highlighted to improve the user friendliness of the application, the amount of text used to instruct the respondent is shortened and the phrasing of questions is changed. The questions were changed because two persons in the test group thought that the questions were focused on the instance of the document, so one copy of the information that is contained in the document. While the questions focus on the information in the document in general, and not only this instance of the information.

4.4 CONDUCTING THE FIELD TEST

The finished questionnaire application is a small application packed in one executable file. No installations or registry entries are required to run the application. Distributing the application is therefore easy from a technical point of view. A low response rate is one of the biggest problems in the use of questionnaires as a research application (Cooper & Schindler, 2006). To reduce the non-response bias different measures are mentioned by Cooper&Schindler (2006). Based on their advice the following measures are applied.

The invitation send to the participants clearly stated the causes for data proliferation, the problems that are associated with it and how completing the questionnaire can help to reduce these problems. This was done to increase the perceived importance of the research topic and importance of completing the questionnaire. The required amount of time to complete the questionnaire was also clearly stated in the invitation. The supervisors of the internship are mentioned in the invitation as sponsors of the research. After the initial invitation two reminders have been sent, the first after seven days and the second after fourteen days. Two days after the initial invitation and two days after the first reminder everybody in the office was asked again to participate, simply by asking them in person.

From the invitation the questionnaire application is opened using a link. By pressing the link a webpage is opened. On this webpage instructions were given to open the actual questionnaire application. The structure of the questionnaire and the tasks of the respondent are displayed on the welcome screen of the questionnaire. In the questionnaire there is a privacy statement which informs the respondent that the files selected in the application are not opened, copied, uploaded or modified and that the application is only collecting the metadata of the documents. Because the server which hosted the questionnaire is only accessible from the Capgemini network, the questionnaire application was also made available in a compressed zip file that could be downloaded location which required no connection with the Capgemini network.

Chapter 5

5 RESULTS METHOD VALUATION

In this chapter the results of the field test are presented. First, the response rate and descriptive statistics of the field test are assessed are presented (5.1). Secondly, the reliability of the questionnaire that has been used in the field test is evaluated (5.2). Thirdly, the different propositions from section 4.1 are tested to determine if they can be corroborated (5.3). Based on these tests, the findings for the field test are presented (5.4).

5.1 DESCRIPTIVE STATISTICS

In this section descriptive statistics about the response of the field test are presented. The user grades of the respondents, the relative response rates of the user grades and the different documents used for valuation are described.

The field test is conducted in the financial service sector of Capgemini NL. Everyone in the sector received an invitation to participate in the field test. In total 654 people were invited, 77 completed the questionnaire. This results in a response rate of 12%. Figure 5.1 shows the distribution of the grades in the responses.

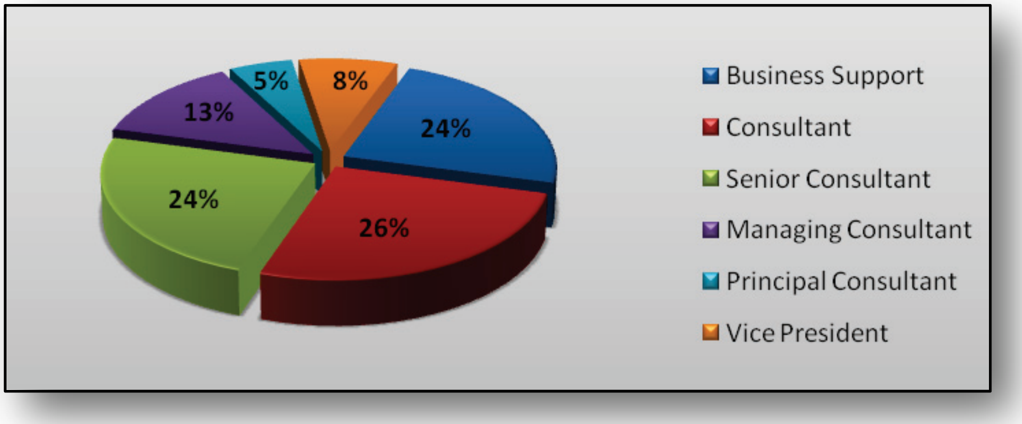


Figure 5.1: Distribution of Grades in Responses

Figure 5.2 shows the relative response rate for each of the grades. The relative response is calculated by dividing the number of people from a certain grade who completed the questionnaire, by the total number of people of this grade in the Capgemini Financial Services sector NL.

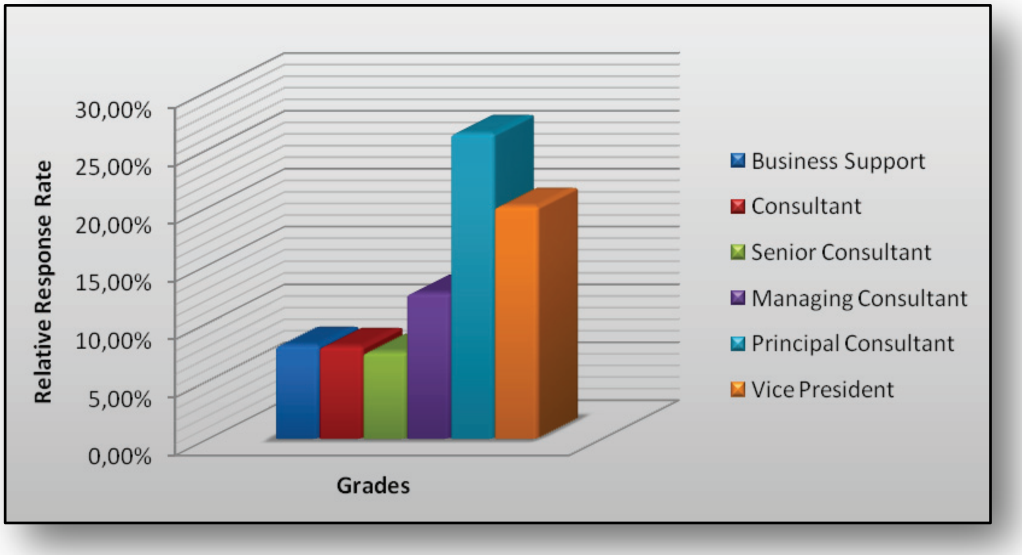


Figure 5.2: Relative Response Rate per Grade

All respondents were asked to complete the questionnaire for at least 5 different documents. In total the 77 respondents assessed the business value of 387 documents. Figure 5.3 shows the distribution of the different document types for which the questionnaire is completed.

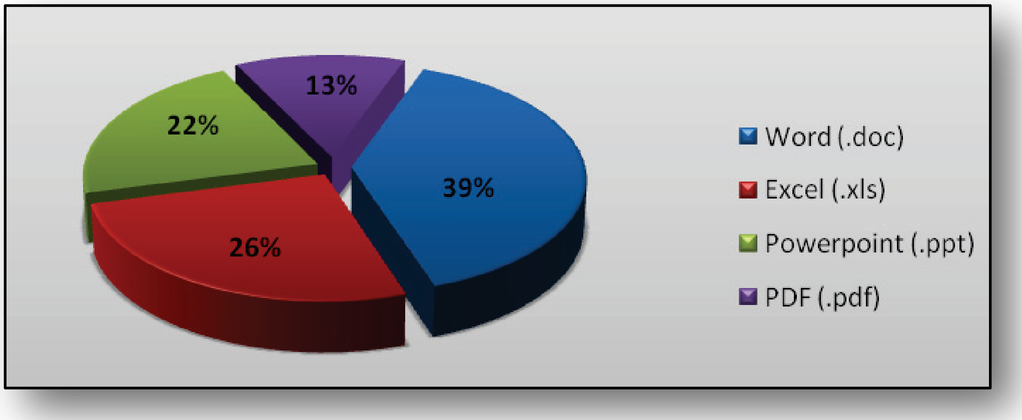


Figure 5.3: Distribution of Document Types

DATA PREPARATION

The results of the completed questionnaires are stored in a text file on the server which acted as a host for the questionnaire. Each row of data in the txt file consisted of a large string value that represents one reviewed document. The large string value has to be transformed into values for different variables. To do so a XML header and footer is added and the file is imported in Microsoft Excel 2007. In the dataset there are four variables that represent a date and time. The review date, file creation date, last access date and last modification date. By subtracting the review date from the other dates the relative age is calculated which is required to perform the calculations.

5.2 RELIABILITY ANALYSIS

The reliability of the results from the field test is assessed using a factor analysis, Cronbach's alpha and item reliability index.

5.2.1 FACTOR ANALYSIS

The IVQ is used to determine the 'document business value' (DBV) of a document. A factor analysis is performed to determine if the questions in the IVQ all load on the DBV construct. In table 5.1 a summary of outcomes for the KMO and Bartlett's test are displayed.

KMO & Bartlett's Test	
Kaiser Meyer Olkin Measure (KMO)	**0,79**
Bartlett's Test of Sphericity	
Approx Chi Square	454,9749
df	10
Sig	**0,000**

Table 5.1: KMO and Bartlett's Test

The KMO statistic greater than 0,5 is considered acceptable and v value above 0,8 is considered great (Field, 2005). The calculated value (0,79), shows that factor analysis is appropriate for the dataset. The Bartlett's test is highly significant ($p < 0.001$) this again confirms that factor analysis is appropriate (Field, 2005).

Communalities		
Item	**Initial**	**Extraction**
Lost	1,000	,689
ReBuilding	1,000	,640
MarketValue	1,000	,551
Legislation	1,000	,582
Time	1,000	,046

Table 5.2: Communalities

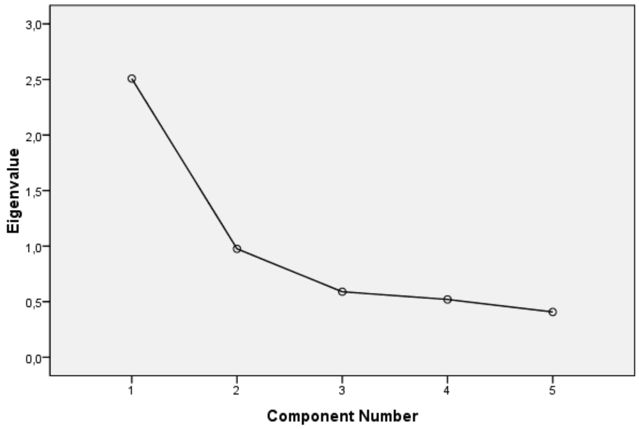


Figure 5.4: Scree plot

Table 5.2 shows the communalities before and after extraction. The average communality is lower than 0,6 (0,50). A scree plot is therefore used to determine the factors. This is allowed because the dataset is large (> 300 cases) (Field, 2005). The scree plot is displayed in figure 5.4. The scree plot shows that there is one factor with a large eigenvalue. The component matrix output of SPSS confirms this, see table 5.3.

Component Matrix	
Item	**Document Business Value**
Lost	,830
ReBuilding	,800
MarketValue	,742
Legislation	,763
~~Time~~	

Table 5.3: Component Matrix

Based on the factor analysis the results from the question in the information value questionnaire regarding the influence of time, are not used. This question does not load on the DBV factor.

5.2.2 CRONBACH’S ALPHA & ITEM RELIABILITY INDEX

To test the reliability of the IVQ as a whole, the cronbach’s alpha is also calculated. To test the reliability of the four items that are selected based on the factor analysis, an item reliability index is calculated.

A Cronbach’s alpha higher than 0,7 is considered reliable (Field, 2005). The Cronbach’s for the four factors from the factor analysis is 0,79. The DBV is therefore a reliable scale.

Item	Corrected Item-Total Correlation
Lost	,666
ReBuilding	,621
MarketValue	,553
Legislation	,574

Table 5.4: Crrected Item-Total Correlation

In Table 5.4 a summary of the item reliability index calculation is presented. In a reliable scale, all items correlate with the total score of the scale. The correlation is shown in the 'corrected item total correlation' column. If the correlation exceeds 0,3 the item is considered reliable (Field, 2005). No item has a smaller correlation than 0,5, they are therefore all reliable.

Item	Cronbach's Alpha if Item Deleted
Lost	,692
ReBuilding	,733
MarketValue	,760
Legislation	,741

Table 5.5: Cronbach's Alpha if Item Deleted

Table 5.5 shows the Cronbach's alpha if one of the items is deleted. Deleting any of the items would result in a lower Cronbach's alpha for the questionnaire, because the Cornbach's alpha for the total questionnaire is 0,79, this confirms the reliability of the items.

5.3 TESTING PROPOSITIONS

After assessing the reliability of the questionnaire and DBV factor, the propositions that have been made in chapter 5 can be tested. There are five propositions that are tested using a quantitative data analysis, see table 5.6.

Propositions	
1a	*"A higher perceived amount of access results in a higher document business value."*
1b	*"The older the document the lower the business value of the document."*
1c	*"A more recent last modification time results in a higher document business value."*
1d	*"The file type of a document can be used to predict the business value of a document."*
2	*"A higher grade of the user results in a higher business value for the documents they use."*

Table 5.6: Propositions

Linear regressions analysis is used to test propositions, 1a, 1b, 1c and 2. To test proposition 1d, one-way independent ANOVA is used. The outcome of proposition 1 is based on the outcomes of proposition 1a, 1b, 1c and 1d and is presented in 5.3.5.

5.3.1 PROPOSITION 1A

The results of the regression analysis for proposition 1a are presented in table 5.7.

Summary Regression Analysis				
ANOVA	F (1, 385) = 31,826	Sig. = 0,000		
Unstandardized Coefficients		Standardized Coefficients		
B	Std. Error	Beta	t	Sig.
0,003	0,001	0,276	5,641	0,000

Table 5.7: Summary Regression Analysis (Perceived Amount of Access – DBV)

The ANOVA summary shows that perceived amount of access of a document is a useable predictor for DBV (p < 0,001). Furthermore, the Beta found under 'Standardized Coefficients is positive. Proposition 1a is therefore corroborated. The R square shows that perceived amount of access explains 7,6% of variance in DBV.

5.3.2 PROPOSITION 1B

The results of the regression analysis for proposition 1b are presented in table 5.8.

Summary Regression Analysis				
ANOVA	F (1, 385) = 8,43	Sig. = 0,004		
Unstandardized Coefficients		Standardized Coefficients		
B	Std. Error	Beta	t	Sig.
- 0,002	0,001	-0,146	-2,903	0,004

Table 5.8: Summary Regression Analysis (Document Age – DBV)

The ANOVA summary shows that document age is a useable predictor for DBV (p = 0,004). Furthermore, the Beta found under 'Standardized Coefficients is negative. Therefore, proposition 1b is corroborated. The R square shows that the age of a document explains 2,1% of variance in DBV.

5.3.3 PROPOSITION 1C

The results of the regression analysis for proposition 1a are presented in table 5.9.

Summary Regression Analysis				
ANOVA	F (1, 385) = 9,568	Sig. = 0,002		
Unstandardized Coefficients		Standardized Coefficients		
B	Std. Error	Beta	t	Sig.
-0,002	0,001	-0,156	-3,093	0,002

Table 5.9: Summary Regression Analysis (Last Modification Time – DBV)

The ANOVA summary shows that the last modification time of a document is a useful predictor for DBV (p = 0,002). Furthermore, the Beta found under 'Standardized Coefficients is negative. Therefore, proposition 1c is corroborated. And the R square shows that the last modification time of a document explains 2,4% of variance in DBV.

5.3.4 PROPOSITION 1D

To test proposition 1d the difference in the means of the DBV for the different document types will be tested on significance using a one-way independent ANOVA test. Table 5.10 summarizes the results of the test.

Summary One-Way Independent ANOVA		
Levene Statistic	Sig. = 0,515	
ANOVA	F (3, 383) = 1,844	Sig = 0,139

Table 5.10: Summary One-Way Independent ANOVA (Document Type – DBV)

The Levene's test is not significant (p > 0,05), this means variances of the DBV of the different document types is not significantly different. This is a prerequisite for performing an ANOVA. However the F-ratio from the ANOVA indicates that the effect of the document type on the DBV is not significant (p > 0,05). Based on this finding is concluded that proposition 1d cannot be corroborated.

5.3.5 PROPOSISTION 1

Proposition 1 is the following;

> *"The behavior of a document predicts the business value of a document."*

This proposition is corroborated when there are causal relations between attributes in the metadata of documents and the DBV of documents. Based on these causal relations an administrator can define policies in the ACE framework. Based on the assumptions and findings in the literature about other valuation methods propositions 1a, 1b, 1c and 1d were defined. The propositions considered possible causal relations between document behavior and DBV. Three of these propositions are corroborated, as is presented above. Proposition 1 is therefore also corroborated.

5.3.6 PROPOSITION 2

The results of the regression analysis for proposition 2 are presented in table 5.11.

Summary Regression Analysis				
ANOVA	F (1, 385) = 6,81	Sig. = 0,009		
Unstandardized Coefficients		Standardized Coefficients		
B	Std. Error	Beta	t	Sig.
0,363	0,139	0,132	2,610	0,009

Table 5.11: Summary Regression Analysis (User Grade – DBV)

The ANOVA summary shows that the grade of a user is a useful predictor for DBV (p = 0,01). Furthermore, the Beta found under 'Standardized Coefficients' is positive. This shows that proposition 2 is corroborated. And the R square shows that grade of a user explains 1,7% of variance in DBV.

5.4 FINDINGS FIELD TEST

In section 4.1 six propositions are defined. A field test is conducted to collect a dataset that is used to determine if these propositions can be corroborated. Based on the analysis of the dataset from the field test, five of the six propositions are corroborated, see table 5.12.

Propositions		Result
1a	*"A higher perceived amount of access results in a higher document business value."*	*Corroborated*
1b	*"The older the document the lower the business value of the document."*	*Corroborated*
1c	*"A more recent last modification time results in a higher document business value."*	*Corroborated*
1d	*"The file type of a document can be used to predict the business value of a document."*	*Not Corroborated*
2	*"A higher grade of the user results in a higher business value for the documents they use."*	*Corroborated*

Table 5.12: Results of the Data Analysis

This demonstrates that it is possible to use file behavior, based on attributes in the metadata of documents, to determine the business value of these documents. This allows an administrator to specify policies in the ACE framework. And use the input of users that are working with these documents in the business. This results in user informed policies for ILM. It also makes the specification of policies for ILM less labor intensive and less complex.

The analysis in this chapter shows that the method that is developed to specify policies in the ACE framework works. There are causal relations between the behavior of documents and the business value of these documents. The goal of this research is to find a way to determine the business value of documents in a MOSS2007 environment in a practical way. The next section in this research therefore focuses on assessing the feasibility of the designed method in practice. To do so, the usefulness and practicality of the method are assessed.

Chapter 6

6 EVALUATING THE DESIGNED METHOD IN PRACTICE

In the previous chapter the results from the field test are analyzed. The results show that it is possible to determine the business value of a document using the behavior of a document. To evaluate how the designed method can contribute in real business problems, experts in the field of ECM and ILM are asked to give their opinion on the designed method. In section 6.1 the goal of this part of the research is introduced. The approach that has been used to achieve this goal is described in section 6.2. The outcomes are described in section 6.3. Based on the outcomes, the findings of this part of the research are discussed in section 6.4.

6.1 GOAL OF THIS SECTION

The goal of this section of the research is to evaluate the usefulness and practicality of the designed method. A method is considered useful if it is serviceable for an end or purpose (Merriam-Webster, 2009b). A method is considered practical if it is capable of being put to use or account (Merriam-Webster, 2010). Experts in the field of enterprise content management can determine whether the designed method is practical and useful.

6.2 APPROACH

This research is explorative and there is limited time available to do the research. Therefore, interviewing the experts is a feasible and effective way to obtain the opinion of the experts.

Three experts of Capgemini NL are interviewed. Semi structured interviews are used to collect the data. Semi structured interviews start with a few specific questions and the rest of the interview is a more natural conversation. A semi structured interview differs from a structured interview in several ways:

- Relies on developing a dialog between the interviewer and participant.
- Extracts more and greater variety of data.
- Achieves greater clarity and elaboration of answers (Cooper & Schindler, 2006).

The interviews are conducted on an individual basis and in a face-to-face setting.

The following three people are interviewed:

- Paul; a managing consultant working on an ECM project in an insurance company.
- Andy; a principal consultant with a sales function in ECM.
- Dick; a principal consultant working as an ILM specialist in government organizations.

(Because of the anonymity of the respondents, all names are replaced with fictional names.)

Section 6.1 discussed usefulness and practicality. To be useful, the method should contribute in resolving a relevant business problem. To be practical, the method should be workable in an organizational environment. Together, the three respondents provide input on the usefulness and practicality of the method.

6.3 RESULTS

In this section the results of the interviews are presented. The results are structured using the questions that are asked in the interview;

1. How can this method help you in your project(s)?
2. What do you consider to be strong points in this method?
3. What do you consider to be the weaker points in this method?
4. Can you think of a useful contribution to my method?

6.3.1 HOW CAN THIS METHOD HELP YOU IN YOUR PROJECTS?

Paul is working in a large insurance company on the integration of knowledge in one central knowledge base. *"The value dimensions which you use in the questionnaire are very important in the insurance company I am working for".* Paul is developing a portal which will be the access point of the knowledge repository. *"I can use the business value of a document to select documents that are interesting to publish on the front page".*

Andy participated in multiple projects where organizations moved their network based storage to a centralized knowledge repository. The owners of documents have to select the documents that they want to retain and the documents that are archived or deleted. These documents are normally all selected individually. *"With this questionnaire I can help business people that have to make a decision about their documents."* The behavior of documents can be used to determine the business value of a document. *"If this is used to semi-automate the migration of documents, a lot of time and effort can be saved".*

Dick worked on numerous ECM projects in government organizations. These organizations are under strict regulations for the archiving of documents. Because of these regulations, ECM in governments makes a distinction between the dynamic and static phase of an electronic document. In the dynamic phase a document is in development and people are collaborating in the document. At the end of the dynamic phase there is a transition, the document enters the static phase and is archived. *"Your questionnaire and policies can be used to determine the transition moment of a document from the dynamic to the static phase".*

6.3.2 WHAT DO YOU CONSIDER TO BE STRONG POINTS IN THIS METHOD?

All respondents are enthusiastic about the idea to use a questionnaire to determine the business value of electronic documents. Paul: *"The questionnaire is very useful to make people aware of the value of documents they are working with"* and *"The questionnaire allows people to easily rank their documents from low value to high value"*. Dick says: *"The participation of owners and users of documents in the valuation process makes this method easy to use and also more saleable in practice".*

Andy worked on multiple storage migration projects where business people were asked to make a selection of documents that they wanted to preserve. He therefore especially likes the idea of ranking: *"Because the questions are suitable and clear, the questionnaire is very useful in practice"* and *"This allows users to make a much more detailed decision when they have to choose which documents they want to keep and which are deleted".* In the storage migration projects there are vast amounts of data that need to be migrated. Andy: *"What I like about your method is that it determines the business value of a document based on available metadata and the causal relations that are found using the questionnaire. I often run into situations where we have to review all documents, document by document".*

In a lot of projects, records managers or IT departments have to determine which data is kept available in systems and which data is archived. Dick: *"The questionnaire allows the owners and users of the data to determine the business value of the data, not IT personnel, this is the way it should be".*

Determining the importance of documents is often a manual and labor intensive task. Using policies to determine the business value of documents is therefore very beneficial. Paul: *"Developing policies is a very time consuming and complex activity, your policy specification method looks promising and is definitely easy to implement".* Andy: *"Automation based on metadata is very useful in practice; I think that this will work".*

The designed method can be used to close the gap between business and archivists. Dick: *"One of the major issues is that business people cannot specify policies for ECM. Therefore archivists specify the policies. Archivists find it very difficult to become part of the business and to become involved. This method can help to close the gap between archivists and the business".*

Organizations make use of enriched metadata, for instance to indicate the processes and departments in which the document is used. The metadata is added using metadata enrichment procedures. Metadata enrichment is a time consuming and complex activity. Paul: *"The questionnaire can be used to guide people in the enrichment of metadata. This distributes the workload of metadata enrichment over more people".*

6.3.3 WHAT DO YOU CONSIDER TO BE THE WEAKER POINTS IN THIS METHOD?

Paul is concerned that the outcomes of the questionnaire might depend heavily on the role and position a person: *"A trader for instance, knows exactly which documents contain valuable information for trading. A legal assistant uses documents which contain valuable information for the compliancy with rules and regulations. A single document used by both persons will be valued completely different by the trader and the legal assistant".*

Andy thinks that it is important which questions are used in the questionnaire: *"Only if the right set of questions is used, the questionnaire is a very useful tool in selecting important documents".*

Paul and Andy are both concerned about the reliability of the method. Paul: *"How many people have to complete the questionnaire before reliable results can be produced with the method?"* Andy: *"How is the reliability of the method ascertained? This is an important prerequisite for the success in a business environment".*

Metadata used in organizations contains much more attributes than the file system metadata that is used in the field test. Contextual metadata is also used. Contextual metadata is metadata that can contain information about the contents of documents, versioning, departments that use the document and business processes in which the document is used and developed. Contextual metadata is very important in ECM, Andy says: *"Your method should include the use of contextual metadata before it is applicable in practice".*

In ECM, document workflows are used to structure the processes in which documents are created and used. Part of workflows is the versioning of documents. Paul is concerned about this: *"Versioning is very important in organizations, with your method there is a risk that an obsolete version of a document is rated higher than its successor. This has to be prevented".*

6.3.4 CAN YOU THINK OF A USEFUL CONTRIBUTION TO MY METHOD?

Andy: *"How can this method be used to determine the behavior of documents? For example, a railway maintenance company is confronted each year with problems that are related to the season. If we can predict which documents will become more valuable, we can make this document more accessible in knowledge portals etc. This would be a great contribution".*

The knowledge portal that Paul is developing will be used by users to search for information in the knowledge base. Paul: *"In an organization we can assign a certain profile to a user based on his position, type of work and department. We use the outcomes of questionnaires of users with the same profile to indicate the documents that can be of high value to the user. We can then use this information about documents to optimize the search results of this user".*

Before a method can be used in practice its reliability has to be proven. Andy sees a possibility to do this in his data migration project where a storage environment is replaced. *"Let the user evaluate a test batch of documents with the method. Based on the causal relations found in the metadata of the test batch, calculate the business value of other documents of the user and provide him with the outcomes of the calculation. The user can then decide whether the outcomes are correct".*

6.4 FINDINGS OF THE EVALUATION IN PRACTICE

This section of the research is conducted to evaluate the usefulness and practicality of the designed method. Useful is defined as; '*serviceable for an end or purpose'* (Merriam-Webster, 2009b), practical is defined as; *'capable of being put to use or account'* (Merriam-Webster, 2010). Before the usefulness and practicality of the designed method is evaluated, the concerns and contributions that are mentioned by the experts are discussed.

CONCERNS

The experts mentioned some concerns about the dependency and reliability of the results, the questions used in the questionnaire, the use of contextual metadata and the versioning of documents. These concerns are discussed below.

The experts worry that the valuation depends heavily on the role and position of the person that completes the questionnaire. The field test shows that persons with a higher grade at Capgemini assign a higher value to the documents they use than people with a lower grade. The value of a document therefore indeed depends on the person that uses the document. The causal relations between document behavior (document age, last modification time, etc.) and business value however do not depend on the user of the document. These causal relations can therefore be used to determine the business value of a document, regardless of the user of a document.

One of the experts thinks that the questionnaire only works if the right set of questions is used. The questions in the information value questionnaire can be changed depending on the organization and type of documents that require valuation. Furthermore, it is also possible assign ‘weights’ to the questions if not all dimensions of value are equally important in a specific situation.

The reliability of the method is a concern for the experts. The reliability of the method is higher when more questionnaires are completed. The causal relations between document behavior and business value become more significant and therefore more useful to determine the business value of the documents. At this moment the document behavior can only be seen as a ‘proxy’ for business value. It allows system administrators and users to make a distinction between documents of high and low value. Calculating a precise number for the business value of a document is not possible at this time, more work and research is needed before a reliable calculation can be made.

Organizations use contextual metadata in ECM activities. According to one of the experts it is critical to include contextual metadata in the method. In the field test of the method, it was technically not feasible to include any contextual metadata. If the contextual metadata can be abstracted from the documents, constructs available in the metadata can be easily included in the analysis. In the same way as file age and file type are abstracted from the file system metadata. The designed method is also capable of including any contextual metadata.

Versioning is very important in organizations. An expert indicated that it is important that an obsolete version may be assigned a higher business value than its successor. This can be ensured by including the version number as one of the constructs in valuation. The version number can be abstracted from the file name or contextual metadata.

CONTRIBUTIONS TO THE METHOD

The experts also mentioned some interesting contributions to the method. These contributions are discussed next.

When issues need to be solved, people start looking for information that is required to solve the issue at hand. The frequency of these issues can for instance depend on the season in a year. If it is predictable which documents become more valuable in a certain season, these documents can be made more accessible. This helps the people that are looking for the documents. The ‘Probability of Further Use’ method could be used as a starting point to predict the future use of documents (L. Turczyk, 2009; L;

Turczyk, et al., 2008; L; Turczyk, et al., 2007). The method designed in this research is not suitable to predict the future behavior of documents.

The value assigned to a document depends on the role and the position of a person in the organization. It can therefore be useful to develop 'profiles' of persons. The profile can be for instance used to sort the search results of a person, placing the documents with the highest value for the person on top of the search results. The profile can also be used for personalized information on intranet WebPages such as knowledge portals. Documents that are assigned with a high business value by users with the same profile can be presented on the front page of the knowledge portal.

USEFULNESS AND PRACTICALITY

The interviews show that the designed method can be used in many different situations:

- To select valuable documents to publish on a knowledge portal.
- To help business people that have to decide which documents should be migrated to a new storage environment and which documents should be deleted or archived.
- To reduce the workload that is associated with the development of storage policies.
- To distribute the work of reviewing documents in data migration projects.
- To determine the moment that a document makes the transition from the dynamic phase to the static phase.
- To reduce the gap between the work of archivists and the business environment.
- It allows users to make more precise decisions than 'keep or delete'.

The designed method can contribute in resolving different relevant business problems and is therefore definitely useful. The interviews also show that the questionnaire can be easily applied in projects it is straight forward and easy to understand. The questionnaire is therefore also practical to use.

The questionnaire is used to find causal relations between document behavior and the business value of documents. These causal relations are then used to determine the business value of other documents. Before this method can be used in practice the reliability of the method has to be tested. One of the experts proposes to test the reliability of the method by performing a valuation of documents based document behavior and present the users with the outcome of the valuation. The users can then give feedback on the valuation which gives an indication of the reliability of the method. This part of the method has practical potential however more work is needed before it can really be used in practice.

In the previous chapters of the thesis the third, fifth and sixth design science research guidelines are applied. The textboxes below summarize the application of these three guidelines.

Design Science Guidelines: Design Evaluation

Third Guideline:	**Design Evaluation**
Description:	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods (Hevner, et al., 2004).

Application in Research

The designed policy specification method is evaluated by conducting a field test and by collecting the opinion of experts about the method. The field test showed that it is possible to determine the business value of a document using the behavior of a document. The evaluation with experts shows that the designed method is useful and practical to apply.

Design Science Guidelines: Research Rigor

Fifth Guideline:	**Research Rigor**
Description:	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact (Hevner, et al., 2004).

Application in Research

All steps in this research use rigorous methods. The literature has been reviewed using a transparent and structured literature review. Based on these outcomes the method has been designed. For the quantitative evaluation of the method a field test is conducted and the outcomes are analyzed using well known statistical techniques. The qualitative evaluation is done with semi structured interviews, using a clearly described approach.

Design Science Guidelines: Design as a Search Process

Sixth Guideline:	**Design as a Search Process**
Description:	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment (Hevner, et al., 2004).

Application in Research

The method that is designed in this research has been based on available relevant literature and expert advice. For the evaluation of the method the entire financial services NL sector was asked to participate. Experts in the field of ECM and ILM from Capgemini also participated in the evaluation. All in all, the resources that were available for the research are exerted to deliver the results.

Chapter 7

7 RESEARCH FINDINGS

In this chapter the findings of this research are reported and discussed. In section 7.1 the central research question is answered and the related conclusions are presented. In section 7.2 the limitations of the research are discussed. In the next sections the contributions of this research to both the academic field (7.3) and practice (7.4) are presented. This is followed by a section with recommendations for further research and the application of the designed method in practice (7.5). The thesis ends with some final thoughts on ILM and the proliferation of data in section 7.6.

7.1 CONCLUSIONS

This research is conducted to find a practical way to determine the business value of electronic documents. The goal was to answer the following research question:

> ***Research Question***
> *"How can the business value of electronic documents be determined in a practical way?"*

This research shows that the business value of electronic documents can be determined using the behavior of the documents in combination with an 'information value questionnaire' that is completed by the users of the documents. The remainder of this section describes how this is concluded.

This research started with a structured literature review to search for suitable data valuation methods. Based on the outcomes of the structured literature review the ACE framework was selected.

The ACE framework uses policies to determine the business value to documents. Three inputs are required for a policy in the ACE framework; a business value, a set of observable attributes and a specific value or value ranges for these attributes. The literature review showed that there are a number of difficulties related to the specification of policies for the ACE framework:

- It is difficult for business people to specify the policies in the ACE framework.

- Specifying the policies is a complex and time intensive activity.

The next phase of the research therefore aimed to develop a policy specification method. The method searches for causal relations between behavior of documents and the business value of documents. To look for these causal relations the business value of a number of documents is determined by hand. The found causal relations are then used to determine the business value of documents for which the business value is unknown.

Business people who own and/or use documents indicate the business value of these documents by answering five questions. A questionnaire called 'Information Value Questionnaire' (IVQ) is used. This questionnaire is shown in Appendix III. The five questions in the questionnaire are combined to a total score between 0 and 20.

The behavior of documents is based on the values of the attributes that are available in the metadata of documents. The attributes used in this research are listed in table 4.1.

A field test conducted at Capgemini showed that there are four causal relations between the behavior of documents and their business value (see also figure 7.1):

- The business value is higher when:
  - The perceived amount of use is higher.
  - The last modification time is more recent.
- The business value lowers as the document becomes older.
- A higher grade of the user that completes the questionnaire results in a higher business value.

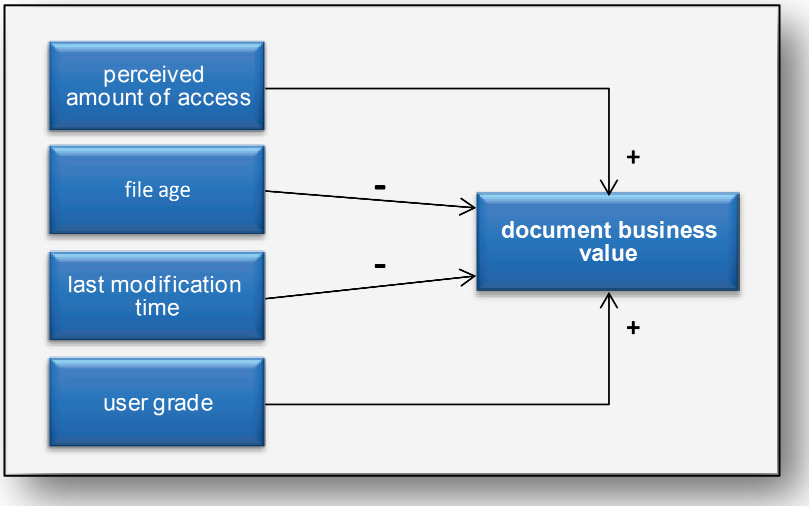


Figure 7.1: Research Findings

Interviews with experts in the ECM and ILM field are conducted to evaluate the usefulness and practicality of the designed method. The experts indicate that there are many possible applications in practice for the designed method. They support the use of a questionnaire because it allows business people to easily quantify the business value of a document. Furthermore, the questionnaire makes people aware the differences in business value of documents. They find the use of the questionnaire useful and practical.

Using the causal relations between file behavior and business value to specify policies and to determine the business value of documents for which the business value is unknown, sounds promising to the experts. The experts however also indicate that this approach needs more testing before if it is reliable enough to use it in practice.

7.2 LIMITATIONS

In this research an explorative study is conducted to develop and validate a method that can be used to determine the business value of documents. Because of the limited time and other resources that were available for the research some immolations had to be made.

First of all, the field test that is conducted in this research had a limited number of respondents. In total 77 people participated in the field test. Together they completed the information value questionnaire for 387 documents. Although this is a reasonable result for the limited time that was available, more respondents will improve the reliability of the outcomes of the data analysis.

Secondly, in the field test people were asked to select at least five different documents for which they wanted to complete the questions. Except for limitation on file types there was no control over the documents that could be selected. The contents of the documents were not available to the researcher. The results of the questionnaire can therefore not be verified.

Thirdly, because of technical limitations in conducting the field test, it was not possible to measure the amount of access of a document. The respondents were therefore asked to indicate the access frequency of a document. This measure is therefore somewhat unreliable, further research in which the amount of access is accurately measured is required to produce better verifiable results.

Fourthly, due to the limited time available for this research, the usefulness and practicality of the designed method is evaluated using interviews with experts. The reliability and thoroughness of the evaluation can be improved. For instance, by applying the designed method in practice and measuring its effects and researching the application more insights about the usefulness and practicality can be obtained.

Fifthly, metadata in organizations is much more extensive than just file system metadata. Because this is an explorative study the research focused on the metadata that is easy to abstract for all electronic documents. This does not mean that other metadata such as contextual metadata will not influence the results of the research.

Sixthly, the method designed in this research provides insights in the causal relations between document behavior and business value. These causal relations can be very usable in practice. At this moment, the method is not yet suitable to determine a precise number for the business value of a document. The document behavior can however be used as a proxy for business value. Making definite decisions based on the business value determined by the method in its current form is not advisable at this time. More research is required before the method is suitable to support 'hard' decisions.

7.3 CONTRIBUTIONS FOR THE ACADEMIC FIELD

Defining policies for ILM purposes is a complex and time consuming activity. This research is the first known research that combines the input of business people combined with statistical techniques to specify policies and determine the business value of documents. It is demonstrated how the behavior of documents can be used to determine the business value of documents.

The structured literature review in this research compares more data valuation methods than other known publications to date. The literature review can be used to quickly gain insight in the different approaches towards data valuation. Furthermore the assessment criteria that have been used to rank the different data valuation methods are a combination of different assessment criteria used in previous publications. This set of assessment criteria provides a framework that can be used to assess the suitability of data valuation method for ILM purposes.

The field test that is conducted in this research showed the file type of a document has no significant causal relation with the business value of a document. File type is therefore not a usable attribute to specify policies or to determine the business value of documents.

The field test also showed that is a strong causal relation between the position of a user of a document and the business value of this document. The ACE Framework is improved by including the position of the user of a document. Other data valuation methods can also be improved by including the position of the user of a document.

7.4 CONTRIBUTIONS FOR PRACTICE

The insights of this research can help practitioners in the field of ILM when developing policies. By using the information value questionnaire they can retrieve the input they need from the business people who are working with the documents or data that they have to manage. Defining policies is currently a very time consuming and complex task. By incorporating the input of the business, more suitable policies can be developed that are less pushed by the IT department and more pulled from the business. This allows practitioners in the ILM field to make the business involved in their work, because as mentioned before, the problem of data proliferation is not a technical problem, it is an organizational problem.

The questionnaire helps practitioners in the ILM field to move towards a business oriented approach. Already at Capgemini during the execution of the field test, people became more aware of different values of documents. They started discussions about the amounts of invaluable data on their own laptops and the different knowledge bases that are used in the organization. With the questionnaire, the business people are stimulated to create a more critical view towards all the documents they use and store. This awareness can be one of the first steps in reducing some of the causes of data proliferation.

7.5 RECOMMENDATIONS

The outcomes of this research will hopefully inspire others to do more research in this interesting field. In this section some recommendations on further research and the possible use of the outcomes of this research in practice are discussed.

This research shows how the business value of documents can be determined using the behavior of documents. What has not been done so far is to implement this method in an enterprise content management system such as MOSS2007. For further research it is recommended to do so, to see how the method can be further improved. As mentioned in chapter one, measuring the business value of documents is the first and critical step towards a successful ILM implementation. More research into the use of file behavior is needed to be able to reduce the proliferation of data.

The only metadata used in the field test of this research is file system metadata. As was mentioned by the ECM and ILM experts, much more enriched metadata is used in organizations. An example of enriched metadata is contextual metadata. Contextual metadata can be used to link documents to departments and business processes in the organization. Researching the use of contextual metadata to look for more causal relations between document behavior and business value of documents can produce interesting additional insights. These insights can then be used to improve existing data valuation methods.

The effectiveness, usefulness and practicality of all the different data valuation methods described in the literature review have never been researched in practice. The evaluation of the method with the experts showed that valuation of documents can be used for many purposes besides ILM. An interesting research topic would be to evaluate the effectiveness, usefulness and practicality of different data valuation methods on the business problems that are mentioned by the ECM and ILM experts:

- What documents are interesting to present on the front page of a knowledge base?
- How can the data valuation methods support users when they have to decide what documents they want to; copy to a new storage environment, what documents they want to archive and what documents they want to delete?
- How can data valuation methods be used to determine when a document transits from the dynamic phase to the static phase?

The questionnaire helps business people to gain insight in the value of documents. It shows the different dimensions of value and the differences of business value between documents. This awareness is required if organizations want to start reducing the proliferation of data. It is therefore recommended to start using the questionnaire in data migration projects.

To improve the method an interesting experiment can be done. Ask the users of the documents on a network drive to evaluate a selection of documents on the drive using the questionnaire. Next, determine the business value of all the other documents with the causal relations between document behavior and business value. Make all documents below a certain threshold business value unavailable to the users and allow the users to indicate if they miss any documents. Then experiment to find out what the optimal threshold value is for which a minimal number of users miss documents and the amount of saved disk space is maximized.

Based on the data from the field test a number of causal relations between document behavior and business value are found. There can however be much more causal relations. Is there for instance also a relation between the document creator and business value? The evaluation of the usefulness and practicality of the designed method designed in this research is done by interviewing ECM and ILM experts. An interesting research topic would be to evaluate the usefulness and practicality from a user’s perspective. The outcomes of this research can be used to further improve the method.

Based on the outcomes of the literature review in which different data valuation methods are compared using assessment criteria, the ACE framework was selected to use as a data valuation method because it fulfills all assessment criteria. This however does not necessarily mean that combination of the ACE framework with another data valuation method can improve the effectiveness of the method. An interesting research topic can be to see what combinations of data valuation are possible and to evaluate the effectiveness of these combinations in different situations in practice.

In this section the seventh design science guideline is followed. Details are described in the textbox below.

Design Science Guidelines: Communication of the Research

Seventh Guideline:	**Communication of the Research**
Description:	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences (Hevner, et al., 2004).

Application in Research

In the final chapter of the thesis the contributions and recommendations of the research are presented. There are contributions given for the academic field and contributions for practitioners. The recommendations presented in this chapter can be used by researchers to do more interesting research on this topic or by practitioners to improve the way ILM is supported by MOSS2007 and to try to reduce the proliferation of data. The managerial issues related to data proliferation have been discussed in chapter one. This explains the relevance of this research for organizations in general.

7.6 FINAL THOUGHTS

ILM in its current form is not the Holy Grail that stops data proliferation. The effects of data proliferation are already seriously influencing business. People have a natural tendency to overflow themselves with information. They therefore spent far too much time looking for information and, more worryingly, the quality of decision making is reduced. The costs and risks that are related to data proliferation are therefore becoming intolerable and immediate action is required. Good problem solving aims to take away the causes of a problem instead of fighting the consequences. ILM is currently still too much a technology pushed concept. Reducing the problem of data proliferation starts with creation of awareness in the business about the causes and the effects of data proliferation. The method that is developed in this research may provide a first step in an approach for ILM that reduces the gap between technology and business. However, much more work can and has to be done. As long as the gap between business and technology in ILM exists, data proliferation and its related problem are here to stay.

8 REFERENCES

AIS (2009). MIS Journal Rankings. *Association for Information Systems* Retrieved 2 march, 2009, from http://ais.affiniscape.com/displaycommon.cfm?an=1&subarticlenbr=432

Bhagwan, R., Douglis, F., Hildrum, K., Kephart, J. O., & Walsh, W. E. (2005). *Time Varying Management of Data Storage.* Paper presented at the Workshop on Hot Topics in System Dependability, Yokohama, 222-232.

Brocke, J. v., Simons, A., & Schenk, B. (2008). *Transforming Design Science Research into Practical Application: Experiences from Two ECM Teaching Cases*. Paper presented at the 19th Australasian Conference on Information Systems.1049 - 1058.

Capgemini (2009). Who We Are Retrieved 2 march, 2009, from http://www.capgemini.com/about/

Chen, Y. (2005). *Information Valuation for Information Lifecycle Management.* Paper presented at the Autonomic Computing, 2005. ICAC 2005. Proceedings. Second International Conference on, 135-146.

Cooper, D., & Schindler, P. (2006). *Business research methods* (ninth ed.). New York: McGraw-Hill Education.

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management, 20*(1), 17-28.

EMC (2003). Data Classification: The Cornerstone for Successful Information Lifecycle Management. *EMC Whitepaper*. Retrieved from http://www.emc.com/collateral/hardware/white-papers/c1059-data-class-cornerstone-successful-ilm-wp.pdf

Eppler, M. J., & Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society: An International Journal, 20*(5), 325 - 344.

Field, A. P. (2005). *Discovering Statistics Using SPSS* (2nd ed.). London: Sage.

Gibson, T., & Miller, E. (1999). *An Improved Long-Term File-Usage Prediction Algorithm.* Paper presented at the Annual International Conference on Computer Measurement and Performance (CMG '99), Reno, NV, 639-648.

Glazer, R. (1993). Measuring the Value of Information: The Information-Intensive Organization. *IBM Systems Journal, 32*(1), 99-110.

Godfrey, J., Hodgson, A., Holmes, S., & Kirsch, L. J. (1997). *Financial Accounting Theory* (3rd ed.). New York: John Wiley and Sons.

Govil, J., Kaur, N., Kaur, H., & Govil, J. (2008). *Data/Information Lifecycle Management: A Solution for Taming Data Beast.* Paper presented at the ITNG 2008. Fifth International Conference on Information Technology: New Generations, 2008, 1226-1227.

Haeusser, B., Osuna, A., Bosman, C., Jahn, D., & Tarella, G. J. (2007). ILM Library: Information Lifecycle Management Best Practices Guide, IBM Redbooks

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *Management Information Systems Quarterly, 28*(1), 75-106.

IBM, G. T. S. (2006). The Toxic Terabyte: *How data-dumping threatens business efficiency*. London

Jin, H., Xiong, M., & Wu, S. (2008). *Information Value Evaluation Model for ILM.* Paper presented at the Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD '08. Ninth ACIS International Conference on, 543-548.

Kaiser, M., Smolnik, S., & Riempp, G. (2008). *Konzeption eines Information-Lifecycle-Management-Frameworks im Dokumenten-Management-Kontext.* Paper presented at the Multikonferenz Wirtschaftsinformatik 2008, Berlin, 483-494.

Lynman, P., & Varian, H. (2003). How Much Information? *School of Information Management and Systems* Retrieved 11 February, 2009, from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Matthesius, M., & Stelzer, D. (2008). *Analyse und Vergleich von Konzepten zur automatisierten Informationsbewertung im Information Lifecycle Management.* Paper presented at the Multikonferenz Wirtschaftsinformatik, München, Germany, 471-481.

Merriam-Webster (2009a). Proliferation. Retrieved Oktober 7, 2009: from: http://www.merriam-webster.com/dictionary/proliferate

Merriam-Webster (2009b). Useful. Retrieved December 6, 2009: from: http://www.merriam-webster.com/dictionary/useful

Merriam-Webster (2010). Practical. Retrieved January 9, 2010: from: http://www.merriam-webster.com/dictionary/practical

Mesnier, M., Thereska, E., Ganger, G. R., & Ellard, D. (2004). *File Classification in Self-* Storage Systems*. Paper presented at the Proceedings of the First International Conference on Autonomic Computing.44-51.

Microsoft (2009a). Microsoft Visual Studio 2008 - Your development happy place Retrieved 10 september, 2009, from http://www.microsoft.com/visualstudio/en-gb/default.mspx

Microsoft (2009b). The Official Microsoft WPF and Windows Forms Site Retrieved 10 september, 2009, from http://windowsclient.net/

Microsoft, C. (2007). *Microsoft Office SharePoint Server 2007 Evaluation Guide*: Microsoft Corporation.

Middleton, R., & Smith, H. (2002). Data Retention Policies After Enron. *Computer Law & Security Report, 18*(5), 333-337.

Moody, D., & Walsh, P. (1999). *Measuring The Value Of Information: An Asset Valuation Approach*. Paper presented at the Seventh European Conference on Information Systems.361-373.

Moore, C., & Karel, R. (2008). *The Five Top Challenges Information And Knowledge Managers Must Master In 2008*. Cambridge: Forrester Research.

MSDN (2009). Windows Presentation Foundation XAML Browser Applications Overview. *.NET Framework Developer Center* Retrieved 10 september, 2009, from http://msdn.microsoft.com/en-us/library/aa970060.aspx

Munkvold, B. E., Päivärinta, T., Hodne, A. K., & Stangeland, E. (2006). Contemporary Issues of Enterprise Content Management. *Scandinavian Journal of Information Systems, 18*(2), 69-91.

Nordheim, S., & Paivarinta, T. (2006). Implementing enterprise content management: from evolution through strategy to contradictions out-of-the-box. [Article]. *European Journal of Information Systems, 15*(6), 648-662.

Ohta, K., Dai, K., Kobayashi, T., Taguchi, R., & Yokota, H. (2006). *Treatment of Rules in Individual Metadata of Flexible Contents Management.* Paper presented at the Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on, 77-82.

Peterson, M., & Pierre St., E. (2004). Information Lifecycle Management Roadmap. *Data Management Forum*. Retrieved from http://www.snia.org/forums/dmf/programs/ilmi/ilm_docs/

php.net (2009). PHP: Hypertext Preprocessor Retrieved 10 september, 2009, from http://www.php.net/

Reiner, D., Press, G., Lenaghan, M., Barta, D., & Urmston, R. (2004). *Information Lifecycle Management: The EMC Perspective.* Paper presented at the 20th International Conference on Data Engineering 10-14.

Sajko, M., Rabuzin, K., & Baca, M. (2006). How to calculate information value for effective security risk assessment. *Journal of Information and Organizational Sciences, 30*(2), 263-278.

Scott, J., Globe, A., & Schiffner, K. (2004). Jungles and gardens: the evolution of knowledge management at J.D. Edwards. *MIS Quarterly Executive, 3*(1), 37-52.

Shah, G., Voruganti, K., Shivam, P., & Alvarez, M. (2006). ACE: Classification for Information Lifecycle Management. *Computer Science IBM Research Report, RJ10372*, (A0602-0044).

Short, J. (2006). *ILM Survey: What Storage, IT and Records Managers Say*. San Diego: ISIC UCSD Research Report.

Smith, H. A., & McKeen, J. D. (2003). Developments in Practice VIII: Enterprise Content Management. *The Communications of the Association for Information Systems, 11*(41), 647 - 659.

Strange, S. (1992). *Analysis of Long-Term UNIX File Access Patterns for Application to Automatic File Migration Strategies*. Berkely, California, USA: University of California.

Tallon, P. P., & Scannell, R. (2007). Information Lifecycle Management. *Communications of the ACM, 50*(11), 65-70.

Tanaka, T., Ushijima, K., Ueda, R., Naitoh, I., Aizono, T., & Komoda, N. (2005). *Proposal and evaluation of policy description for information lifecycle management.* Paper presented at the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, 261-267.

Turczyk, L. (2009). *Information Lifecycle Management - Eine Methode zur Wertzuweisung von Dateien.* Technischen Universität Darmstadt Darmstad.

Turczyk, L., Frei, C., Liebau, N., & Steinmetz, R. (2008). *Eine Methode zur Wertzuweisung von Dateien in ILM.* Paper presented at the Multikonferenz Wirtschaftsinformatik, München, Germany, Turczyk, L., Groepl, M., Liebau, N., & Steinmetz, R. (2007). *A Method for File Valuation in Information Lifecycle Management.* Paper presented at the 13th Americas Conference on Information Systems, Keystone, Colorado, 1122-1133.

Verma, A., Pease, D., Sharma, U., Kaplan, M., Rubas, J., Jain, R., et al. (2005). *An Architecture for Lifecycle Management in Very Large File Systems.* Paper presented at the 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies 160-168.

Wikipedia (2009a). Adobe Flash. *Wikipedia, the free encyclopedia* Retrieved 10 september, 2009, from http://en.wikipedia.org/w/index.php?title=Adobe_Flash&oldid=315344500

Wikipedia (2009b). Sandbox (computer security). *Wikipedia, the free encyclopedia* Retrieved 10 september, 2009, from http://en.wikipedia.org/w/index.php?title=Sandbox_(computer_security)&oldid=313675730

Wrozek, B. (2001). Electronic Data Retention Policy. *SANS GIAC - GSEC Security Essentials*

Zadok, E., J. Osborn, A. Shater, C. Wright, K. Muniswamy-Reddy, J. Nieh (2004). *Reducing Storage Management Costs via Informed User-Based Policies.* Paper presented at the IEEE Conference on Mass Storage Systems and Technologies, Maryland, 101-105.

9 APPENDICES

Overview of appendices

APPENDIX I: CAPGEMINI

Capgemini has its headquarters in Paris France. Capgemini operates in more than 36 countries. Currently there are more than 88,000 people working for Capgemini in Europe, North America, and the Asia Pacific region (Capgemini, 2009).

The structure of the Capgemini organization in the Netherlands is presented in figure 1.

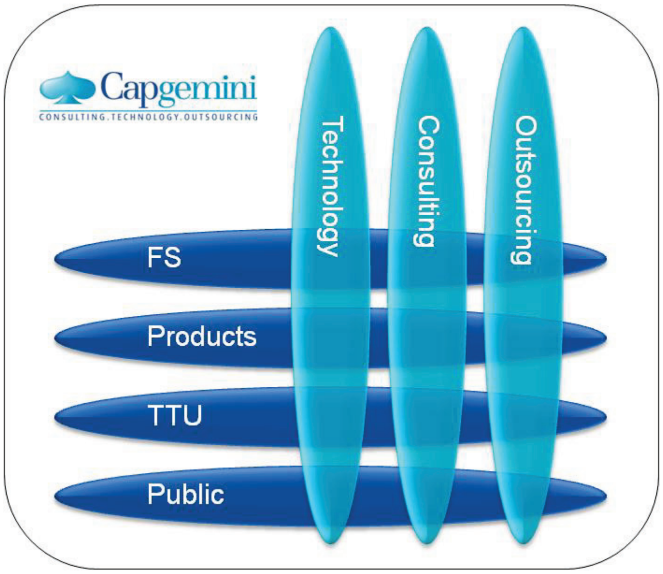


Figure I.1: Structure Capgemini NL

Management and support roles aside, the employees of Capgemini NL are grouped around three disciplines; technology, consulting and outsourcing.

The disciplines of Capgemini are focused around four sectors; Financial Services (FS), Products, Transport Telecommunication and Utilities (TTU) and Public.

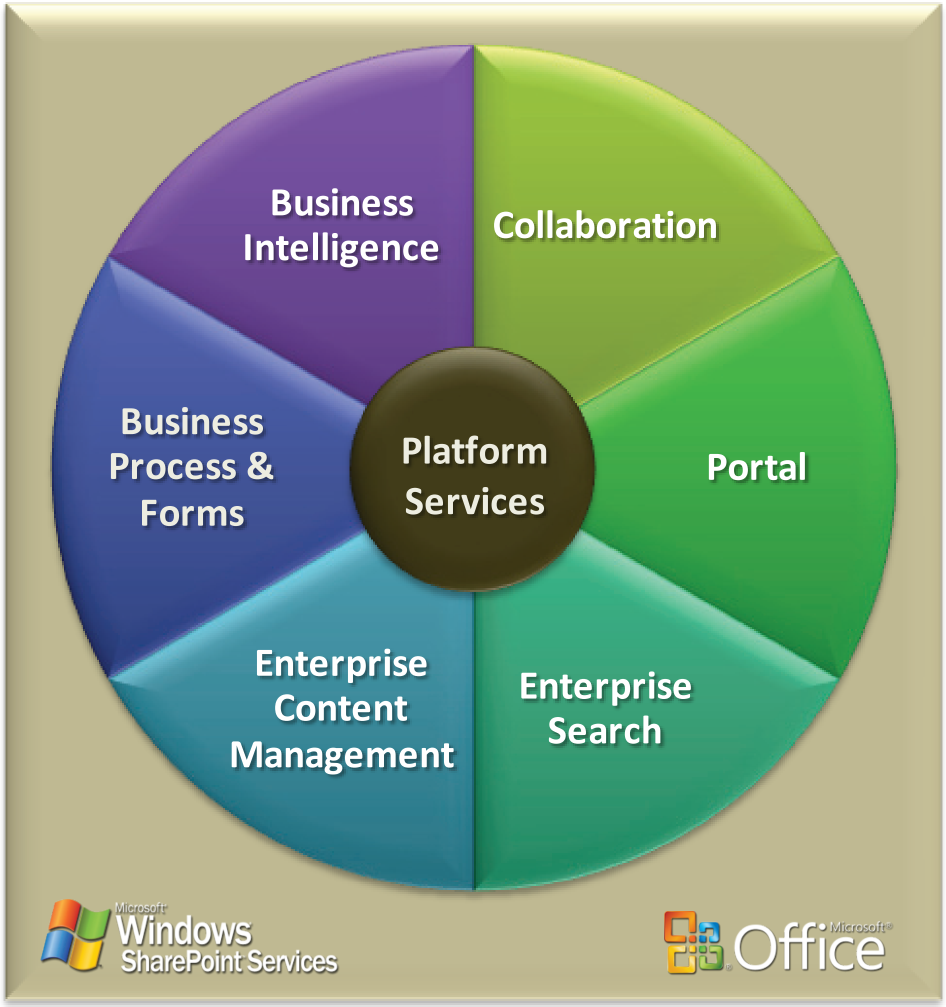
This research is executed within technology discipline of the financial services sector.

APPENDIX II: MICROSOFT OFFICE SHAREPOINT SERVER 2007

According to Microsoft, Microsoft Office SharePoint Server 2007 helps organizations gain better control and insight over their content, streamline their business processes, and access and share information. In addition, it gives IT professionals the tools they need for server administration along with application extensibility and interoperability.

MOSS2007 provides a single, integrated location where employees can efficiently find organizational resources, access corporate knowledge, and leverage business insight to make better-informed decisions.

When considering MOSS2007 for enterprise Web solutions, there are six major feature areas to explore, as represented in the following figure (C. Microsoft, 2007):



The feature areas of MOSS2007 are as follows:

1. **Collaboration** The enabling technologies that allow teams to work together effectively, providing intuitive, flexible, and secure mechanisms for sharing information through the use of wikis and blogs, collaborating on and publishing documents, maintaining task lists, conducting surveys, developing and maintaining site templates customized for specific business uses, and implementing workflows.
2. **Portal** The facilities that provide the capabilities to personalize the user experience of an enterprise Web site, to target content to various audiences based on sets of rules, to automatically facilitate intuitive navigation through the Web site while tailoring the navigation to the individual rights of the user, to deliver comprehensive site content management and structural facilities, and more.
3. **Enterprise Search** The critical ability to quickly and easily locate relevant content distributed across a wide range of sites, document libraries, business application data repositories, and other sources, including files shares, various Web sites, Microsoft Exchange public folders, and Lotus Notes Databases — and to find the appropriate people who can help answer questions or be involved in projects.
4. **Enterprise Content Management** The facilities for the creation, publication, and management of content, regardless of whether that content exists in discrete documents or is published as Web pages. Content management scenarios include document management, records management, and Web content management.
5. **Business Process and Forms** The ability to rapidly and effectively implement forms-based business processes, from design to publication to user access, by using standard Web browsers or a rich client application such as Microsoft Office InfoPath 2007. Also includes the ability to connect with structured systems such as databases and line-of-business applications, and the ability to access that information in a number of ways.
6. **Business Intelligence** The ability to deliver information critical to business objectives through a wide range of mechanisms, from server-based spreadsheets accessing business data in real time and performing sophisticated analyses to the presentation of key performance indicators (KPIs) through enterprise Web sites.

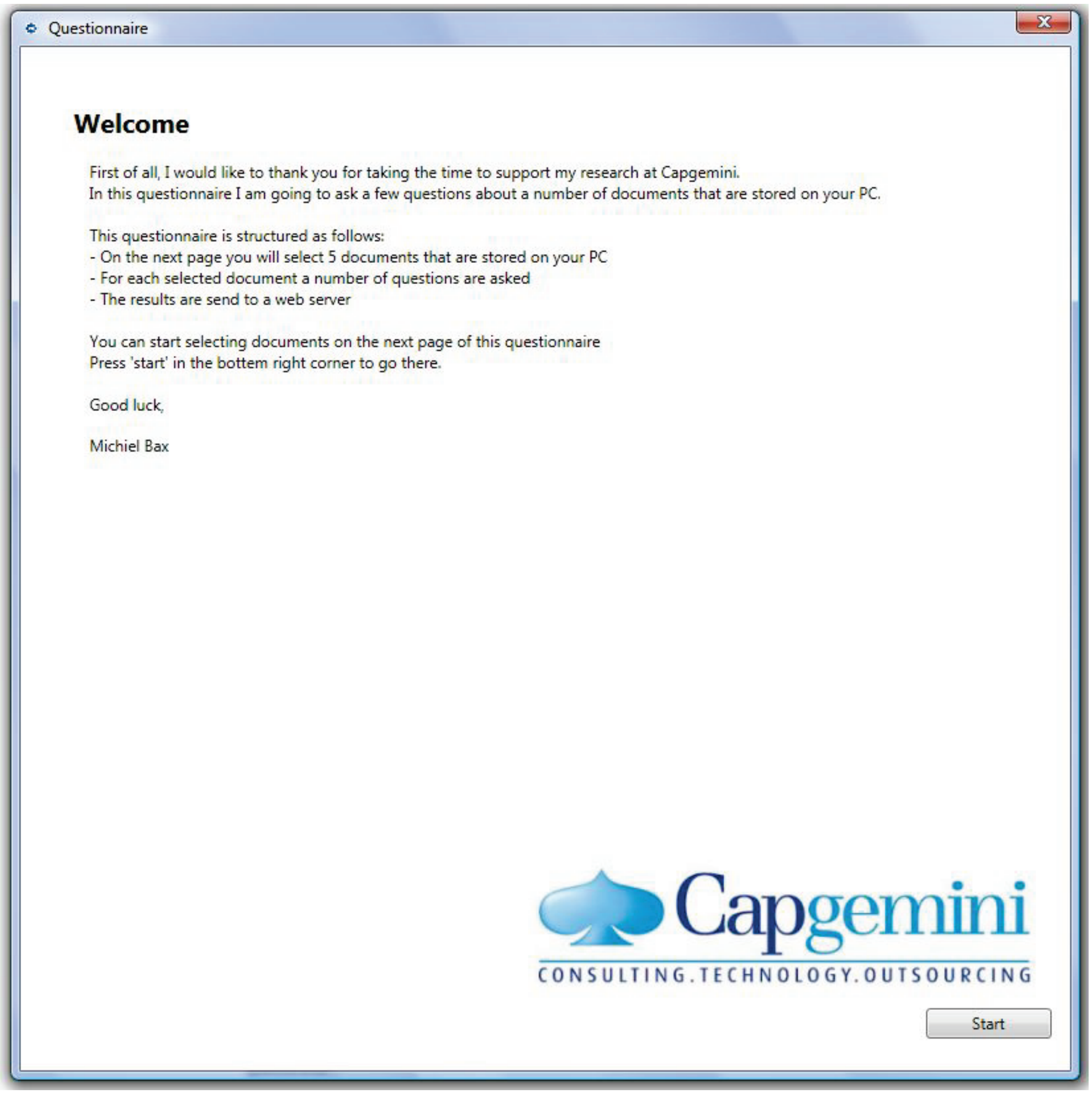
APPENDIX III: INFORMATION VALUE QUESTIONNAIRE

Dimensions	Questions	Scores
Lost	**What happens if we do not have this information anymore?**	
	Nothing Special	0
	Some processes are late, but not essentially	1
	Its imperfection is noticeable, but replaceable	2
	New unnecessary costs appear without this information	3
	Bigger halt and wrong decisions are threatening – new urgent production is necessary	4
(Re)building	**Cost of replacing information or production of a new copy?**	
	Negligibly small	0
	Cost exist but they are low	1
	Higher costs are incurred	2
	Cost is hardly tolerable	3
	Intolerably high costs	4
Market Value	**What happens if the competitor has the same information?**	
	Nothing	0
	Competitor has all available unimportant information about our company	1
	Competitor has insight in our business processes	2
	Competitor can reach the company	3
	Competitor gets a competitive advantage	4
Legislation	**Is there any obligation for keeping the information and are there consequences if the information is lost?**	
	There are no obligations	0
	It is necessary to keep the information for a brief period	1
	The company has to keep the information but there are no consequences	2
	Keeping the information is obligatory and the company can meet sanctions	3
	Keeping the information is obligatory and sanctions are strict	4
Time	**How fast does the information value fall in the course of time?**	
	Very quickly (1-3months)	0
	Quickly (6 months)	1
	After 1 year	2
	After a few years	3
	Does not fall at all	4

Figure III.1: questionnaire for the assessment of information value (Sajko, et al., 2006)

APPENDIX IV: THE QUESTIONNAIRE APPLICATION

A questionnaire has been used to generate a dataset. This dataset contains the metadata attributes of documents and the values of these attributes. For each of the documents the respondent was asked to fill out the Information Value Questionnaire (IVQ) that can be found in Appendix B. A sixth question was added to the questionnaire to ask the respondent to give an indication of the number of times the information in the document is used. The questionnaire application is developed in the Microsoft Windows Presentation Foundation program language using Microsoft Visual Studio 2008. The questionnaire is distributed as an executable file hosted in a server in the Capgemini network. More details about the design an application of the questionnaire can be found in section 4.3. In this appendix screenshots of the questionnaire can be found as it was presented to the respondents.



Welcome

First of all, I would like to thank you for taking the time to support my research at Capgemini.
In this questionnaire I am going to ask a few questions about a number of documents that are stored on your PC.

This questionnaire is structured as follows:
- On the next page you will select 5 documents that are stored on your PC
- For each selected document a number of questions are asked
- The results are send to a web server

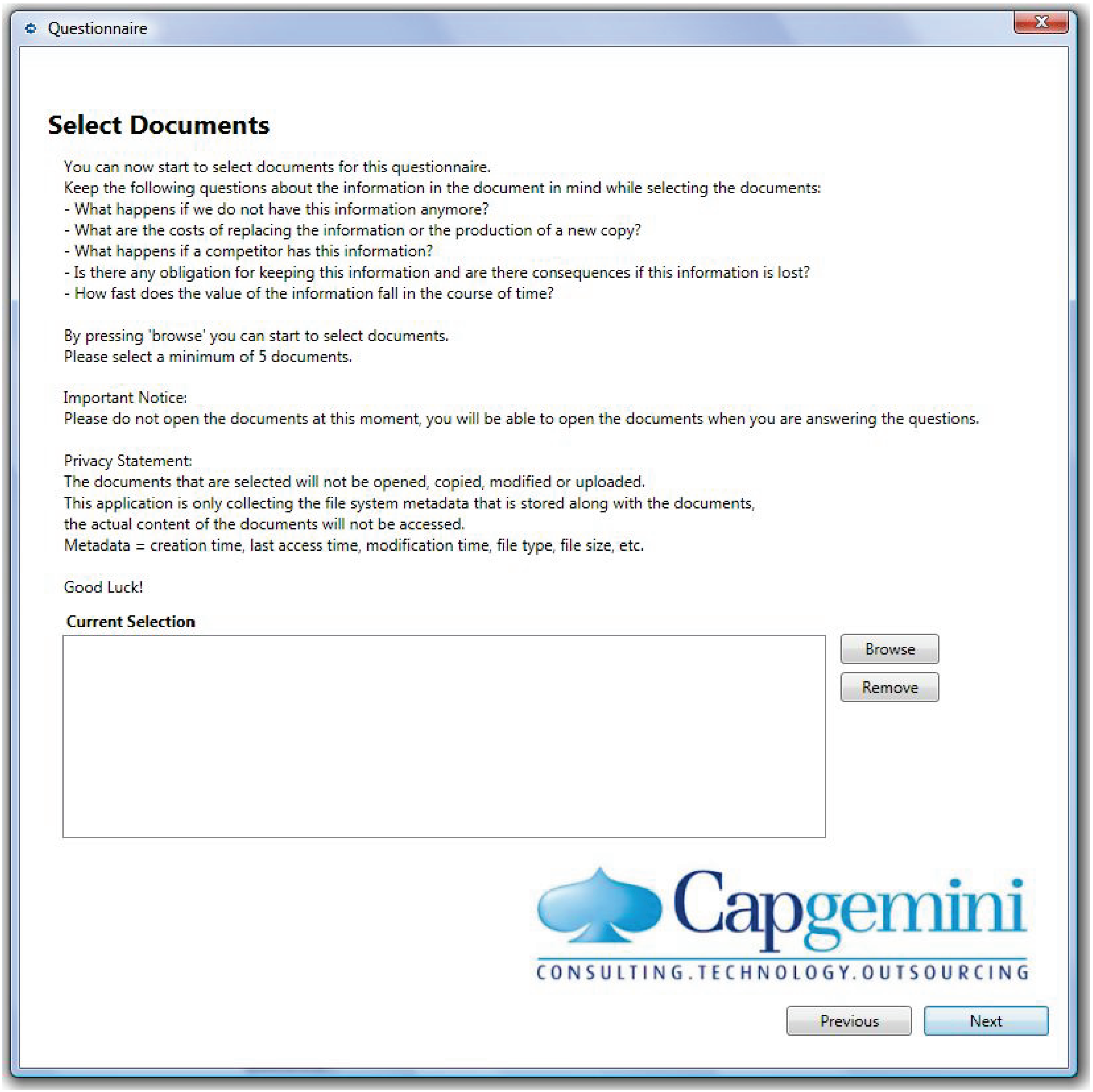
You can start selecting documents on the next page of this questionnaire
Press 'start' in the bottom right corner to go there.

Good luck,

Michiel Bax

Figure IV.2: Welcome Page of the Survey

On the second page of the questionnaire the respondent is asked to select five documents on his or her computer. To guide the respondent in selection process, the five questions of the IVQ are presented here as well. The respondent is asked to select documents which contain information that is related to these questions.



Select Documents

You can now start to select documents for this questionnaire.
Keep the following questions about the information in the document in mind while selecting the documents:
- What happens if we do not have this information anymore?
- What are the costs of replacing the information or the production of a new copy?
- What happens if a competitor has this information?
- Is there any obligation for keeping this information and are there consequences if this information is lost?
- How fast does the value of the information fall in the course of time?

By pressing 'browse' you can start to select documents.
Please select a minimum of 5 documents.

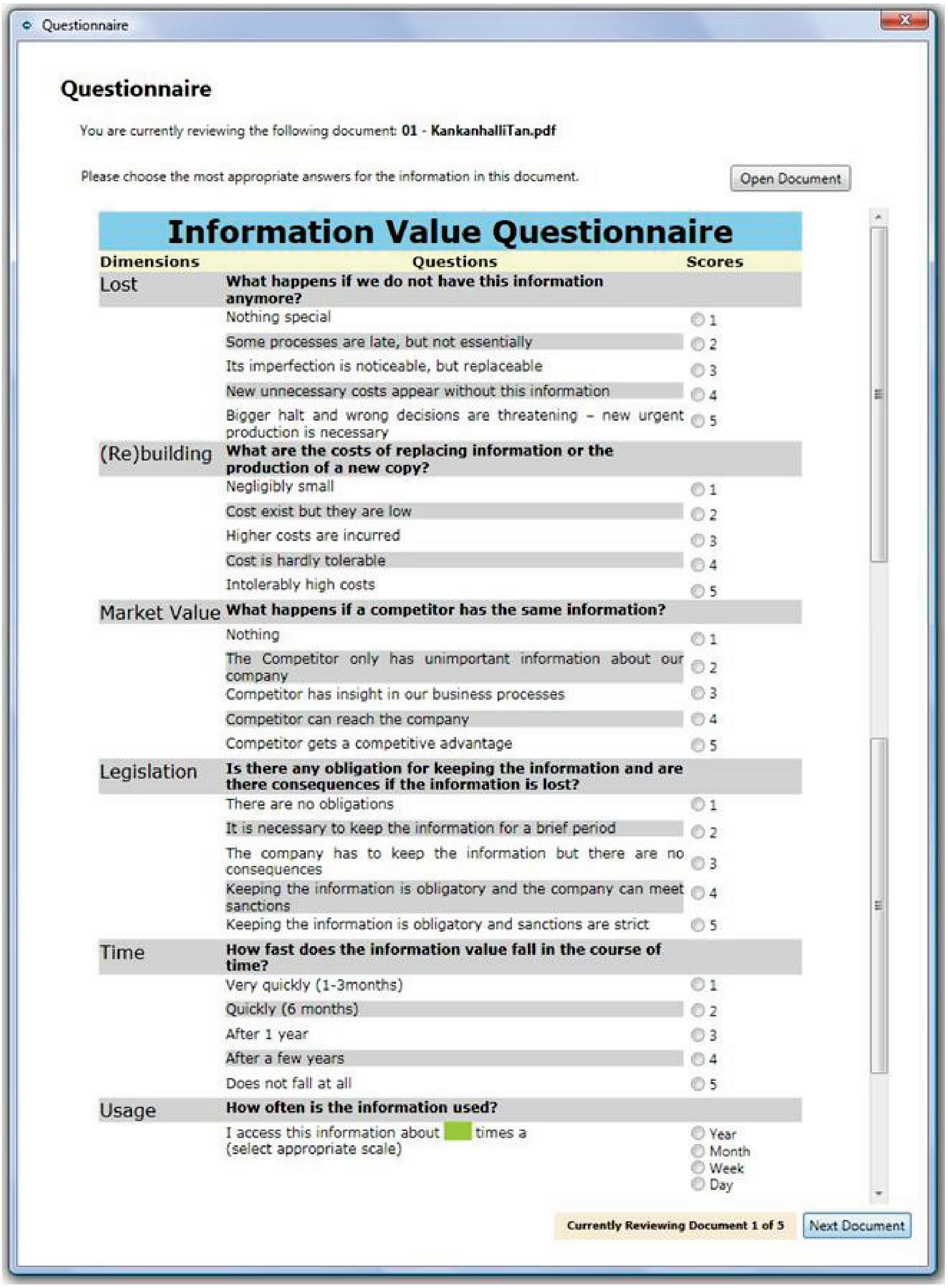
Important Notice:
Please do not open the documents at this moment, you will be able to open the documents when you are answering the questions.

Privacy Statement:
The documents that are selected will not be opened, copied, modified or uploaded.
This application is only collecting the file system metadata that is stored along with the documents,
the actual content of the documents will not be accessed.
Metadata = creation time, last access time, modification time, file type, file size, etc.

Good Luck!

Figure IV.3: Selecting the Documents

When the documents are selected the respondent is asked to fill out the IVQ for each of the documents. To help the respondent it is possible to view the document directly from the survey. A counter displaying the progress is also present.



Questionnaire

You are currently reviewing the following document: **01 - KankanhalliTan.pdf**

Please choose the most appropriate answers for the information in this document.

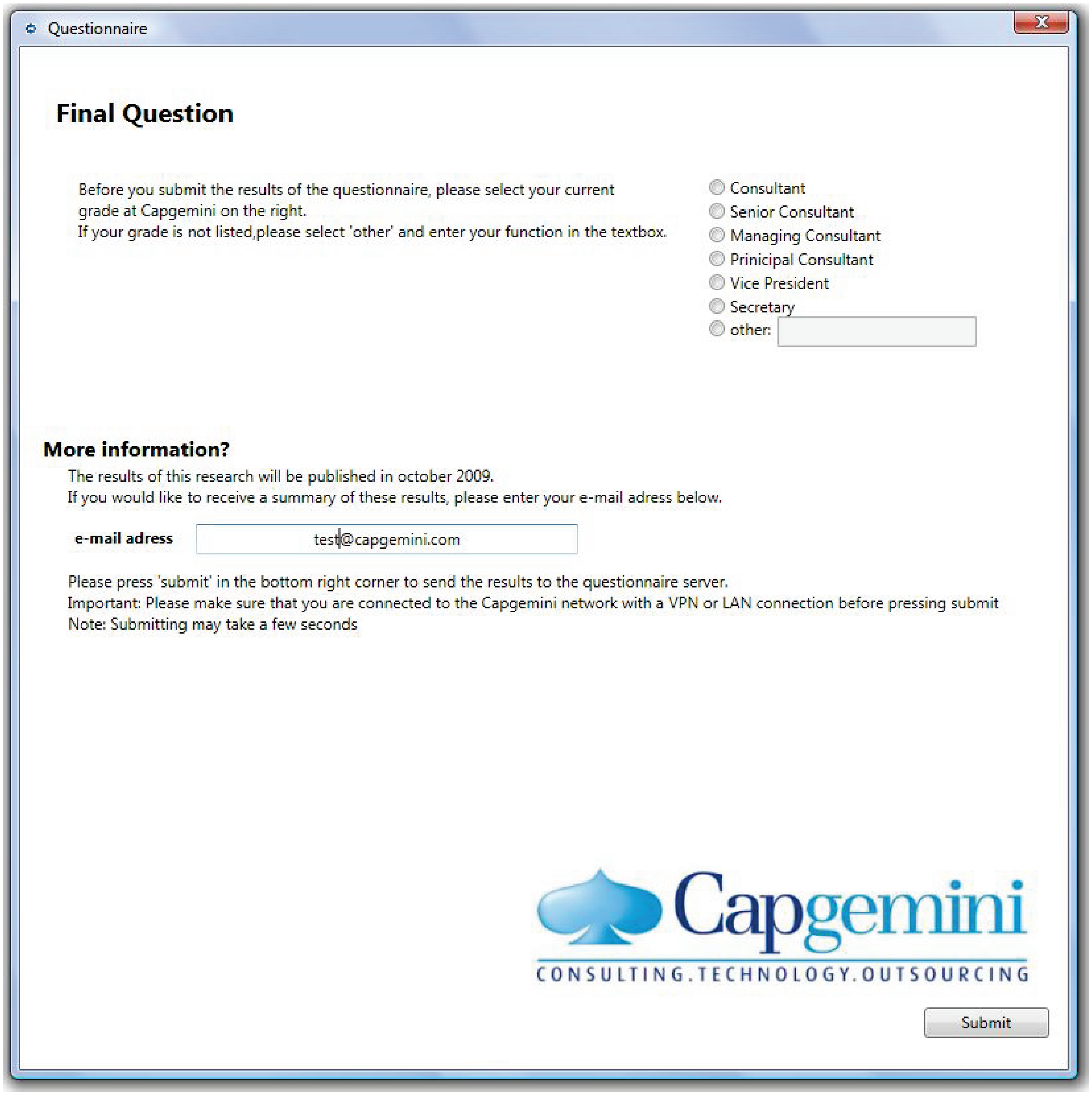
Information Value Questionnaire

Dimensions	Questions	Scores
Lost	**What happens if we do not have this information anymore?**	
	Nothing special	○ 1
	Some processes are late, but not essentially	○ 2
	Its imperfection is noticeable, but replaceable	○ 3
	New unnecessary costs appear without this information	○ 4
	Bigger halt and wrong decisions are threatening – new urgent production is necessary	○ 5
(Re)building	**What are the costs of replacing information or the production of a new copy?**	
	Negligibly small	○ 1
	Cost exist but they are low	○ 2
	Higher costs are incurred	○ 3
	Cost is hardly tolerable	○ 4
	Intolerably high costs	○ 5
Market Value	**What happens if a competitor has the same information?**	
	Nothing	○ 1
	The Competitor only has unimportant information about our company	○ 2
	Competitor has insight in our business processes	○ 3
	Competitor can reach the company	○ 4
	Competitor gets a competitive advantage	○ 5
Legislation	**Is there any obligation for keeping the information and are there consequences if the information is lost?**	
	There are no obligations	○ 1
	It is necessary to keep the information for a brief period	○ 2
	The company has to keep the information but there are no consequences	○ 3
	Keeping the information is obligatory and the company can meet sanctions	○ 4
	Keeping the information is obligatory and sanctions are strict	○ 5
Time	**How fast does the information value fall in the course of time?**	
	Very quickly (1-3months)	○ 1
	Quickly (6 months)	○ 2
	After 1 year	○ 3
	After a few years	○ 4
	Does not fall at all	○ 5
Usage	**How often is the information used?**	
	I access this information about 🟩 times a (select appropriate scale)	○ Year ○ Month ○ Week ○ Day

Currently Reviewing Document 1 of 5

Figure IV.4: The Information Value Questionnaire in the Survey

When the IVQ is completed for each of the five selected documents the respondent is asked to select his or her current grade in the Capgemini organization. If the respondent wants to be informed about the conclusions of the study an e-mail address can be submitted. Now the respondent presses 'submit', and the results of the survey are sent to the server.



Final Question

Before you submit the results of the questionnaire, please select your current grade at Capgemini on the right.
If your grade is not listed,please select 'other' and enter your function in the textbox.

○ Consultant
○ Senior Consultant
○ Managing Consultant
○ Prinicipal Consultant
○ Vice President
○ Secretary
○ other:

More information?

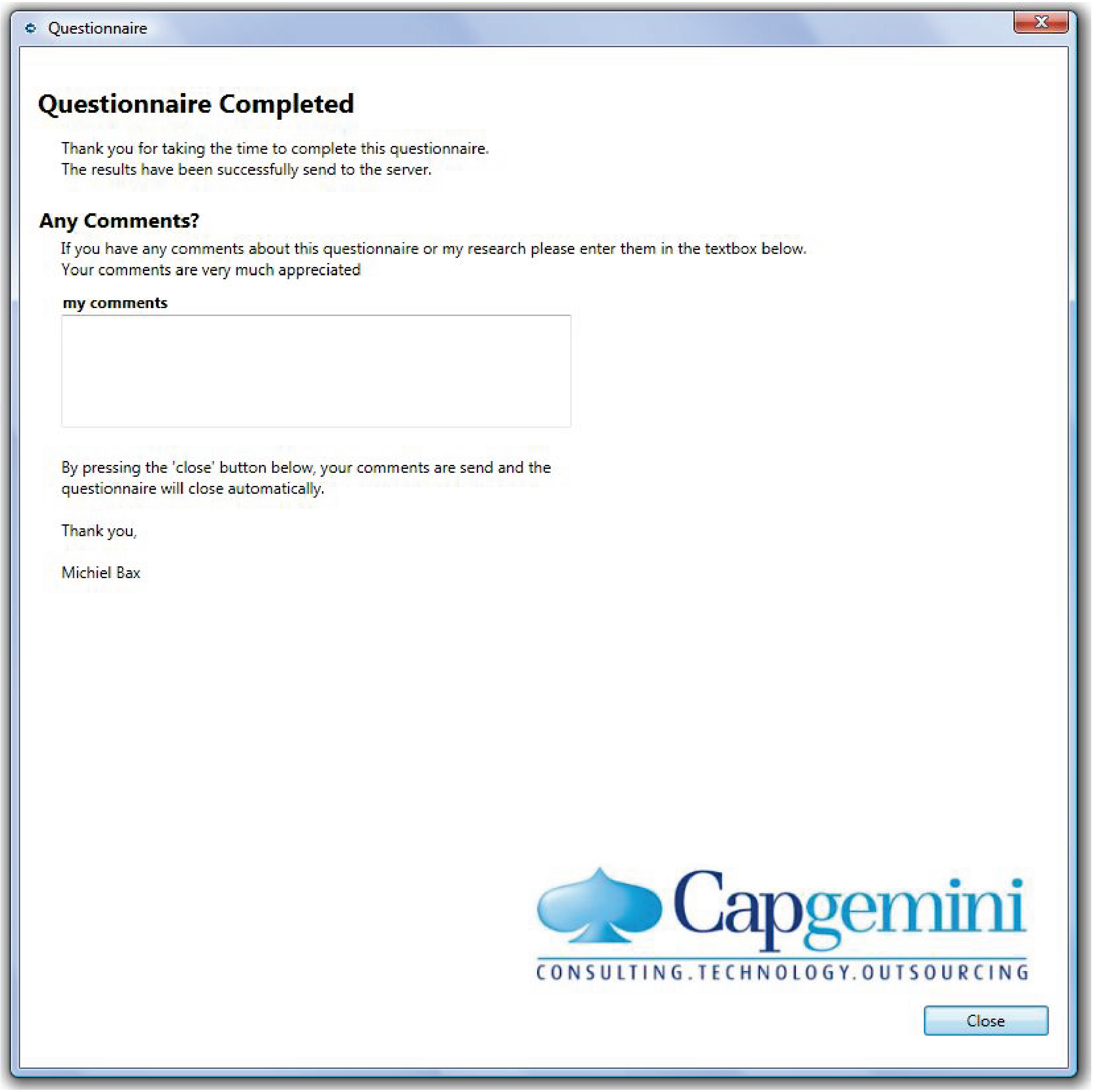
The results of this research will be published in october 2009.
If you would like to receive a summary of these results, please enter your e-mail adress below.

e-mail adress: test@capgemini.com

Please press 'submit' in the bottom right corner to send the results to the questionnaire server.
Important: Please make sure that you are connected to the Capgemini network with a VPN or LAN connection before pressing submit
Note: Submitting may take a few seconds

Figure IV.5: Final Question

The questionnaire is now completed. If the respondent has any comments, these can be entered on the last page of the questionnaire. When the questionnaire application is closed, the comments are sent to the researcher by mail.



Questionnaire Completed

Thank you for taking the time to complete this questionnaire.
The results have been successfully send to the server.

Any Comments?

If you have any comments about this questionnaire or my research please enter them in the textbox below.
Your comments are very much appreciated

my comments

By pressing the 'close' button below, your comments are send and the questionnaire will close automatically.

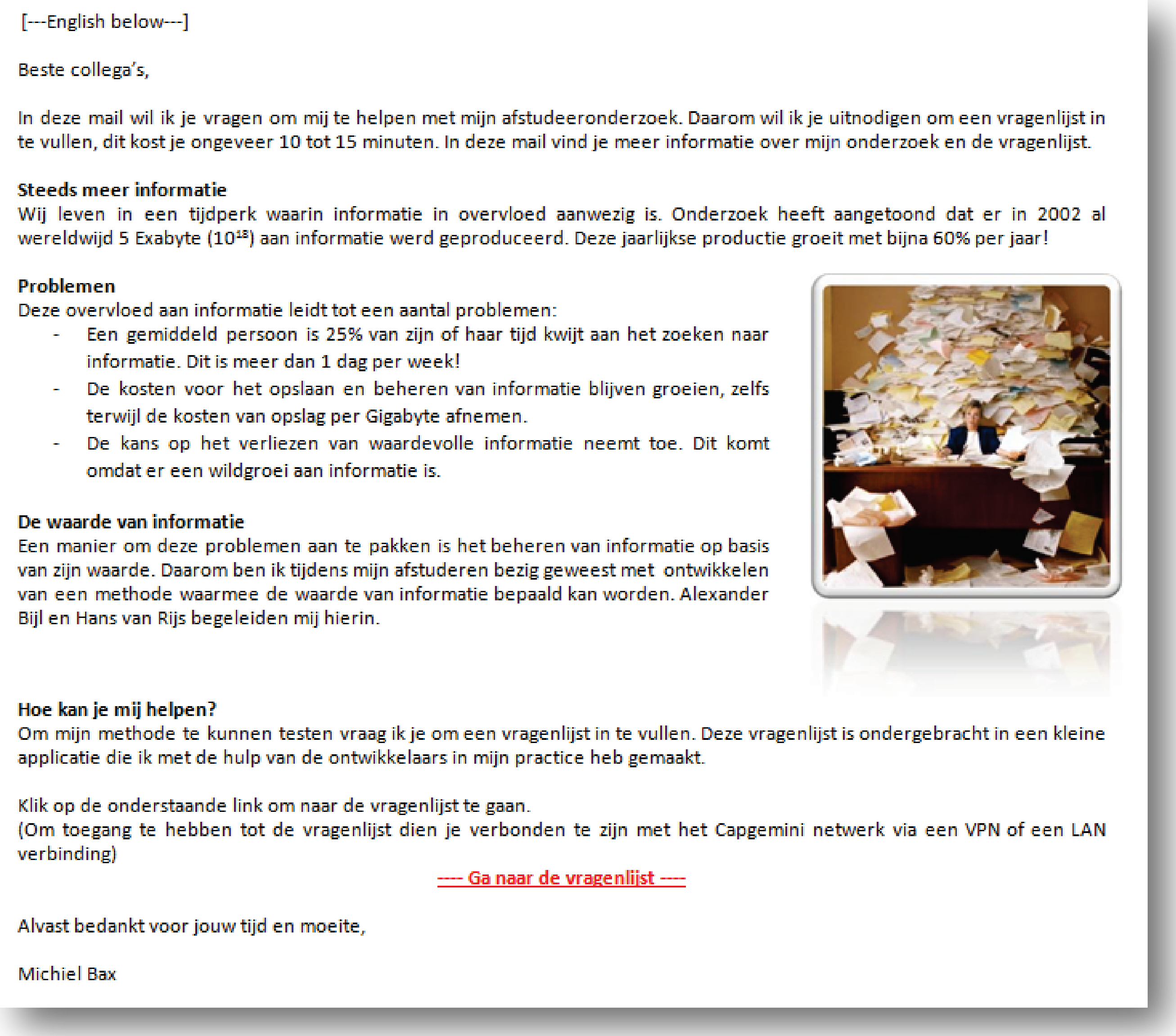
Thank you,

Michiel Bax

Figure IV.6: Final Page of the Survey

APPENDIX V: INVITATION FOR THE QUESTIONNAIRE

All respondents that have been invited to participate in the research by completing the questionnaire received the invitation that is found below in figure D.0.1. This invitation was sent by e-mail. An English version of the invitation can be found in figure D.0.2.



[---English below---]

Beste collega's,

In deze mail wil ik je vragen om mij te helpen met mijn afstudeeronderzoek. Daarom wil ik je uitnodigen om een vragenlijst in te vullen, dit kost je ongeveer 10 tot 15 minuten. In deze mail vind je meer informatie over mijn onderzoek en de vragenlijst.

Steeds meer informatie
Wij leven in een tijdperk waarin informatie in overvloed aanwezig is. Onderzoek heeft aangetoond dat er in 2002 al wereldwijd 5 Exabyte (10^{18}) aan informatie werd geproduceerd. Deze jaarlijkse productie groeit met bijna 60% per jaar!

Problemen
Deze overvloed aan informatie leidt tot een aantal problemen:

- Een gemiddeld persoon is 25% van zijn of haar tijd kwijt aan het zoeken naar informatie. Dit is meer dan 1 dag per week!
- De kosten voor het opslaan en beheren van informatie blijven groeien, zelfs terwijl de kosten van opslag per Gigabyte afnemen.
- De kans op het verliezen van waardevolle informatie neemt toe. Dit komt omdat er een wildgroei aan informatie is.

De waarde van informatie
Een manier om deze problemen aan te pakken is het beheren van informatie op basis van zijn waarde. Daarom ben ik tijdens mijn afstuderen bezig geweest met ontwikkelen van een methode waarmee de waarde van informatie bepaald kan worden. Alexander Bijl en Hans van Rijs begeleiden mij hierin.

Hoe kan je mij helpen?
Om mijn methode te kunnen testen vraag ik je om een vragenlijst in te vullen. Deze vragenlijst is ondergebracht in een kleine applicatie die ik met de hulp van de ontwikkelaars in mijn practice heb gemaakt.

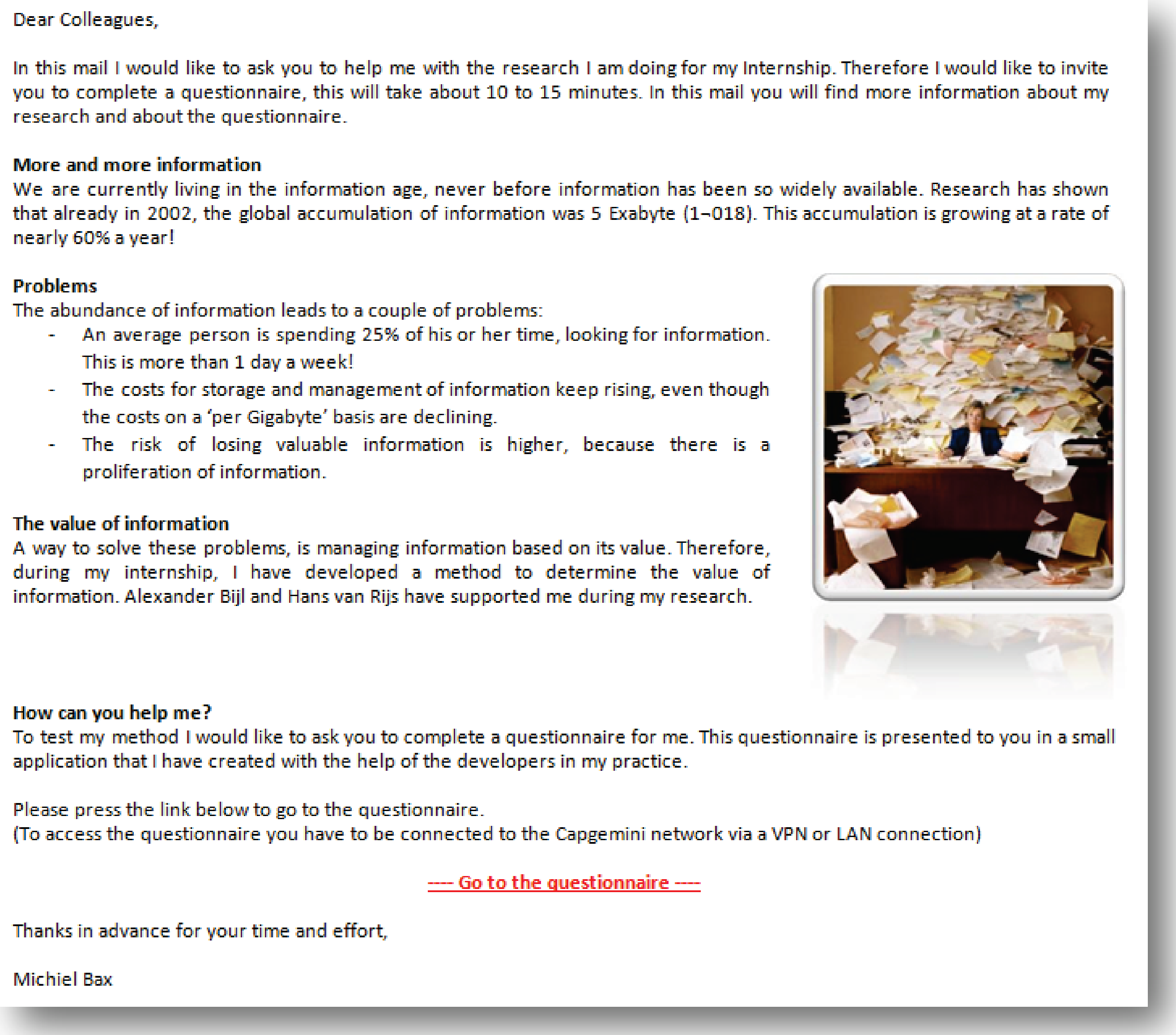
Klik op de onderstaande link om naar de vragenlijst te gaan.
(Om toegang te hebben tot de vragenlijst dien je verbonden te zijn met het Capgemini netwerk via een VPN of een LAN verbinding)

---- Ga naar de vragenlijst ----

Alvast bedankt voor jouw tijd en moeite,

Michiel Bax

Figure V.7: Dutch invitation for the Survey



Dear Colleagues,

In this mail I would like to ask you to help me with the research I am doing for my Internship. Therefore I would like to invite you to complete a questionnaire, this will take about 10 to 15 minutes. In this mail you will find more information about my research and about the questionnaire.

More and more information
We are currently living in the information age, never before information has been so widely available. Research has shown that already in 2002, the global accumulation of information was 5 Exabyte (1–018). This accumulation is growing at a rate of nearly 60% a year!

Problems
The abundance of information leads to a couple of problems:

- An average person is spending 25% of his or her time, looking for information. This is more than 1 day a week!
- The costs for storage and management of information keep rising, even though the costs on a 'per Gigabyte' basis are declining.
- The risk of losing valuable information is higher, because there is a proliferation of information.

The value of information
A way to solve these problems, is managing information based on its value. Therefore, during my internship, I have developed a method to determine the value of information. Alexander Bijl and Hans van Rijs have supported me during my research.

How can you help me?
To test my method I would like to ask you to complete a questionnaire for me. This questionnaire is presented to you in a small application that I have created with the help of the developers in my practice.

Please press the link below to go to the questionnaire.
(To access the questionnaire you have to be connected to the Capgemini network via a VPN or LAN connection)

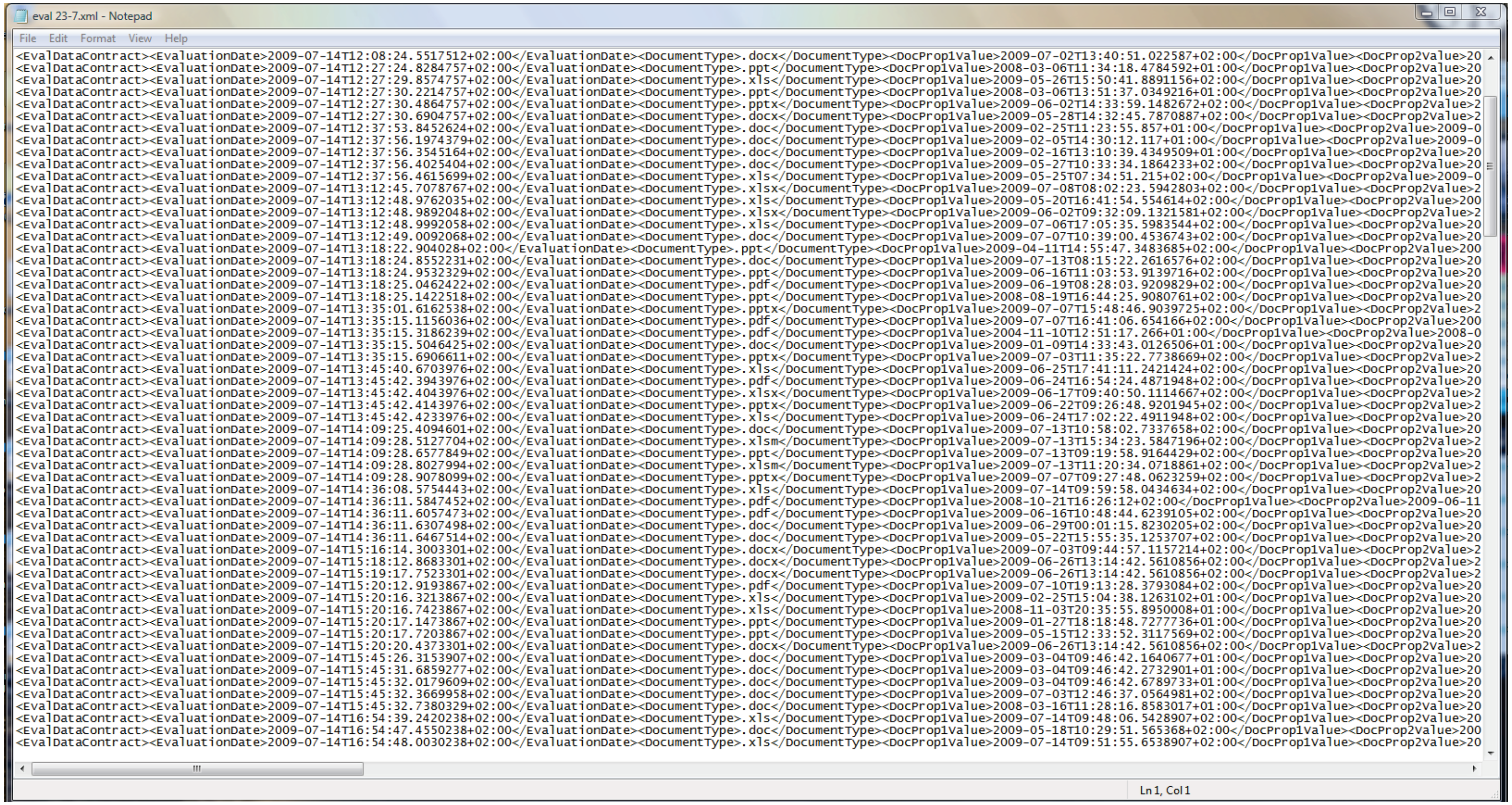
---- Go to the questionnaire ----

Thanks in advance for your time and effort,

Michiel Bax

Figure V.8: English Invitation for the Survey

APPENDIX VI: RESULTS TXT FILE



```
<EvalDataContract><EvaluationDate>2009-07-14T12:08:24.5517512+02:00</EvaluationDate><DocumentType>.docx</DocumentType><DocProp1Value>2009-07-02T13:40:51.022587+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T12:27:24.8284757+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2008-03-06T11:34:18.4784592+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T12:27:29.8574757+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-05-26T15:50:41.8891156+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T12:27:30.2214757+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2008-03-06T13:51:37.0349216+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T12:27:30.4864757+02:00</EvaluationDate><DocumentType>.pptx</DocumentType><DocProp1Value>2009-06-02T14:33:59.1482672+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T12:27:30.6904757+02:00</EvaluationDate><DocumentType>.docx</DocumentType><DocProp1Value>2009-05-28T14:32:45.7870887+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T12:37:53.8452624+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-02-25T11:23:55.857+01:00</DocProp1Value><DocProp2Value>2009-0
<EvalDataContract><EvaluationDate>2009-07-14T12:37:56.1974379+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-02-05T14:30:12.117+01:00</DocProp1Value><DocProp2Value>2009-0
<EvalDataContract><EvaluationDate>2009-07-14T12:37:56.3545164+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-02-16T13:10:39.4349509+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T12:37:56.4025404+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-05-27T10:33:34.1864233+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T12:37:56.4615699+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-05-25T07:34:51.215+02:00</DocProp1Value><DocProp2Value>2009-0
<EvalDataContract><EvaluationDate>2009-07-14T13:12:45.7078767+02:00</EvaluationDate><DocumentType>.xlsx</DocumentType><DocProp1Value>2009-07-08T08:02:23.5942803+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T13:12:48.9762035+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-05-20T16:41:54.554614+02:00</DocProp1Value><DocProp2Value>200
<EvalDataContract><EvaluationDate>2009-07-14T13:12:48.9892048+02:00</EvaluationDate><DocumentType>.xlsx</DocumentType><DocProp1Value>2009-06-02T09:32:09.1321581+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T13:12:48.9992058+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-07-06T17:05:35.5983544+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:12:49.0092068+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-07-07T10:39:00.4536743+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:18:22.904028+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2009-04-11T14:55:47.3483685+02:00</DocProp1Value><DocProp2Value>200
<EvalDataContract><EvaluationDate>2009-07-14T13:18:24.8552231+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-07-13T08:15:22.2616576+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:18:24.9532329+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2009-06-16T11:03:53.9139716+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:18:25.0462422+02:00</EvaluationDate><DocumentType>.pdf</DocumentType><DocProp1Value>2009-06-19T08:28:03.9209829+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:18:25.1422518+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2008-08-19T16:44:25.9080761+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:35:01.6162538+02:00</EvaluationDate><DocumentType>.pptx</DocumentType><DocProp1Value>2009-07-07T15:48:46.9039725+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T13:35:15.1156036+02:00</EvaluationDate><DocumentType>.pdf</DocumentType><DocProp1Value>2009-07-07T16:41:06.654166+02:00</DocProp1Value><DocProp2Value>200
<EvalDataContract><EvaluationDate>2009-07-14T13:35:15.3186239+02:00</EvaluationDate><DocumentType>.pdf</DocumentType><DocProp1Value>2004-11-10T12:51:17.266+01:00</DocProp1Value><DocProp2Value>2008-0
<EvalDataContract><EvaluationDate>2009-07-14T13:35:15.5046425+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-01-09T14:33:43.0126506+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:35:15.6906611+02:00</EvaluationDate><DocumentType>.pptx</DocumentType><DocProp1Value>2009-07-03T11:35:22.7738669+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T13:45:40.6703976+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-06-25T17:41:11.2421424+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:45:42.3943976+02:00</EvaluationDate><DocumentType>.pdf</DocumentType><DocProp1Value>2009-06-24T16:54:24.4871948+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T13:45:42.4043976+02:00</EvaluationDate><DocumentType>.xlsx</DocumentType><DocProp1Value>2009-06-17T09:40:50.1114667+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T13:45:42.4143976+02:00</EvaluationDate><DocumentType>.pptx</DocumentType><DocProp1Value>2009-06-22T09:26:48.9201945+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T13:45:42.4233976+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-06-24T17:02:22.4911948+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T14:09:25.4094601+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-07-13T10:58:02.7337658+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T14:09:28.5127704+02:00</EvaluationDate><DocumentType>.xlsm</DocumentType><DocProp1Value>2009-07-13T15:34:23.5847196+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T14:09:28.6577849+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2009-07-13T09:19:58.9164429+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T14:09:28.8027994+02:00</EvaluationDate><DocumentType>.xlsm</DocumentType><DocProp1Value>2009-07-13T11:20:34.0718861+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T14:09:28.9078099+02:00</EvaluationDate><DocumentType>.pptx</DocumentType><DocProp1Value>2009-07-07T09:27:48.0623259+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T14:36:08.5754443+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-07-14T09:59:58.0434634+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T14:36:11.5847452+02:00</EvaluationDate><DocumentType>.pdf</DocumentType><DocProp1Value>2008-10-21T16:26:12+02:00</DocProp1Value><DocProp2Value>2009-06-11
<EvalDataContract><EvaluationDate>2009-07-14T14:36:11.6057473+02:00</EvaluationDate><DocumentType>.pdf</DocumentType><DocProp1Value>2009-06-16T10:48:44.6239105+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T14:36:11.6307498+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-06-29T00:01:15.8230205+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T14:36:11.6467514+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-05-22T15:55:35.1253707+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:16:14.3003301+02:00</EvaluationDate><DocumentType>.docx</DocumentType><DocProp1Value>2009-07-03T09:44:57.1157214+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T15:18:12.8683301+02:00</EvaluationDate><DocumentType>.docx</DocumentType><DocProp1Value>2009-06-26T13:14:42.5610856+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T15:19:17.7523301+02:00</EvaluationDate><DocumentType>.docx</DocumentType><DocProp1Value>2009-06-26T13:14:42.5610856+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T15:20:12.9193867+02:00</EvaluationDate><DocumentType>.pdf</DocumentType><DocProp1Value>2009-07-10T19:13:28.3793084+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:20:16.3213867+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-02-25T15:04:38.1263102+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:20:16.7423867+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2008-11-03T20:35:55.8950008+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:20:17.1473867+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2009-01-27T18:18:48.7277736+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:20:17.7203867+02:00</EvaluationDate><DocumentType>.ppt</DocumentType><DocProp1Value>2009-05-15T12:33:52.3117569+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:20:20.4373301+02:00</EvaluationDate><DocumentType>.docx</DocumentType><DocProp1Value>2009-06-26T13:14:42.5610856+02:00</DocProp1Value><DocProp2Value>2
<EvalDataContract><EvaluationDate>2009-07-14T15:45:26.3153907+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-03-04T09:46:42.1640677+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:45:31.6859277+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-03-04T09:46:42.2732901+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:45:32.0179609+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-03-04T09:46:42.6789733+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:45:32.3669958+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-07-03T12:46:37.0564981+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T15:45:32.7380329+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2008-03-16T11:28:16.8583017+01:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T16:54:39.2420238+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-07-14T09:48:06.5428907+02:00</DocProp1Value><DocProp2Value>20
<EvalDataContract><EvaluationDate>2009-07-14T16:54:47.4550238+02:00</EvaluationDate><DocumentType>.doc</DocumentType><DocProp1Value>2009-05-18T10:29:51.565368+02:00</DocProp1Value><DocProp2Value>200
<EvalDataContract><EvaluationDate>2009-07-14T16:54:48.0030238+02:00</EvaluationDate><DocumentType>.xls</DocumentType><DocProp1Value>2009-07-14T09:51:55.6538907+02:00</DocProp1Value><DocProp2Value>20
```