Data quality and data alignment in E-business

Data Quality and Data

Alignment in E-Business

Ir. Bas H.P.J. Vermeer

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Vermeer, Bas H.P.J.

Data quality and data alignment in E-business/ by Bas Henri Peter Johan Vermeer. – Eindhoven: Technische Universiteit Eindhoven, 2001. – Proefschrift.

ISBN 90-386-0923-X
NUGI 684

Keywords: E-business / EDI / Data quality / Data alignment / Interorganizational systems / Information Integration / Interoperability / Interorganizational business process

Research Sponsored by EAN Nederland, Amsterdam, The Netherlands, and Deloitte & Touche Bakkenist, Amsterdam, The Netherlands

Cover: Ben Mobach
Printed by Eindhoven University Press Facilities

Data Quality and Data

Alignment in E-Business

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de Rector Magnificus, prof.dr. M. Rem, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 9 april 2001 om 16.00 uur

door

Bas Henri Peter Johan Vermeer

Geboren te Tilburg

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. A.F.L. Veth

en

prof.dr. P.M.A. Ribbers

Dit boek draag ik op aan hen die mij dierbaar zijn,

Anouk
Joost
Renske

Preface

In Dutch we have a saying: “Een goed verstaander heeft aan een half woord genoeg”, which translates to something like: “a good listener only needs half a word”. Exactly this is what computer systems cannot do. We have to spell it out for them word for word, before they can communicate fluently with each other.

In short, this has been the subject of my interest for the last 6 years. During these years, one person has been a tremendous inspiration to me, who is Ton Veth. Since I was Ton’s first PhD student, he always made time for me. We spoke each other almost weekly, and I think we have filled at least 500 whiteboards with interesting ideas. I think I am his only student that can bear to smell his pipe for more than half a day. Trust me, it takes years to achieve that. What really puzzles me is that Ton keeps on inspiring me, although now that takes place on the business side. Maybe we just get along very well.

The person who got me here was Piet van de Vlist. When I graduated in 1994, he immediately offered me a job at Bakkenist, with a dissertation at the University included. I guess Piet has a nose for students that have an interest in research. During my research, Piet performed the ungrateful task of tempering my enthusiasm, bringing back some of my wild ideas to rationally constructed arguments. Thank you Piet, for learning me how to do this.

Next I would like to thank the other two members of my small committee, Piet Ribbers and Kuldeep Kumar. Piet encouraged me to go to HICSS, which was not only a fantastic experience, but also provided the basis for my data quality chapter. Kuldeep made me become aware of the diversity in information systems research. This helped me to place my design research into the world of theory developing and testing research of my colleagues at the Edispuut. Finally it provided the basis for my methodology chapter.

Furthermore, I would like to thank Hans Wortmann, who was not on my small committee, but who paid a large contribution to this research, and Jacques Theeuwes, for his silent suggestions, whenever I was stuck in the research process.

I thank all the members of the Edispuut, a fantastic group of PhD students on the subject of EDI and E-commerce. The way this group helps each other with both content oriented and methodological comments is an example of doing value added research.

Of course I thank all the members of the *vakgroep* I&T (currently *capaciteitsgroep* I&T, what's in a name?), and especially all of my roommates through the years, first Rob van Stekelenborg, and later Tommy Dolan, and Remko Helms. I have many pleasant memories of the secretariat, where a large part of the social life in the vakgroep takes place. Since my beloved room D2 was only 2 steps away, I have had numerous discussions with Ineke, Karin and later Ada as well as all the vakgroep members each time I went to the printer. Being an AIO, this was quite a lot of course. What amazes me is that the mutual bond is so close, while so many people come and go each year.

My special thanks go out to EAN Nederland who sponsored my research, and especially to Hein Gorter de Vries. Hein opened many doors, so that I could do my research. Thanks to his contribution, I was able to perform the field study in the food sector, and develop the DAL method in the CBL case, and the EAN-DAS case respectively. Furthermore, I thank Deloitte & Touche Bakkenist, who offered me the opportunity to do my research next to a normal job as a consultant.

One person whose help has been invaluable at the end is Henri van Hirtum. A man I have never met, instantly freed up his agenda to read and edit my thesis. Thank you very much for making this thesis available beyond the community of Dutch-English speakers.

This research would not have been possible without the support of my family. Throughout my research I spent much time at my parents in Prinsenbeek or my family-in-law in Drunen without actually being there. I thank you for your patience. However, the largest support ultimately comes from the persons closest to you. How often I told Anouk, Joost and Renske that after this weekend it all would be over, I don't know. Only their patience and understanding helped me to complete this task.

Once Anouk told me that men shall never know the true meaning of being in labor. From my experience over the last 6 years, I think I finally know.

Bas Vermeer
's-Hertogenbosch, 7 februari 2001

Contents

Contents

1. Introduction

1.1 EDI in E-business

Technically it is rather easy to set up a website and start a Business-to-Consumer (B2C) E-commerce facility. It is far more difficult tot get that facility properly integrated with back office systems and to achieve integration with Business-to Business (B2B) partners backwards in the supply chain.

Integration between B2B partners also means interconnection of IT-systems. To achieve that, it is necessary to align the data so that appropriate data quality is assured. That is what this thesis is about.

This research started some years ago, when Edifact and ANSI-X12 based EDI (Electronic Data Interchange) were the dominant technologies to interchange inter-company data using dedicated value added networks (VANs). Today companies might choose to use XML-technology and the Internet to exchange their information. But the problem remains the same: they still need to align their data to assure appropriate data quality.

1.1.1 What is E-business?

E-business is defined as B2B E-commerce (Veth 2000). This means that several Internet technologies are used to integrate business partners in supply chains to optimize processes. This raises two questions:

1. Why do business partners in supply chains need to be connected?
2. How should this be done?

Why integrated supply chains?
The traditional supply chain consists of independently operating companies that focus on processing a small range of products for their immediate customer. These chains are in most cases product-oriented and the challenge to be dealt with is the distribution of finished products to the market. Therefore these chains are called push-chains. Traditionally each part of these chains is rather inward focused. This holds for optimization of the internal business processes as well as for the information systems. It is well known that even in the case that each part of the chain operates in an optimal way, the overall performance of such a chain is sub-optimal and suffers from the bullwhip effect.

As the demand for product variety increases, combined with increased competition on a global scale, organizations are forced to broaden their assortments, compress their product life cycle, shorten their time to market and improve their customer service (Browne et al. 1995).
In the new paradigm for supply chain management the focus is on the market requirements and the challenge for the chain as a whole is to meet these requirements in an optimal way.
Chains operating this way are called pull-chains or demand-chains.

To deal with these new requirements, companies extend their organizational boundaries to cooperate in a network of supply chain partners that deliver value to the ultimate consumer (Browne et al. 1995, Miles and Snow 1992). This change in perspective towards the ultimate consumer resulting in more integrated supply chains is shown in Figure 1-1 (adapted from Veth 1996). Hence, business partners in the supply chain use E-business to integrate the supply chain.

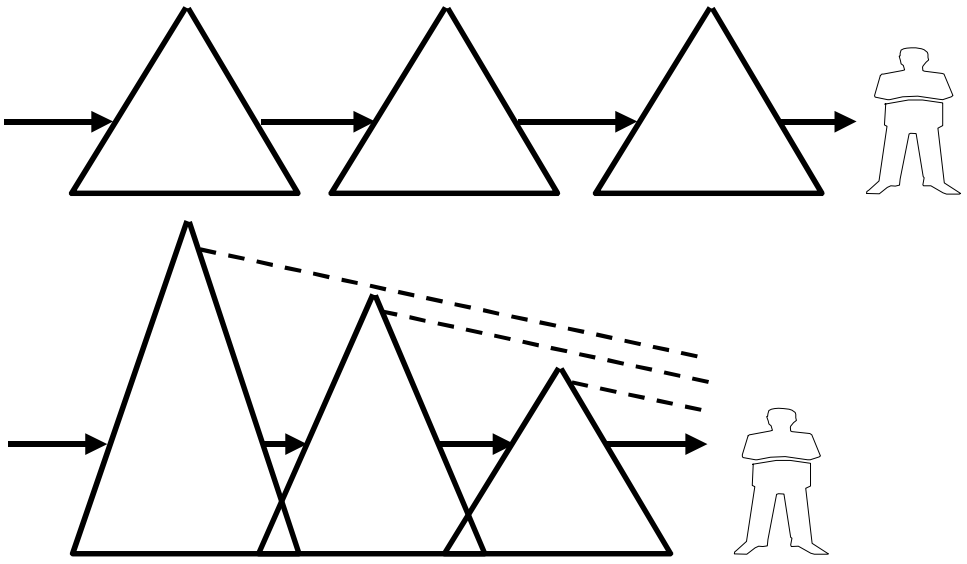


Figure 1-1: From traditional supply chains to integrated demand chains (Veth 1996)

It should be noted that in practice the market is not confronted with a relatively simple, linear, supply chain, but much more with a business network, including not only production and distribution services but also with transportation- and financial services. An individual company in general will be involved in many different supply chains. From this perspective it would be better to use the term business network management instead of supply chain management.

E-business to integrate supply chains
The next question is how modern information and communication technology can help to integrate the supply chain? From Figure 1-1 the following points of applications can be derived:

- Supporting the market to specify products and service requirements in such a way that chain processes can be optimized;
- Monitoring market behavior in order to optimize forecasting, for example by Point Of Sales scanning (POS), loyalty programs, or Customer Relationship Management systems (CRM);
- Exchanging market- and planning information throughout the chain, allowing for the application of Advanced Planning Systems (APS);
- Enhancing business-to-business communication in order to optimize inter-organizational processes, for example the E-procurement systems;
- Enhancing intra-business operations for example with the Enterprise Resource Planning systems (ERP), allowing for chain integration.

As can be derived from this list, most applications make extensive use of modern, internet-based, communication technology. This technology is not only used to optimize existing processes, but allows for complete inter-business process redesign. This can be illustrated by the trend in electronic purchasing or E-procurement, receiving currently much attention.

In 1983, Kraljic (1983) introduced a comprehensive purchase portfolio, which defined four strategies for possible buyer-supplier integration, using a product oriented approach (see also van Weele 1992). Veth and van Weele (Veth, 2000) use the Kraljic portfolio to explain which E-business technologies may be used in each of these four strategies. The adapted portfolio is shown in Figure 1-2.

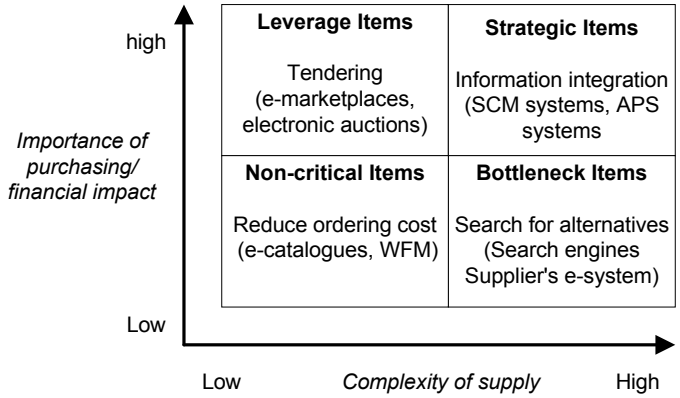


Figure 1-2: Strategies for buyer-supplier integration using e-business technologies

The portfolio explains that:

- In case of Non-critical items (low value items, with low supply risk), the purchasing strategy should focus on reducing the ordering costs. Both electronic catalogues, and Work Flow Management (WFM) are E-business technologies that offer convenient and fast searching of products for all departments in the organization and efficient and effective processing of orders.
- In case of Bottleneck items (low value items with a high supply risk), the purchasing strategy should focus on securing the resource and lowering the supply risk. Securing the resource can be done through concluding long-term contracts and then using the supplier's E-commerce system to order the products. Lowering the supply risk can be achieved through searching the Internet for alternatives.
- In case of Leverage items (high value, low supply risk), the purchasing strategy should focus on divide and rule. E-business technologies such as auctions and electronic tendering may support this process.
- In case of Strategic items (high value, high supply risk), the purchasing strategy should focus on strategic partnerships through information integration. This means that both partners integrate their respective business processes and supporting IT systems to improve reaction time, to lower transaction costs, and to reduce throughput time. E-business technologies such as Supply Chain Management (SCM) systems or Advance Planning and Scheduling (APS) systems are used to integrate the local ERP systems of the business partners in an overall system.

1.1.2 The role of EDI in E-business

EDI is defined as the automated electronic exchange of structured and standardized messages between computers of different organizations (van der Vlist, 1991). EDI focuses on direct system-to-system communication, without human intervention. Today's application of EDI in the automotive and the food sector shows how EDI is effectively used to integrate all Supply Chain Management (SCM) functions in the business chain. A good example is Efficient

Consumer Response (ECR, see Kurt Salmon Associates 1993), an EDI based supplier-retailer integration concept that effectively integrates over 30 IT applications in the food supply chain.

The question is how EDI fits in E-business? Wortmann (in Veth, 2000, pp. 10) provides a simple overview of several technologies used in E-business.

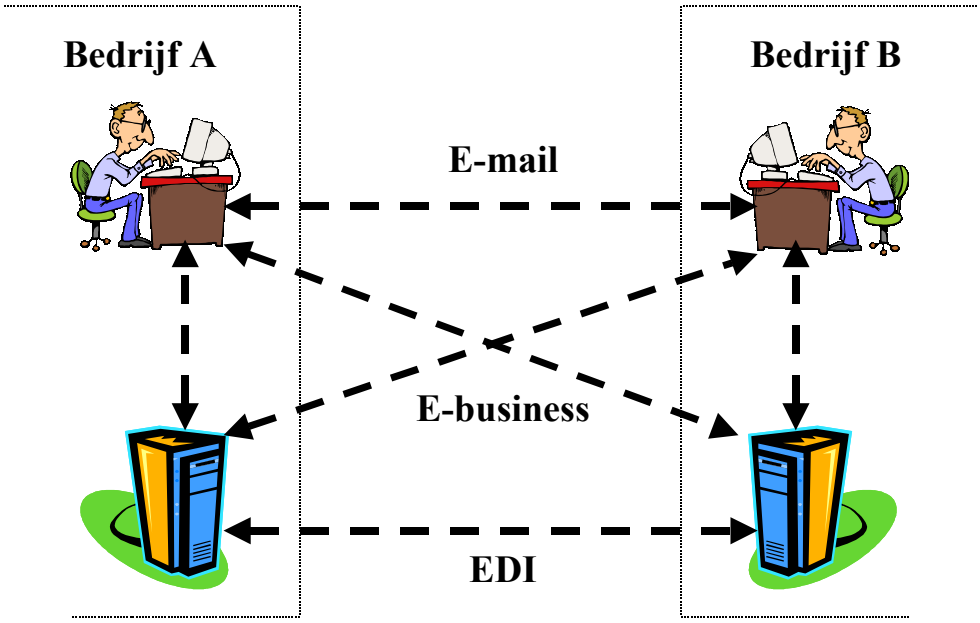


Figure 1-3: EDI in E-business

According to Wortmann, 3 types of IT-enabled communication between business partners exist:

- Human-to-human communication (E-mail);
- System-to-system communication (EDI);
- Human-to-system communication (website).

Originally, when supply chains were integrated, only human-to-human communication, and system-to-system communication existed. With the rise of the Internet, web technology was introduced, which enabled people to use applications from their transaction partners (human-to-system communication). Thus, EDI only focuses on system-to-system communication, whereas E-business focuses on all types of interorganizational communication.

Both Kraljic and Wortmann show that EDI is one of several technologies that together enable E-business. Wortmann shows that EDI is specifically focusing on system-to-system communication. Kraljic shows that the application area for EDI is in the area of information integration, where the integration of computer systems of B2B partners reduces the cost of frequently ordering (strategic) items significantly.

1.1.3 Focus of this thesis

Now we have explained the role of EDI in E-business we can define the focus of this thesis more clearly:

> *In this thesis we will study the problem of information integration for frequently ordered (strategic) items. Here, information integration is defined as integration of computer systems of business partners through system-to-system communication, for instance through using EDI.*

1.2 Problem statement: Insufficient product data quality

Exchange of electronic data (whether Edifact, XML, or XML-EDI is used) is the key-enabling technology for SCM or APS systems. We will start with a short review of the EDI–literature to understand benefits and problems of information integration based on electronic message exchange.

Using EDI in interorganizational business processes promises many positive impacts. EDI results in more *efficient* business processes through reduction in paper handling related activities, fewer data-entry errors, elimination of the data-entry function, a reduction of throughput time and hence in a reduction of inventory costs (Emmelhainz 1993). Furthermore, EDI results in more *effective* business processes (Venkatraman 1994, Clark & Stoddard 1996). For instance, Davenport (1993) demonstrates in his IT-Process-Productivity relationship model that IT initiatives (such as EDI) generate new process design options, which leads to more process productivity.

Although several studies regarding operational costs and benefits of EDI confirm these positive effects, (Kekre & Mukhopadyay 1992, Anvari 1992, Mackay 1993, Wrighly 1994, Jelassi and Figon 1994, Mukhopadhyay et al. 1995) several others do not (Benjamin et al. 1990, Carter 1990, McCusker 1994, Riggins and Mukhopadyay 1994). In several studies, insufficient product data quality is mentioned as a possible explanation for the observed disappointing results of EDI. For instance, in a study of the impact of EDI in the European automotive industry, Reekers and Smithson (1994) report several less than satisfactory impacts that could result from insufficient product data quality:

> *"In many cases EDI data insufficiently reflects reality".*

Ribbers (1995) reports that Technische Unie, a large Electrotechnical wholesaler, experienced major EDI implementation problems such as: (1) Problems with translation of article codes between suppliers, (2) Interpretation problems, because a large supplier did not make a distinction between different truck loads, and (3) Data synchronization problems, because not all codes of suppliers were known to Technische Unie. These problems are similar to the problems that were reported in a field study in the food sector, which concluded that:

> *"Many problems with scanning and EDI emerge from insufficient product data quality between the databases of suppliers and retailers" (Vermeer 1996:1, see also Section 3.2.1).*

A case study between a pharmaceutical supplier and wholesaler confirmed that:

> *"Many EDI problems resulted from major inconsistencies in the article data of the participants' databases" (Vermeer 1996:2, see also Section 3.2.2).*

The question is in what way insufficient product data quality impedes the positive effects of EDI. Below we present two examples that were reported in the grocery field study and the pharma case as an illustration of the data quality problem.

A large food retailer experienced major invoicing problems. The company received an invoice specifying boxes of shampoo with 20 bottles a box. However, its warehousing system only registered 10 bottles a box. Hence, the company refused to pay the bill. However, the supplier was certain it delivered 20 bottles a box. After investigating the problem, the retailer found the supplier was right. The investigation showed that bottles were taped together in pairs. In the warehouse, only one of the bottles was scanned, which resulted in the registration of only half the bottles received.

Example 1: Food retailer

A large pharmaceutical wholesaler was confronted with major scanning problems in the incoming goods process of its newly automated distribution center. More often than not, the product name displayed on the hand scanners was different from the name on the actual product packages. For instance, when the name on the product was Valium, the hand scanner displayed Diazepam. This led to much confusion: is the wrong product delivered or is the name on the scanner wrong? It turned out to be neither: both names are referring to the same product. Valium is the registered product name of the supplier. Diazepam is the name of the working substance in the product. The reason the wholesaler used the second name, was that in the early days of its information system the working substance was used to describe the product. This was done on the premise that the working substance was what the customers, the chemists, wanted to know. Although they now register both names, the hand scanners still display the original product description in the database.

Example 2: Pharmaceutical wholesaler

The examples show that the EDI problems in the business process mainly result from inconsistencies in the *context* of the communication between both senders and receivers. Because there are several inconsistencies in the product data in the context of the communication process (such as an unclear definition of what the unit of shampoo is, or what exactly is meant with the product name), EDI orders are incorrectly interpreted and hence lead to confusion and finally errors in the operational business process. Apparently, the use of EDI requires a higher degree of quality in the context of the communication process. We will define the degree of quality in the context of the communication process as context quality. With a high degree of context quality we mean that both the definitions of product data and the product data itself are aligned between the contexts of different organizations in the supply chain. Only when the degree of alignment between the product data of senders and receivers is high, EDI orders will contain fewer errors and hence will result in improved business process performance.

The role of data quality in an EDI-enabled interorganizational business process, as a result of insufficient product data alignment in the context of the communication process, is the central problem in this thesis.

1.3 Definitions of important concepts

Before we address the research objective and research questions, we will first provide the definitions of seven important concepts that are used throughout this thesis. These definitions will help the reader to understand the research objective and research questions unambiguously.

- Communication. We make a distinction between three levels of communication, where each level addresses a different aspect of the complete communication process. On the lowest level, communication is defined as the transmission of electric signals over noisy channels. On the middle level, communication is defined as the exchange of messages between information systems. Finally, on the highest level, communication is the exchange of signs between human beings that are interpreted using a shared norm system. Here, a sign stands for a physical reality according to an interpretant. For instance, the sign Java means a computer language to an IT expert, while it means coffee to non-experts. Finally, a shared norm system is defined as the shared believes (in terms of rules, definitions, or facts) of a social group of people.

- Context. In the definition of communication we have introduced the concept of a shared norm system. We will define the context of a communication process as the shared norm system of a sender and a receiver that is used to interpret the information in a message. Specifically, the context consists of the rules, definitions and data that are used to interpret a message.

- Data quality. The information in a message has a certain data quality. We make a distinction between three views on data quality, depending on what the information in the message is compared with. In the reliability view, data quality is defined as conformance to reality. In the relevance view, data quality is defined as the conformance to the user of the data (='fitness for use'). Finally, in the context view, data quality is defined as the conformance of the data with the shared context of sender and receiver.

- Data alignment. Normally, in a communication process, sender and receiver exchange messages to coordinate their business processes. We refer to this as transaction communication. As discussed before, both sender and receiver have a context that they use to interpret the information in the transaction communication. Whenever a change in one context occurs, an update is sent to the other party's context so that interpretation errors (of messages in the transaction communication) will not occur. The process of making agreements about mutual data and exchanging updates between contexts, is defined as *data alignment*. An example of data alignment is the communication of the price update of a certain product so that the new invoice of that product will not be rejected.

- Context quality. Since, according to the context view of data quality, information depends directly on the shared context, the quality of information will depend directly on the quality of the shared context. We will define this quality of the shared context as context quality. Context quality depends on the degree of *alignment* of the contexts

of the communicating parties. In an ordering process, this context quality can be operationalized as the degree of alignment between product databases of sender and receiver.

- Difference between context quality and data quality. Context quality and data quality are two separate constructs. Context quality is defined as the degree of quality of the shared context (an example is the percentage of exact record matches between the product databases of sender and receiver). Data quality is defined as the quality of the data in a certain message exchange between sender and receiver (an example is the percentage of correct EDI orders). When we define 'correct' in terms of conformance to the context, we are referring to the context view of data quality. This is not equal to context quality. Here, the context view of data quality means that the correctness of EDI orders is interpreted against the shared agreements between sender and receiver. This difference is shown in Figure 1-4.

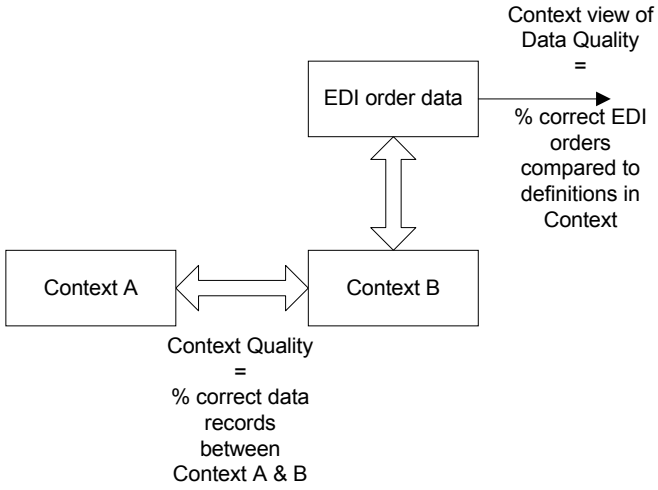


Figure 1-4: Context quality vs. context view of data quality

- Data (or fact update) distribution structure. We will define a data distribution structure as a set of procedures and functions that arranges the delivery of *fact updates* (such as product updates) to the right customer at the right time with the right quality in a network of multiple suppliers and receivers. In this thesis, data distribution structure and fact update distribution structure are used interchangeably.

1.4 Research objective and research questions

The research objective and the research questions are based on the results of three explorative cases studies we conducted to study the problem of insufficient product data quality in an EDI-enabled interorganizational business process. We will describe these results in detail in Chapter 3.

1.4.1 Research objective

The broad objective of this research is as follows:

> Develop a method that improves the degree of alignment in the context of the communication process, which will increase the context quality and hence the applicability of EDI in large interorganizational business networks.

To realize this objective in a methodologically responsible manner, we have limited this objective in three ways:

- This research will focus on the *food (or grocery)* sector. Although there are several other sectors where this problem exists, the food sector was one of the first that experienced this problem. Furthermore, the food sector is one of the most automated sectors. Our suspicion that large-scale automation may be an important cause of the problems justified our choice of the food sector.
- Since we suspect that large-scale automation is an important condition for this problem to arise, this further justifies our choice for the food sector.
- This research will focus on *product data quality*, as an operation of context quality. Although other sources of context information could be selected (e.g. information about sender/receivers, information about general pricing strategies, etc.) product information covers the largest part of the context of the EDI communication process.
- The research will focus on the *sector* level of intercompany organization. We distinguish between four different levels of intercompany organization as is illustrated in Figure 1-5.

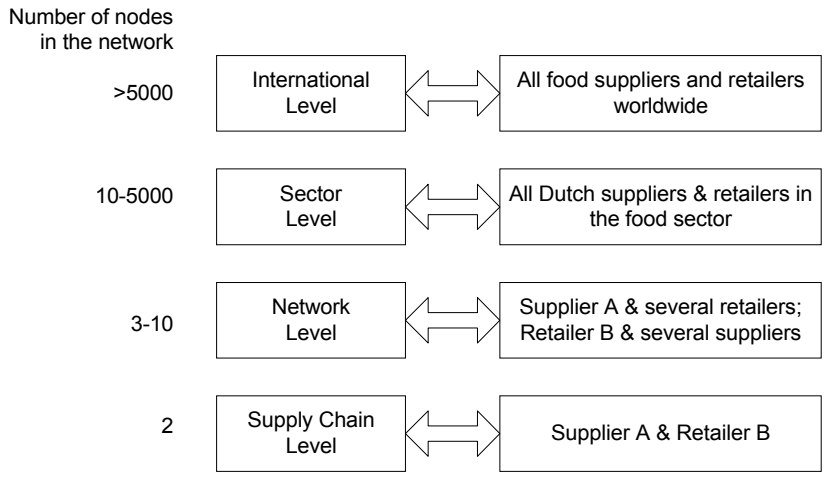


Figure 1-5: Four levels of intercompany organization

As we can see from Figure 1-5, we have defined the *supply chain level* as consisting of 2 companies, namely a specific supplier and retailer, for instance, the companies A and B. *The*

network level consists of 3-10 companies that work closely together in a specific network, normally a specific company with several suppliers and customers. We have defined the *sector level* as consisting of all suppliers and retailers in a specific sector, for instance all suppliers and retailers in the food sector. Such a network normally consists of hundreds to thousands of companies. Finally, we have defined the *international level* as all companies in a specific sector worldwide, such as all food suppliers and retailers in the world.

We choose to focus the research on the sector level because of two reasons: Firstly, the explorative research analysis (which is described in Chapter 3) shows that the main causes of insufficient product data quality in applying EDI lie on the sector level. Although issues on the network and supply chain level also play an important role, we believe that we have to start on sector level and from there narrow the method down to the network and finally the supply chain level. Secondly, apart from the organizations themselves, there is an important interest group that wants to solve the data quality problems with EDI, namely the E-business standardization agencies. As a representative of the sector, they have a specific interest in the advancement of EDI-based E-business in the sector.

This leads to the following limited objective for this research:

> Develop a method that improves the degree of alignment between *product* databases in the *food* sector, which will increase the product data quality and hence the applicability of EDI in interorganizational business processes for *the food sector*.

1.4.2 Research questions

To solve this problem, we first have to understand the relation between context, data quality and an EDI enabled business process. Therefore, we formulated the following two research questions:

1. What is the relation between context and data quality?
2. How does data quality affect an EDI-enabled business process?

Once we understand how data quality affects an EDI-enabled business process, we can design a data alignment method that solves the problem of insufficient product data quality in EDI enabled business processes. To do this, we will use the results of the analysis of the problem in Chapter 3. Here we found two causes for insufficient product data quality: (1) the lack of agreements on the sector level of the supplier-retailer network and (2) the lack of a data distribution structure between organizations in a supplier-retailer network.

Therefore the third and fourth research questions are:

3. Which design steps or guidelines should be applied on *sector* level to help organizations in a supplier-retailer network to specify agreements on the semantics of product data?
4. Which design steps or guidelines should be applied to help organizations in a supplier-retailer network to develop a data distribution structure for aligning the product data across the network?

This data alignment method is especially designed for E-business agencies, with interest in B2B integration in a specific sector. Examples of these agencies are ECR NL, ECP.NL (Electronic Commerce Platform the Netherlands) and EAN.

1.5 Research products

This thesis will deliver the following research products. First, a proposition will be developed and tested about the effect of data quality on an EDI enabled business process. Secondly, a method is developed which helps organizations in a supplier-retailer network to specify agreements about product data and to develop a data distribution structure to align product data across the network. This method is especially useful for E-business agencies, with interest in B2B integration in their sector. Lastly, based on the method, a formal model for defining a data distribution structure is developed that optimizes the structure of the network through allocating the different functions for data alignment in the network.

1.6 Contribution of this research

1.6.1 Theoretical contribution

This research makes a contribution to both the data quality and data integration research field. Firstly, data quality theory is extended with the notion of context. In addition to the existing definitions of data quality as 'conformance to reality' and 'conformance to user requirements', we add a definition of data quality as 'conformance to shared context'. This view is important to understand how context plays a role in explaining how data quality affects an EDI enabled business process.

Secondly, data integration theory is extended with the notion of data alignment. In the data integration field, the central problem is how to integrate heterogeneous data sources. This problem normally translates into the question of database schema integration: How should we translate between different, heterogeneous database schemas? We will show that in large interorganizational networks (as they exist on sector level), data alignment (=How to synchronize data between many different databases?) becomes an important problem. In the current data integration theory, this problem is not an issue, because this theory mainly focuses on data integration between a few (2 –5) heterogeneous sources. In that case, the alignment problem can be solved through an automatic database update mechanism. However when many (10 – 1000) different data sources exist, with constantly changing requirements and both manual and electronic update procedures, data alignment becomes an important issue that must be addressed (see Chapter 6 and 7).

1.6.2 Practical contribution

This research provides two practical contributions. Firstly, the research enables E-business agencies to design a structure for data alignment, thus solving product data quality problems with B2B integration for specific sectors (such as the food sector). Secondly, this research helps companies as well as B2B network designers to understand and ultimately solve interoperability problems when participating in B2B networks.

1.7 Outline of this thesis

As discussed in Section 1.4, we will address two issues in this thesis:
(1) The role of data quality in an EDI enabled business process;
(2) The development of a *design method*, to solve the problem of insufficient product data quality, to improve an EDI-enabled business process.

Figure 1-6 (on page 14) gives an overview of this thesis, explaining how we will address both issues. In this chapter (Chapter 1), we have introduced the research problem and defined our research objectives and questions. In Chapter 2, we will discuss the methodology that we extensively followed in this thesis. This is important, because we used two different methodological approaches to solve the two issues. We used the theory development and testing methodology for the first issue, while we used the positive design methodology for the second issue.

In Chapter 3, we will present the results of the exploratory phase of our research where we analyzed three explorative cases. The first two cases explored the issue of insufficient product data quality in an EDI-enabled business process, while the last case explored the causes of this problem and possible solutions for this problem. This has resulted in the research objective and questions that we addressed in Section 1.4. After Chapter 3 we address the issue of explaining the role of data quality in an EDI-enabled business process and solving this problem through data alignment separately (see Figure 1-6).

With respect to the *'role of data quality' issue*, we will review the theory on data quality and communication to explain the role of data quality in an EDI enabled business process, in Chapter 4. We will first examine the relation between context and data quality, since our practical findings show that context plays an important role. Next, we will use the context view of data quality to explain how context quality affects an EDI enabled business process. In Chapter 4.5 this results in the formulation of a proposition about the role of context quality in an EDI enabled business process. This proposition will then be tested in two data quality cases in Chapter 5.

With respect to the *'solving through data alignment' issue*, we analyzed theoretical solutions in the data integration and data distribution fields, to search for possible alternatives for our design method in Chapter 6. This leads to a list of requirements for our data alignment method at the end of Chapter 6. In Chapter 7, we will describe the first version of the Data Alignment through Logistics (DAL) method that we constructed from the results of the literature search in Chapter 6 and the results of the explorative cases in Chapter 3. This DAL method is further developed and tested in two case studies in Chapter 8. Here we will describe how we applied the method and what we learned through the process of critical reflection. The resulting second version of the DAL method provides the basis for formulating the data alignment problem in terms of formal model. Therefore, we added an extra chapter, Chapter 9, where we will transform the DAL method into a Binary Integer Programming model. This thesis concludes with Chapter 10, wherein we will first present the conclusions about each issue, after which further research is discussed.

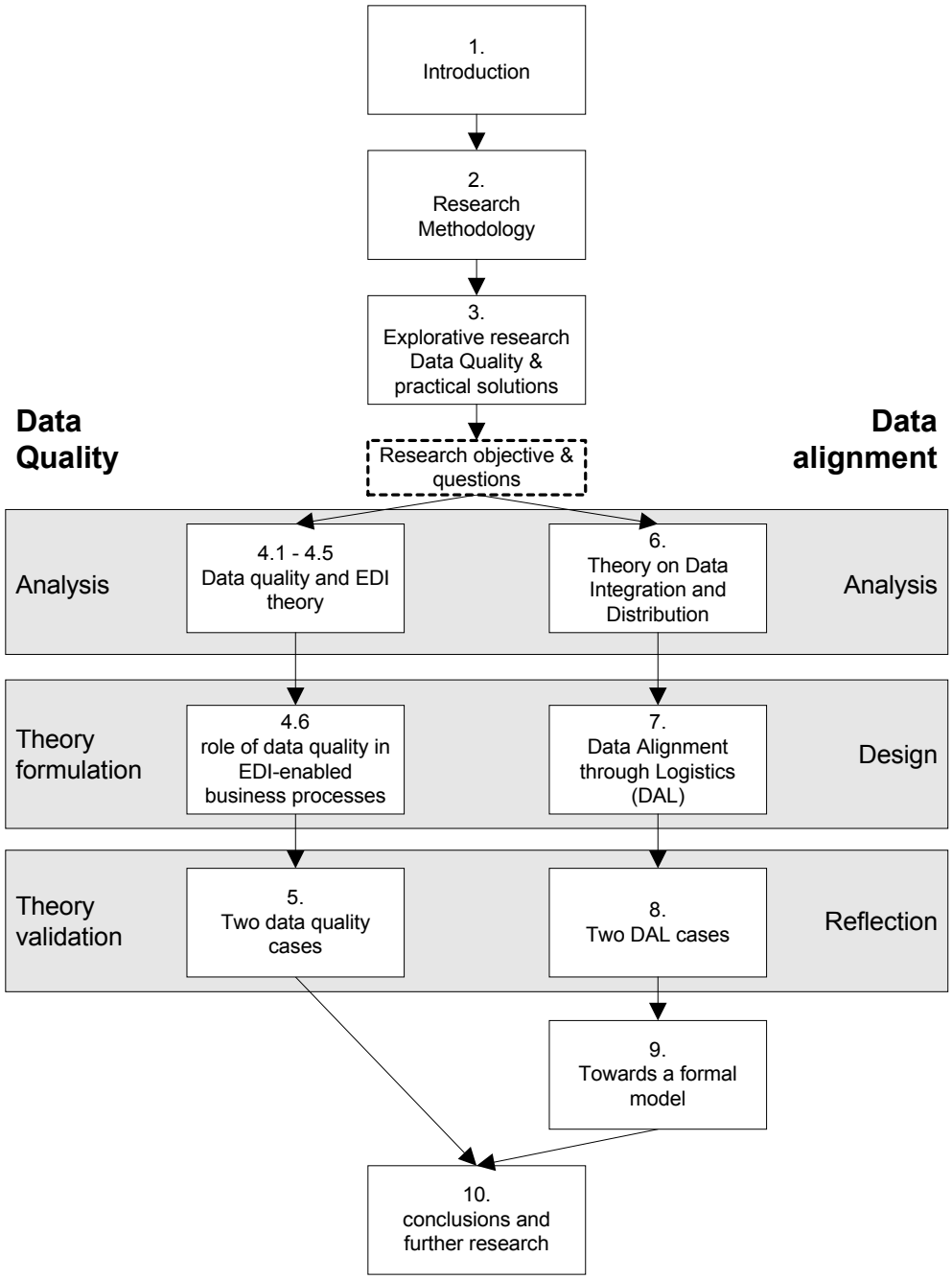


Figure 1-6: Outline of this thesis

2. Research Methodology

2.1 Introduction

Our research is conducted in the Information Systems (IS) field. The IS field is a young field and often characterized as diverse (Robey 1996, Benbasat & Weber 1996). According to Benbasat & Weber (1996), this diversity is found in (a) the problems addressed, (b) the theoretical foundations and reference disciplines that guide the research and (c) the methods used to collect and interpret the data. Despite this diversity, until recently only one dominant perspective guided the IS research. As Orlikowski and Baroudi (1991) put it:

> *"We examined 155 information systems articles published from 1983 to 1988 and found that, although this research is not rooted in a single arching theoretical perspective, it does exhibit a single set of philosophical assumptions regarding the nature of the phenomena studied by information systems researchers, and what constitutes valid knowledge about those phenomena".*

Analyzing the articles by research design, time frame of study and epistemology (epistemology refers to the assumptions about knowledge and how it can be obtained) Orlikowski and Baroudi conclude that 96.8 % of the articles are based on a positivist epistemology, using mostly surveys and laboratory experiments as research instruments in cross sectional, single snapshot studies. Because of this diversity, they suggest that a single perspective for studying information systems phenomena is unnecessarily restrictive. Therefore, they argue to include two other research perspectives, the interpretive and the critical (we will explain these three perspectives in the following section). They conclude that much can be gained if a plurality of research perspectives is effectively employed to investigate information systems phenomena (see also Lyytinen & Klein 1985).

This plea for wider diversity is currently more strongly adopted in the IS community. Several studies in leading MIS journals were published adopting interpretive perspectives (Markus 1994, Lee 1994, Myers 1994, Walsham 1995) or critical perspectives (Ngwenyama & Lee

1997). Recently, several articles about methodological issues when conducting studies with such perspectives have been published, such as the seven principles of interpretive field research (Klein & Myers 1999), the knowledge interest framework of Habermas to set up a critical social methodology Ngwenyama 1991, Lyytinen & Klein 1985) and the five requirements for conducting a critical social study (Ngwenyama 1991, pp. 272).

Currently, much attention is paid to the use of qualitative research methods in the IS field (Cash & Lawrence 1989, Nissen et al. 1991, Myers 1997) as opposed to quantitative research methods. Quantitative research methods were originally developed in the natural sciences to study natural phenomena. Examples of quantitative methods are surveys and laboratory experiments. Qualitative research methods were developed in the social sciences to enable researchers to study social and cultural phenomena. Examples of qualitative methods are case study research, ethnography and action research. Since these methods originated from the social sciences, which is often associated with the interpretive (or critical) perspective, many IS researchers regard these methods as non-positivist and therefore not valid. However, Myers argues that:

> *"Qualitative research is not synonymous with the interpretive perspective. Qualitative research may or may not be interpretive, depending upon the underlying philosophical assumptions of the researcher. Therefore, qualitative research can be positivist, interpretive, or critical (see Figure 2-1). For example, case study research can be positivist (Yin, 1984,1994), interpretive (Walsham, 1993) or critical".*

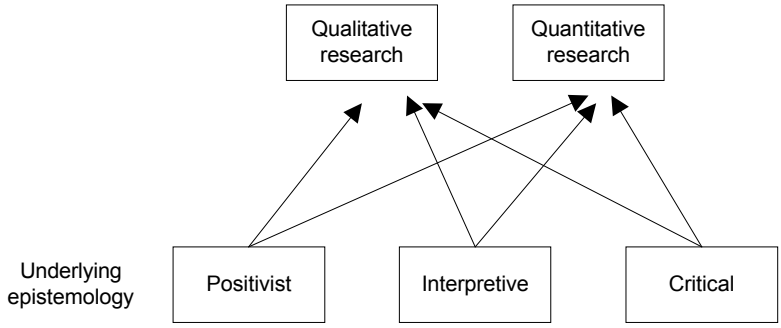


Figure 2-1: Perspectives on qualitative research

We think that the distinction between the research methods used and the underlying philosophical perspective is important to understand our research methodology. Therefore, we will first address the character of our research in Section 2.2. The character of the research describes both the philosophical perspective and the objective of our research. Next, we will address the research strategy in Section 2.3. The research strategy describes how the research methods are used to develop each type of research product. Finally, in Section 2.4 we will in detail describe the research method we used and define several criteria to validate the results.

2.2 Character of the research

To define the character of our research, we use two parameters:

- The underlying epistemology of the research;
- The type of knowledge that is generated.

We will first discuss each of these parameters, after which we will combine them into a table to clarify what the exact background of this research is.

2.2.1 Three philosophical perspectives

Following Chua (1986), Orlikowski and Baroudi suggest three philosophical perspectives.

Positivist perspective
Positivists generally assume that reality is objectively given and can be described by measurable properties, which are independent of the observer (researcher) and his or her instruments (Myers 1997). Positivist studies are premised on the existence of a priori fixed relationships within phenomena, which are typically investigated with structured instrumentation. Since these relationships are fixed, positivists seek to generalize their findings to construct general theories. Positivist studies serve primarily to test theory, in an attempt to increase predictive understanding of phenomena (Orlikowski & Baroudi 1991).

Interpretivist / constructionist perspective
Interpretivism asserts that reality, as well as our knowledge thereof, are social products and hence incapable of being understood independent of social actors (including the researchers) that construct and make sense of that reality (Orlikowski & Baroudi 1991). The underlying premise of the interpretive researcher is that individuals act towards things on the basis of meanings that things have for them, that meanings arise out of social interaction and that meanings are developed and modified through an interpretive process (Boland 1991). Interpretivism is based on phenomenology, which is committed to understand social phenomena from the actor's perspective (Taylor and Bogdan 1984 p2). Unlike phenomenology, interpretivism goes one step further and tries not only to understand but also to interpret immediate events in the light of previous events, private experience and whatever else the researcher finds pertinent to the situation under investigation (Gummesson 1991 p150). In interpretive studies, generalization from the setting (usually only one or a handful of field studies) is not sought; rather, the intent is to understand the deeper structure of a phenomenon, which it is believed can then be used to inform other settings (Orlikowski & Baroudi 1991).

Critical perspective
Orlikowski and Baroudi (1991) describe critical studies as aiming to critique the status quo, through the exposure of what are believed to be deep-seated structural contradictions within social systems, and thereby to transform these alienating and restrictive social conditions. Myers (1997) says that:

> *"Critical researchers assume that social reality is historically constituted and that it is produced and reproduced by people. Although people can consciously act to change their social and economic circumstances, critical researchers recognize that their ability to do so is constrained by various forms of social, cultural and political domination. The main task of critical research is seen as being one of social critique, whereby the restrictive and alienating conditions of the status quo are brought to light."*

Ngwenyama (1991) provides five fundamental assumptions of Critical Social Theory (CST):

1. People are the creators of their social world, and as such can change it if they wish.
2. Knowledge of the social world is value laden because all scientific knowledge is a social construction and all social constructions are value laden.
3. Reason and critique are inseparable. Reason means the capacity to understand the social world, to criticize it, and to search for and present alternatives to it.
4. Theory and practice must be interconnected, because we want to use knowledge to improve the human condition.
5. Reason and critique must be reflexive *in practice*. Reflexive means that CST must concern itself with the validity conditions of knowledge and change which it produces. This, however, is done through collaboration with those who will be affected, opening it for critical debate.

Many of these assumptions are based on the view of knowledge of Jürgen Habermas (1968). According to Habermas, knowledge is defined through human action. Habermas makes a distinction between two types of action (Lyytinen & Klein 1985):

- Purposive rational action, which is directed to achieving success;
- Communicative action, which takes place through language. Communicative action can be divided in normal communicative action, which is oriented to create mutual understanding and discursive communicative action, which aims to resolve conflicts through open rational discourse. This second mode of communicative action is the principal means in a social situation to validate claims and guard against systematic distortions (for instance power distortions).

According to Habermas, these three types of human action (purposive rational, normal communicative and discursive communicative), determine the cognitive strategies that guide systematic inquiry. Depending on the Knowledge Interest of the researcher, a different cognitive strategy is used to develop new knowledge. Habermas introduces the Knowledge Interest framework, which shows this view on knowledge and how it is obtained (see Table 2-1).

Knowledge Interest	Social action	Mediating elements	Sciences	Purpose	Process
Technical	Purposive-rational	Work systems	Empirical Analytic	Explanation, Prediction, Control	Scientific method, Verification
Practical	Communicative action	Cultural institutions, Natural language	Historical-hermeneutic, Geistenwissen-schaften	Understanding of meaning, Expansion of intersubjectivity	Idiographic method, Dialogue rules of hermeneutics
Emancipatory	Discursive action	Power, Unwarranted constraints	Critical sciences, Psycho-analysis, Philosophy	Emancipation, Rational consensus, Mundigkeit	Reflective method, Criticism of assumptions

Table 2-1: The Knowledge Interest framework of Habermas (Klein & Lyytinen 1985)

According to this framework:

- The technical Knowledge Interest is related to understanding the natural world. The technical Knowledge Interest uses instrumental rationality (based on purposive-rational human action) as a cognitive strategy to develop technical knowledge. Purpose of the technical Knowledge Interest is to explain, predict and control the natural world. Products of the technical Knowledge Interest are, amongst others, technology (e.g. procedures) and problem solving methods validated with regards to effectiveness.
- The practical Knowledge Interest is related to the quest for self-understanding. The practical Knowledge Interest uses communicative rationality (based on normal communicative action) as a cognitive strategy to develop social knowledge. Validation is achieved through the dialogue rules of hermeneutics.
- The emancipatory Knowledge Interest is related to our concern for freedom from distortions. The emancipatory Knowledge Interest uses dialectic rationality (based on discursive action) for critical reflection and analysis of instrumental rationality, with regards to their rightness. Products of this KI are norms for justice and freedom.

Because CST recognizes these different knowledge interests, the critical perspective allows different methods of inquiry depending on the specific knowledge interest (Ngenyama 1991). Hence, CST adopts a multi paradigm approach (see also Klein & Lyytinen, 1985, pp. 231).

Differences between the three perspectives
Important differences between the three perspectives are as follows. One, the positivist perspective is value free, which means that the observed reality exists independent of the researcher. Interpretive and critical researchers believe that the ideas and values of the actors including the researcher directly influence the observed facts and hence that facts are never value free. Two, in the positivist worldview, the researcher is detached from the phenomena of interest and aims to explain and predict external reality. This implies that people are not active makers of their physical and social reality. In both the interpretive and critical world view the researcher cannot be detached from the phenomena, because the results of the research clearly can and do change the nature of these phenomena. Thus, in these perspectives the researcher actively changes (social) reality. In the critical perspective the objective to change social reality through uncovering hidden assumptions is crucial. Three, positivist studies disregards historical and contextual conditions, which play a vital role in critical and interpretive studies respectively. Three important differences between interpretive and critical studies are that:

1. Interpretive studies, just as positive studies suppose rather stable and cooperative social relations, whereas critical studies explicitly incorporate conflicting situations. These situations are used to critique the current situation.
2. Interpretive studies use the hermeneutic method for inquiry, whereas critical studies adopt pluralistic inquiry methods, combining the empirical analytical method for the technical world and the hermeneutic and reflective method for the social world. Here the empirical analytical method is based on the scientific approach of theory formulation and testing. The hermeneutic method is based on creating an understanding of a complex social situation, through interpretation of its parts. The reflective method is based on critical examination of the validity conditions of knowledge and change that it produces.
3. Critical studies require some form of intervention because of its emancipatory objective.

Typology of perspectives
Based on the description of the three perspectives and their differences, we constructed a typology of perspectives, which is shown in Table 2-2.

Epistemology	Purpose	Process	
Positivism	Explanation, Prediction, Control	Scientific method using empirical analytical cycle	
Interpretivism	Mutual understanding	Idiographic method using hermeneutic cycle	
Criticism	Emancipation	Technical knowledge interest	Scientific method using empirical analytical cycle
		Practical knowledge interest	Idiographic method using hermeneutics
		Emancipatory knowledge interest	Reflective method based on criticism of assumptions

Table 2-2: Typology of perspectives

2.2.2 Two types of knowledge

As we can see from the discussion in the previous section, virtually all research in the IS field is aimed at theory, whether it is the testing of theories (that describe an objective reality as in the positivist view) or the creation of social theories (attempting to understand social situations as in the interpretive or critical view). This excludes an important part of the IS field, which is specifically interested in the *design* of information systems. Unlike their colleagues from the natural and social science disciplines, engineers are not so much interested in how things are, but in how things *ought* to be. To do this, engineers have to engage in a creative process to find acceptable solutions for the design problem. Since in the traditional disciplines science is considered value free, engineering was not regarded as scientific. Therefore, the emphasis in engineering education until the sixties lay in teaching the scientific principles from the traditional disciplines, such as thermodynamics and Newton's law. Next to that, engineering schools spent considerable time with providing their students with hands on experience (Suh 1990, p19).

In 1969, the highly respected professor Herbert Simon disassociated himself from this view on engineering by introducing engineering as a new scientific discipline, with its own scientific objectives and methods. In his book, Simon (1969, 1996 third edition) introduces the *sciences of the artificial*, next to the existing natural sciences. With artificial Simon means man-made as opposed to natural. According to Simon:

> *" Natural science is knowledge about natural objects and phenomena. We ask whether there cannot also be artificial science, knowledge about artificial objects and phenomena".*

Van Aken (1994:1&2) explains the need for this specific type of research, namely design research, as follows:

> *" It does not matter how many books about physics a person reads, they will never tell him how the wing of an aircraft should be designed. This can only be learned through studying the object and process knowledge from aircraft engineering schools".*

In 1977, Suh came with a similar plea for design research:

> *"What they (engineering schools) should have done, but did not do, was to change the research emphasis of these empiricism-dominated fields (of engineering) in order to establish a scientific base and develop these fields into disciplines. We needed a stronger engineering science base…"*

Suh responded to his own plea through developing the axiomatic approach to design. He introduced two general design principles, the *independence* and *information* axioms, that should help designers in making better designs.

Design knowledge is different from conventional theories. Conventional theories describe how things are, often through the use of causality models. Design knowledge consists of design models and heuristic statements. Van Aken (1994:2) defines a design model as an operation guideline, applicable for a specific application domain. For example, a design model in business engineering describes the different ways in which business processes should be designed or controlled in, for instance, assembly to order situations. Van Aken defines heuristic statements through comparing them with algorithmic statements.

> *"Algorithmic statements are of the form: When in this type of situation you do this, than that happens (deterministic algorithmic statement) or …, than in n% of the situations that happens (stochastic algorithmic statement). In contrast, heuristic statements are of the form: " When in this type of situation you do this, than it is* advantageous *to apply this design model."*

Thus, heuristic statements define guidelines and principles by which to operate. In summary, design knowledge differs from conventional theories in that it contains design models and heuristics. These models and heuristics describe how a designer should operate to attain some desired situation.

Although the type of knowledge that is generated is different, design researchers have the same objective as traditional researchers: They want to develop generalized knowledge (about design) that is applicable by engineering professionals. They do this through empirically testing which design models and heuristics are applicable in which types of situations using scientific methods. Thus, although the type of knowledge that is generated is not value free, the methods by which the design models and heuristics are tested is empirical and can be value free, depending on the researchers philosophical assumptions.

2.2.3 An IS research characterization matrix

Based on the two parameters we described above, namely the *epistemology*, which guides the method of inquiry in the research and the *type of knowledge*, which is the product of the research, we can define a matrix that helps us to characterize this research. This model was derived through an extensive review of methodology publications in the IS field, which we will

discuss first. Next, the IS research characterization matrix, which follows from this review, is presented.

Classification of IS methodology publications
Based on an extensive review of IS methodology publications, we constructed an IS methodology publication matrix, which is shown in Table 2-3. We will discuss the publications in each block of the matrix below.

	Type of knowledge	
Epistemology	Conventional theories	Design models & heuristics
Positivism	Whetten (1989) Bacherach (1989) Ives et al. (1980) Hamilton & Ives (1982)	Simon (1969) Suh (1990) Van Aken (1994:1,2)
Interpretivism	Winograd & Flores (1987) Boland (1985, 1991) Walsham (1993) Glasier & Straus (1967) Klein & Myers (1999) Peirce (1960)	Jones (1977) Alexander (1971) Schön (1983, 1987) Dorst (1997) Checkland (1981) Andersen et al. (1990)
Criticism	Ngwenyama (1991, 1997) Jönsson (1991) Forester (1992)	Lyytinen (1985) Ngwenyama (1991) Hirschheim et al. (1995) Weigand et al. (2000)

Table 2-3: Classification of IS methodology publications

Positivist, theory testing research
The publications in the upper left part of the matrix describe the traditional positivist view on research aimed at *testing* theory. Although this view allows that theory can be constructed, testing and not construction is the ultimate goal. The first two publications are from the organizational field and were taken from a forum of six papers, published in the Academy of Management Review. Wetten (1989) discusses the essential ingredients of a value added theoretical contribution. Bacherach (1989) introduces a set of ground rules in the form of a framework for evaluating theories. He defines theory as:

> *"a statement of relationships between units observed or approximated in the empirical world. Approximated units mean constructs, which cannot be observed directly. Observed units mean variables, which are operationalized empirically by measurement."*

The objective of research is to evaluate theory on aspects such as the utility and the falsifiability of the theory and the validity of the variables, constructs and relationships that are expressed in the theory. The last two publications (Ives et al. 1980 and Hamilton & Ives 1982) are from the IS field and contain similar descriptions of theory testing.

Interpretive, theory constructing or testing research
The publications in the middle left part of the matrix describe the evolving interpretive view on research aimed at constructing (social) theory. In the IS field Winograd and Flores (1987) and Walsham (1993) are important readings about interpretivism in IS research. According to Walsham, interpretive methods are:

> *"aimed at producing an understanding of the context of the information system and the process whereby the information system influences and is influenced by the context".*

Within interpretive research, several perspectives exist that differ in the way data is collected (method of inquiry) and the extent to which the acquired knowledge can be generalized.

We will present a short discussion of four different perspectives:

1. Philosophical hermeneutics (Gadamer 1976). Philosophical hermeneutics is based on the hermeneutic circle, which describes how we make sense of the world. The idea of the hermeneutic circle suggests that we come to understand a complex whole from preconceptions about the meanings of its parts and their interrelationships (Klein & Myers 1999). According to this circle, we move from a pre-understanding (which are the biases and prejudices we have about a new social text) via interpretation of this text (which is the process of understanding parts of the text using our pre-understanding) to understanding of the complete text, which in itself forms a new pre-understanding (Hirschheim et al. pp. 151). An important contribution comes from Klein & Myers (1999) who present seven principles that define how interpretive studies, based on the hermeneutic circle can be conducted and evaluated.
2. Phenomenology (Husserl 1982). Phenomenology is concerned with:

   > *" the structures of meaning that give sense and significance to our immediate experience (Boland 1985).*

   Through description of observed phenomena (such as the design or use of information systems), a true and accurate understanding of such phenomena is established. Although phenomenology originally is not based on hermeneutics, Boland extends phenomenology with the hermeneutic circle. A principle characteristic is that phenomenology does not aim at producing generalizable theory, but only true and accurate descriptions.
3. Grounded theory (Glaser and Straus, 1967). Grounded theory is defined as:

   > *"a theory that is inductively derived from the study of the phenomenon it represents. That is, it is discovered, developed and provisionally verified through systematic data collection and analysis of data pertaining to that phenomenon" (Straus & Corbin 1990).*

   In contrast with phenomenology, after a grounded theory is constructed, it may be generalized and tested.
4. Semiotics (Peirce 1960, Klein & Truex 1995). Semiotics is the study of signs. Specifically, semiotics is concerned with the exchange and interpretation of signs in social situations. Examples of signs are a sentence, a text, body language, a symbol (icon, stop sign), etc. Thus, theories are developed through the analysis of signs in a social environment.

Although we presented these perspectives as distinctive, they overlap in several ways.

Critical, theory constructing or testing research
The lower left part of the matrix describes the emerging critical view on research aimed at constructing or testing theory. Jönsson (1991, pp. 376) describes critical theories as follows:

> *Both critical & interpretive studies are dependent upon the understanding of the language of the individuals they study. This makes the empirical material temporally and spatially limited. Generalization is achieved by drawing out the principles of the observed action. Here, interpretivism stops. The critical perspective goes on to analyze what structures condition that these specific principles have been established in this specific setting.*

Thus, critical theories are defined in terms of principles of observed action and structures that conditioned current patterns of action. This view is consistent with Ngwenyama (1991 pp. 274), who defines critical theories as maps of action situations derived from critical reflective inquiry.

Two interesting case studies based on CST are Ngwenyama (1997) and Forester (1992). In the first case, a new theory of information richness is developed based on the CST framework. The interesting point is that the presence of communication richness in e-mail messages could only be explained using a CST-based social action framework, where users could critique the validity of rightness of what is communicated (which is considered to be rich communication). Positivist explanations (based on features of the process of communication) or interpretive explanations (based on mutual understanding) could not do this. In the second case study, a brief interaction between city planners is analyzed using discourse analysis. The study shows that only through reading (in terms of critical examination) the social world behind the words of the actors, the social situation can be understood.

Positivist, design research
The publications in the upper right part of the matrix describe the positivist view on design research. This view typically emanates from the work of Simon, who describes the design process as rational problem solving. Normally, the design process is characterized as follows: first there is a need, which is translated in Functional Requirements (FRs). These FRs are translated in Design Parameters (DPs) in a *creative* process. Finally, the proposed solution is *analyzed* to determine whether the proposed solution is correct and rational (Suh 1990).

Simon connected these design steps to the process of rational problem solving. The FRs represent the objective that the artifact should fulfill. The DPs define the problem space. Through an analytical process in which the DPs are optimized (or satisficed[1]), the best solution is selected. As Simon describes it:

> *" …the search for a solution through a vast maze of possibilities (within the problem space)… Successful problem solving involves searching the maze selectively and reducing it to manageable solutions".*

Through comparing design with rational problem solving, Simon connects the science of design with the discipline of mathematical optimization methods. According to Simon, a real science of design should be modeled on the natural sciences. It should be:

---

1 Simon introduces the concept of satisficing as opposed to optimizing. Optimizing means finding the best solution. Satisficing means finding solutions that are good enough. This is often the case in designing, since the number of possible design solutions is so large that it is not economically or otherwise feasible to find the best.

> *"the discovery of a partially concealed pattern in a rigorous and objective way. Design science could then become a body of intellectually tough, analytic, formalizable partly empirical and teachable doctrine about the design process".*

Van Aken (1994:1,2) describes the need to introduce design research in portions of the business- and information sciences. He claims that generalized statements (resulting from organizational or IS theories) are often not very relevant in practice. To justify the validity of these statements, they often have a very limited scope. Therefore, although they are true, they have a limited impact in business (engineering) practice. According to van Aken, the business professional is better served with specific design knowledge in the form of conceptual models or heuristics, than this type of general statements.

Interpretive design research

The publications in the middle right part of the matrix describe the interpretivist or constructionist view on design research. At its conception, this perspective fundamentally breaks with the rational problem solving perspective. According to Dorst (1997), this perspective:

> *"originated in architecture, where there was a feeling that the most important aspects of design practice could not be captured or supported by this (the rational problem solving) methodology…. Most fundamental were the suggestions by Jones (1977) and Alexander (1971), who proposed an alternative view on design and methodology, based upon a phenomenologically inspired epistemology and ontology of science. In books like 'The Timeless Way of Building' (Alexander 1979) architectural design is seen as a subjective and deeply human activity. The subjective knowledge humans attain through constructing their own realities is considered to be essentially personal and inherently non-generalizable. If categorical statements cannot be made about this knowledge, then general methods cannot exist".*

Although Jones and Alexander opened a new perspective on design methodology, the impact of their work has been limited.

In 1983, Donald Schön published his book 'The reflective practitioner' about the behavior of professionals, which was based on a constructionist perspective. In the constructionist perspective people actively construct a view of reality based on their perceptions of objective reality. Thus, although Schön believed that the positivist view on design was not useful to understand design and develop design methods, he clearly believed in more general methods for design. Schön wanted to do this through addressing the activity of design much more from the viewpoint of the designer. According to Schön, the activity of design does not consist of following a number of predefined steps. Rather, the designer is thrown into a design situation in which he has to develop a view of the design task and make several respective design decisions about 'what to do now' within a limited time period. Dorst (1997) describes that:

> *"To Schön every design task is unique, a 'universe of one'. Therefore, one of the basic problems for designers is to determine how such a single unique task should be approached. This problem has always been relegated to the 'professional knowledge' of experienced designers and was not considered describable or generalizable in any meaningful way. However, this does not satisfy Schön….Schön proposes an alternative view of design practice, based on the idea that 'a kind of knowing is inherent in intelligent action' (Schön 1983, p50). This 'action oriented', often implicit knowledge cannot be described within the paradigm of technical rationality….. What can be thought about and taught is the explicit reflection that guides the development of one's knowing in action habits. This he calls reflection-in-action…… According to Schön (1983), the designer executes a number of 'move-testing experiments' (involving action and reflection) to construct a view of the world on his/her experiences.… In this 'reflective conversation with the situation',*

> *designers work by naming the relevant factors in the situation, framing a problem in a certain way, making moves toward a solution and evaluating these moves. The frames are based on an underlying background theory, which corresponds with the designer's view about design problems and his/her personal goals. This background theory is not subject to change within design projects".*

Thus, Schön seeks to develop generalized design knowledge, but this knowledge is specifically described from the viewpoint of the designer who has to perform a specific design task in certain design situations. This view on design knowledge has much in common with the interpretive view on theory development.

Two interesting examples of studies that belong to this segment of the matrix are the Soft System Methodology (SSM) from Checkland (1981) and the Professional Work Practices approach (Anderssen et al. 1990). In their classification of IS development methodologies, Hirschheim et al. (1995, pp. 93) classify both SSM and the Professional Work Practices approach as part of the social relativist paradigm (which is based on an interpretive epistemology), because both methods primarily aim at creating mutual understanding based on genuine participation.

The interesting part about SSM is that it enables multiple views (root definitions in Checkland's terms) of the same system for multiple stakeholders, where each view *names and frames* the perception of the system (and what it is supposed to do) according to each specific stakeholder. Because of its focus on defining different views, SSM is specifically useful in the requirements definition phase of IS development, where the problem uncertainty is still high. The interesting part about the Professional Work Practices approach as described by Andersen et al. (1990) is that it is based on the assumption that:

> *"development research must be anchored in a thorough understanding of the actual work habits of practicing system developers (Hirschheim et al. 1995, pp. 130).*

Hence, both SSM and the Professional Work Practices approach contain many similarities with Schön's reflective practice.

Critical design research

Finally, the publications in the lower right part of the matrix describe the critical view on design research. Hirschheim, Klein and Lyytinen (1995), who in their book discuss conceptual and philosophical foundations of IS development and data modeling, could not find IS development approaches that follow the ideals of the critical perspective. According to Hirscheim et al. (1995, pp. 39):

> *"It (a critical IS design method) focuses on emancipation and adopts features of the previous generations (Hirschheim et al. describe 7 generations of IS development where each generation is largely influenced by one of the four philosophical paradigms they describe). It takes its motivation from the work of Habermas (1984) Theory of Communicative Action. It too conceives of systems development as a social process and sees the need for sense making (what is called mutual understanding), but where it differs is in its orientation toward emancipation, which is striven for through the use of rational or emancipatory discourse. Communication comes to the fore in this approach and hence vehicles are developed to overcome obstacles to free and undistorted communication. The goal of systems development is a system which would not only support emancipatory discourse but also mutual understanding for all its users."*

Based on the work of Lyytinen (1985), Ngenyama (1991) and Hirschheim et al. (1995), we propose the following criteria for critical design research:

1. The basic objective of critical design research is to develop models or methods that improve the human condition.
2. The models or concepts that are developed should be based on multiple paradigms depending on the specific knowledge interest. More specifically, depending on the phase of the design process, different knowledge interests apply, which ask for the application of a different paradigm. For instance, in the definition and analysis phase of the problem situation, the practical knowledge interest is prevalent, and hence methods or concepts that support mutual understanding and sense making should prevail. In the construction or solution phase a more positive perspective is effective, because the technical construction of the system is the main issue. Finally, in the implementation and use phase, both the interpretive and critical perspective should be adopted, because both sense making and rational discourse are prevalent. This criterion is consistent with the work of Dorst (1997) who concludes that the constructionist perspective tends to be used by designers in the problem analysis phase, while the rational problem solving perspective is adopted in the design phase.
3. Methods and concepts that are the result of critical design research should result in systems that support both rationale and critique and reflection in practice. More specifically, these systems should be able to create mutual understanding, and if conflicts about social norms arise, they should be able to open a rational discourse to resolve the conflict. An example of such functionality is described by van Reiswoud (1996) and Weigand (2000). For example, Weigand describes a workflow management system, which is based on norms about how the system should work. If the situation arises, the norms that define the workflow management system can be challenged, after which the system supports the creation of a rational discourse to resolve the norm that is challenged.
4. Critical design research requires full participation of the designer(s), the researcher and the users of the models or methods. Therefore, several researchers (Ngwenyama 1991, Jönsson 1991) propose the action research method as the preferred method for doing critical research (and specifically critical design research). Important reasons are: (1) that action research requires that the researcher is involved in the action situation, (2) that an intervention is a fundamental part of the research, (3) that generalized knowledge of action is produced through reflective inquiry, and (4) that action research seeks to create alternatives to the status quo, promote learning and change at the level of norms and values. However, in accordance with Myers (1997), although action research is preferred, this does not mean it is the only method that can be used.

IS research characterization matrix

Based on the review of IS methodology publications, we constructed the following IS research characterization matrix, which is shown in Table 2-4.

	Type of knowledge		
	IS theories	IS Design methods, models and heuristics	
Epistemology	that consist of……..	aimed at …..	through……..
Positivism	causal, one directional theories	conforming to requirements	FR-DP-optimize
Interpretivism	bidirectional maps; cognitive maps; conceptual models	creating mutual understanding	name-frame-move-reflect
Criticism	bidirectional maps; cognitive maps; conceptual models	improving the human condition	name-frame-move-reflect & FR-DP-optimize

Table 2-4: IS research characterization matrix

This table shows we make a distinction between IS research that aims at producing knowledge in the form of IS theories and knowledge in the form of IS design methods, models and heuristics. If the aim is to develop and test IS theories, the resulting theory differs depending on the philosophical perspective that the researcher takes. Thus, researchers who adopt a positive stance will develop causal, one-directional theories. If the researcher adopts an interpretive or a critical stance, the developed theories tend to be in the form of bi-directional maps, cognitive maps or conceptual models.

If the aim is developing IS design methods, models and heuristics, then the resulting method, model or heuristic also differs depending on the philosophical perspective the researcher takes. If the researcher takes a positive stance, the objective of the developed method is aimed at conforming to requirements. This aim is achieved through following the rational problem solving cycle (Simon 1969), which consists of the following steps (Suh 1990): Define functional requirements (FR), develop the different design parameters (DP) and construct an optimal fit of design parameters that conform to the functional requirements. If the researcher takes an interpretive stance, the aim of the method is to create mutual understanding. This aim is achieved within the method through the users who are naming and framing problems, move towards its solution and reflect on the outcome. This cycle is repeated continuously until a consensus is reached. Finally, if the researcher adopts a critical perspective, the aim of the method (or concept or heuristic) created is to improve the human condition. This is achieved within the method through both adopting the name-frame-move-reflect cycle and the FR-DP-optimize cycle, depending on the specific phase of the method.

2.2.4 Classification of the character of our research

We classify our research as **design research** with a **positivist perspective**.

Our research is **design oriented** because the objective of our research is to develop a method that solves the problem of insufficient data quality of product information on sector level, which will increase the applicability of EDI in interorganizational business processes for the grocery sector.

Furthermore, our research is based on a **positive epistemology** because the objective of the method is to provide a general method for data alignment to be used by data integration professionals. This method will describe problem representations, objectives and steps *independent* of the designer. Thus, we are not so much concerned with practical knowledge about what a designer should do when confronted with a certain data integration problem in specific situations (specifically how he should frame his problems, together with his users, in certain situations). Rather, we want to develop generally applicable representations, concepts and maps that each designer could use when confronted with data integration and data alignment problems. Therefore, we will model the knowledge that is contained in the method specifically after Simon's rational problem solving approach.

2.3 Research strategy

The research strategy defines both the product and the process of the research. The research product describes the specific type of knowledge that is obtained and hence the criteria for evaluating the results. The research process describes the different steps that were taken to obtain the knowledge and therefore provide criteria for evaluating the validity of the research findings.

2.3.1 The research product

To evaluate our design method, we will use Simon's curriculum for design, which consists of the following topics (Simon 1996 p134):

THE EVALUATION OF DESIGNS
1. Theory of evaluation: utility theory, statistical decision theory.
2. Computational methods.
   a. Algorithms for choosing *optimal* alternatives such as linear programming computations, control theory, dynamic programming.
   b. Algorithms and heuristics for choosing *satisfactory* alternatives.
3. THE FORMAL LOGIC OF DESIGN: Imperative and declarative logics.

THE SEARCH FOR ALTERNATIVES
4. Heuristic search: factorization and means-end analysis.
5. Allocation of resources for search.

6. THEORY OF STRUCTURE AND DESIGN ORGANIZATION: hierarchic systems.
7. REPRESENTATION OF DESIGN PROBLEMS.

Based on Simon's curriculum for design, we constructed the following criteria for a design method:

1. A design method should have a problem representation (Simon point 7).
2. A design method should state objectives and constraints (Simon point 1).
3. A design method should describe different research paths for instance through describing different scenarios (We deducted this from Simon point 4).

4. A satisficing design method should have an explicit progress evaluation function; an optimizing method should have an optimization function (Simon point 2 and 5).
5. A design method should state how the results are evaluated (Simon point 1 and 3).
6. A design method should be hierarchic (Simon point 6).

We will use these criteria to evaluate our method.

2.3.2 The research process

To describe the research process, we used the five inquiring systems defined by Churchman (Churchman 1971, de Vreede 1995):

1. Leibnitzian: expanding scientific knowledge by formal deduction from elementary forms of knowledge.
2. Lockean: expanding scientific knowledge by induction from sensing experiences, endowing them with properties and combining them with previous experiences.
3. Kantian: expanding scientific knowledge by formal deduction as well as by informal experiencing through a set of 'a priori sciences'; a blend of Leibnitzian and Lockean.
4. Hegelian: expanding scientific knowledge 'objectively' by identifying conflicting interpretations of observations and going beyond this conflict through synthesis.
5. Singerian: expanding scientific knowledge by adapting it 'endlessly', inductively and multidisciplinarily based on new observations.

We will use the Singerian inquiry strategy by conducting a series of case studies for our research objective. Van Aken (1994:1) uses this strategy in describing the serial case study strategy. According to van Aken:

> *"Analogous to criminal justice where one witness is no witness, in engineering science one case study is no case study. Therefore, design knowledge is normally developed in a series of case studies. To do this, two types of case studies are important: the exploratory serial case study and the evolving serial case study. In the exploratory serial case study the problem situation is described and best practices are identified as well as the problems that arise with less practices. These best practices are then brought back to their professional essence and their effects are studied through mutual comparison. ….He (the researcher) himself does not develop these methods (they are only explored).*
>
> *In the evolving serial case study he does develop these methods. He selects a typical case and goes through the normal regulative circle (the regulative circle consists of the following steps: problem identification, analysis, design, action and evaluation). After the case, he reflects on the results and uses the abstracted solution (that is, the design solution without the specific context of the first case) in the next case. This he continues to do until the design method has matured and he has enough empirical evidence about the effects of the method".*

The results of the exploratory serial case study are described in Chapter 3. The two cases that are used to develop our method are based on the evolving case study strategy. We will use the following starting points in our evolving case study strategy:

- In each case the evolving method is constructed through explicit reflection on the case at hand.

- Each case explicitly states indications and contra-indications to determine the class of problems for which the method is applicable.
- Each new case is selected based on the indications and contra-indications of the former case.
- Case studies start out with a strong exploratory nature, trying to establish which factors determined the final design in the case. In a later stage, the design method that evolved in the exploratory cases is tested in a strictly designed case (this design is based on the indications for usage of the method).

2.4 The research method

A variety of both qualitative and quantitative research methods exist (e.g. Miles & Huberman 1984). For an excellent overview of references in the IS field, we refer to Myers (1997). Because of our research strategy, we chose the case study as our primary research method.

Den Hartog and van Sluis (1995, p135) define three different types of case studies based on their objective. The objective of the case study can be:

- theory development (Eisenhardt 1989);
- theory testing (Lee 1989, Yin 1994);
- problem solving.

An excellent article about case study research in theory development studies is from Kathleen Eisenhardt (Eisenhardt 1989). Based on the work on qualitative methods from Miles and Huberman (1984), case study research (Yin 1984) and grounded theory building (Glaser & Strauss 1967), she provides an eight-step road map to *building* theories from case studies. In our opinion, important aspects are: the selection method of the case, the use of multiple data collection methods, the combined use of qualitative and quantitative data and if possible, the use of cross-case analysis.

For theory testing case studies, important publications are from Yin (1984) and Lee (1989). In his book about case study research, Yin provides numerous guidelines for improving the validity of using the case study instrument. Specifically, Yin suggests the following:

1. Define the type of the study. Is the case study exploratory or explanatory, aimed at testing theories?
2. Define the units of analysis in the case. Depending on the design of the case, several hierarchical levels may exist where data is collected and analyzed. For instance, job satisfaction could be analyzed on personal, group, department, organization and sectoral level. Yin specifically warns not to use the data from one level to make statements on another level, which compromises the validity of the case results.
3. Choose the case study design. Yin makes a distinction between a single and multiple case design and between holistic (=one unit of analysis) and embedded (= multiple units of analysis). There are three rationales for choosing a single case study design (Yin, 1993 p40-42): (1) if the case represents a critical case in testing a well-formulated theory, (2) when the case represents an extreme or unique case, and (3) when the case has a revelatory nature. According to Yin (1994, p45):

"The multiple case design has the advantage that the evidence is considered to be more compelling and is therefore regarded as more robust".

As well, the holistic design is advantageous when (Yin 1994, p42):

No logical subunits can be identified and when the relevant theory underlying the case is itself of a holistic nature. Potential problems arise, however, when a global approach allows an investigator to avoid examining any specific phenomenon in operational detail. Another typical problem with the holistic design is that the entire case study may be conducted at an abstract level... or may shift, unbeknownst to the researcher, during the course of the study… A major pitfall of the embedded design occurs when the case study focuses only on the subunit level and fails to return to the larger unit of analysis."

4. Define the criteria for judging the quality of research designs. These criteria include tests for (Yin 1994, p33):

   - Construct validity: establishing correct operational measures for the concepts being studied;
   - Internal validity (for explanatory or causal studies only and not for descriptive or exploratory studies): establishing a causal relationship, whereby certain conditions are shown to lead to other conditions, as distinguished from spurious relationships;
   - External validity: establishing the domain to which a study's findings can be generalized;
   - Reliability: demonstrating that the operations of a study – such as the data collection procedures – can be repeated, with the same results.

5. Define the case study protocol. A case study protocol contains the instruments that are used to collect the data, but also the procedures and general rules that should be followed in using the instrument. Thus, the case study protocol defines which instrument is used when and how to collect data from which data sources. We will return to these guidelines below.

Another important publication specifically for the IS field is the article of Lee, who describes a scientific methodology for MIS case studies. In his article Lee shows how the four requirements of a theory according to Popper can be fulfilled. Specifically Lee suggests to formulate several rivalry theories, on the basis of which several logically consistent predictions can be made, which then are falsified or confirmed through empirical testing.

We could not find articles that specifically deal with the methodological problems of using case studies for problem solving. Since problem solving is the objective of our case studies, we used the guidelines of Eisenhardt and Yin to describe our case studies and to define the criteria for evaluating our cases.

We identified the following criteria for describing and evaluating the case studies:

1. Case design
   The case design describes:

   a. The type of study: exploratory or testing; holistic or embedded?
   b. The case selection: why did we choose this case?
   c. The case objective and the case questions.

d. The case design: Through identifying the unit(s) of analysis, the system boundaries and their relations, the structure of the case is determined.
e. The criteria for interpreting the findings.

2. Case methodology
The methodology basically describes:

a. The data sources (people, documents, databases etc.).
b. The research instruments (interviews, data field comparison etc.).
c. The case study protocol: via which steps were the case results constructed from the data sources, using the research instruments.

In our case descriptions we will only present the case study protocol, which is assumed to bind the data sources and the research instruments together.

3. Case results
In the case results we will describe our research findings.

4. Conclusions
The conclusions cover the following aspects:

a. Case conclusions regarding the contents of the case. The case conclusions will cover two aspects:
   - What have we learned in the case? This means we will *reflect* on the results.
   - What have we learned regarding design methods? Does the method specify a way of representation, objectives and constraints; scenarios and evaluation criteria?
b. Methodology conclusions.

We will describe each case based on the criteria identified above.

2.5 Conclusions

In this chapter we provided an extensive discussion on research methodology in Information Systems research. The reason for doing this is that the IS field is still young, which results in a variety of methodologies based on different philosophical assumptions. In our opinion, it is only possible to make an unbiased evaluation of the research results when the philosophical perspective of the researcher is made clear to other researchers. Otherwise, the interpretation of the research results is troubled with the own philosophical assumptions of the evaluator. Therefore, we used a large part of this chapter to construct an IS research characterization matrix, which IS researchers may use to classify their research.

Once the character of the research is clear, the reader must understand what research strategy the researcher used and which methods were used in this strategy. In this way, the reader is able to check if the research strategy and method(s) that the researcher has chosen are consistent with the character of the research.

The research perspective, strategy and method of this research are summarized in Figure 2-2.

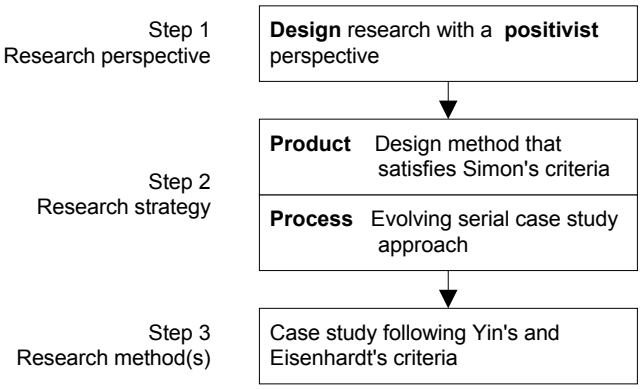


Figure 2-2: Our research perspective, strategy and method

Using Figure 2-2, we conclude that our research is classified as design research with a positivist perspective. The product of our research is a design method, based on Simon's criteria. To determine this product, we will use the process of the evolving serial case study. Finally, we will use Yin's and Eisenhardt's criteria for case study research to evaluate the results of our case studies.

3. Explorative Research

3.1 Introduction

In this chapter we will present the results of the practical analysis of the problem situation, as we described in Section 1.2. The analysis consisted of three explorative cases and an analysis of the conclusions from these cases. In Section 3.2 we will first describe the three investigative cases. The results of the three cases are then further analyzed in Section 3.3. Finally, in Section 3.4 conclusions of the explorative research follow.

3.2 Three investigative cases

The first case describes the problems with EDI and scanning technology in the Dutch food sector, which we will describe in Section 3.2.1. The second case in Section 3.2.2 describes the results of measuring the actual product data quality between a manufacturer and a pharmaceutical wholesaler, as an example of the seriousness of the product data quality problem. In the third case in Section 3.2.3, we will discuss several solutions for solving the product data quality problem, which are currently used in practice. Finally, in Section 3.2.4 the conclusions from the cases will follow.

3.2.1 Which problems exist? The food case

In the beginning of 1995, we conducted a field study, where we investigated the problems with scanning and EDI technology in the Dutch food sector (Vermeer 1996:1). The field study consisted of interviews with three food manufacturers and three food retailers. The manufacturers were selected from the top 10 of largest manufacturers for the Dutch food sector, of which all of them are multinational companies. The three food retailers were selected from the top 6 of largest Dutch food retailers.

In the interviews, we asked each company which typical problems they experienced with respect to problems with scanning and EDI. The results are shown in Table 3-1.

Problems	# companies from study that recognize this problem
1. Orders with outdated product numbers	3
2. Missing numbers in product look up tables (PLU) or product databases	5
3. Indistinctnesses about packaging units	5
4. Processing of order updates because of product- or packaging unit problems	2
5. Lagging behind with entering price-lists	2
6. Reception of different goods than ordered	4
7. Price differences between expected and actual invoice	6
8. Problems because of differences in internal codes and standard EAN product identification codes	4

Table 3-1: Frequency of scanning & EDI problems in the field study (Vermeer 1996:1)

The table shows that all problems are related to insufficient *product data quality*. Especially, price differences between the expected and actual invoice, missing product numbers in Product Look Up tables (PLUs) and indistinctnesses about packaging units are data quality problems that almost all interviewees recognize.

This study shows that companies in the food sector contribute problems with scanning and EDI largely to product data quality problems. Although the interviews were conducted with only 6 organizations from the Dutch food sector, the interviewed companies belong to the largest companies of the sector, thus giving these results more credibility.

3.2.2 How serious is it? The pharma case

To test the seriousness of the problem, we compared the content of the databases of a large pharmaceutical wholesaler and one of its largest suppliers (Vermeer 1996:2). The wholesaler is one of the three wholesalers in the Dutch pharmaceutical sector who together supply over 70% for the Dutch market. The supplier is one of the ten largest pharmaceutical companies in the world.

Both companies had extensive experience with EDI. At the end of 1995, the wholesaler communicated 15% of its orders to its suppliers using EDI. The supplier exchanged EDI messages with at least 5 of its customers. Together, these companies exchanged 100% of their mutual order lines with EDI. Also, both companies already used the order and invoice message extensively and were together doing a project to implement the EDI article message.

Firstly, we compared the EAN article codes in the databases of both supplier and wholesaler. An EAN article code is a standardized, internationally recognized product identification code, used for item coding. It appeared that the wholesaler related to that specific supplier had almost

300 EAN codes registered, compared to only 150 original EAN codes at the supplier. From the 150 extra EAN codes, 23% were out of date, 30% were wrong, 35% were interim codes (these were EAN codes the wholesaler used before the supplier had its own EAN codes) and 6% were invalid. Thus, *50%* of all EAN codes of the wholesaler were invalid in some way.

Secondly, we compared the attributes of the EAN codes that were identical. The results are shown in Figure 3-1.

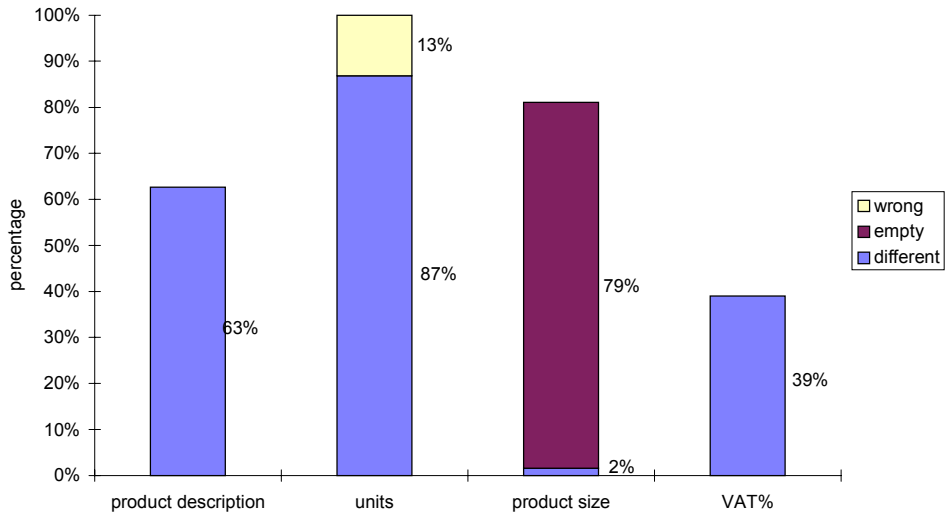


Figure 3-1: Attribute comparison between supplier and wholesaler (Vermeer 1996:2)

The results show that from the product descriptions 63% were different. This explains the problems that occurred during the scanning process, which we introduced in the example of the pharmaceutical wholesaler in Chapter 1 (Example 1-2). From the trade units, 87% were different, while 13% were completely wrong. This difference can be explained because the supplier uses the box as the trade unit, while the wholesaler uses the number of strips in the box as the trade unit**.** This may have serious consequences, as was shown in the example of the food retailer in Chapter 1 (Example 1-1). Furthermore, it appeared that the wholesaler only incidentally registered product size, which resulted in many empty database fields. Finally, the VAT percentage was different in 39% of the cases, which the firms had never noticed. This explained many invoicing problems between wholesaler and supplier.

This database comparison clearly shows that differences between databases can be significant, even though parties have extensive experiences with EDI.

3.2.3 What can we do about it? The Electrotechnical case

In the Electrotechnical case, we studied several practical solutions to solve the interorganizational product data quality problem. The results are based on interviews with four organizations from the Electrotechnical sector, a study of the impact of IT on the position of the wholesaler in four sectors, including the Electrotechnical sector (Coenjaerts et al. 1995), and desk research.

The four organizations we interviewed were:

- A large manufacturer of lighting products;
- A large Electrotechnical wholesaler;
- Instalnet, a Dutch Electrotechnical interest group for the standardization and implementation of EDI;
- The Electrotechnical sub department of EAN Nederland, the national Dutch EDI standardization organization.

Additionally, we studied several reports about the central ETIM article database (Kuijer 1991, 1992, Uneto 1993, Uneto 1994) and the use of the EDI article message (Instalnet 1994). The results of this study will be discussed below (for more details we refer to Jansen 1997).

The Electrotechnical business chain
The Electrotechnical (ET) sector supplies products for two major markets: the consumer market and the industrial market. For the consumer market, Electrotechnical products are manufactured and distributed to the consumers through department stores, supermarkets and specialized illumination stores. In the industrial market, illumination parts are supplied to the building industry, the maintenance market and the governmental market. These markets are characterized by (building) projects. This case focuses on the Electrotechnical supply chain for the industrial market.

This chain consists of two main parts: design and construction, and supply. In the first part, principals, their architects, constructors and contractors cooperate with the Electrotechnical installation companies in designing and installing the Electrotechnical subsystems in the building project. In the supply part, ET manufacturers deliver their products to the installers through the ET wholesalers. The ET supply chain for the industrial market is represented in Figure 3-2.

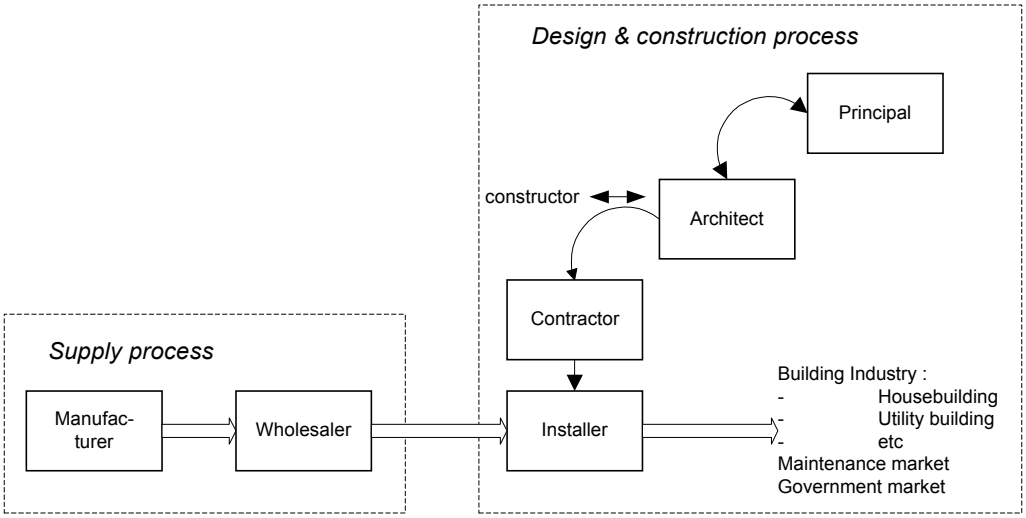


Figure 3-2: The ET business chain

Analogous with the distinction of the two parts in the ET business chain, we distinguish two main flows of product information:

- Technical product information, to support the design process;
- Commercial product information, to support the ordering process by selection information, logistical product information and price information.

In the ET supply chain we found that two initiatives for electronic information distribution exist:

- Centralized distribution via a sector wide product database;
- Decentralized distribution via EDI article messages.

Central product database
In the past few years, the installer's representative organization UNETO developed the Electrotechnical Information Model (ETIM). This model is a framework that contains a generic description of all business processes of the typical installer. The objective of ETIM is that software developers use it as a reference model, to guarantee optimal cooperation between the information systems of different installers. One result of this initiative is the development of the ETIM classification structure, a uniform technical classification of all products in the ET sector. The objective of this classification structure is to facilitate product search for installers.

A second result of the ETIM initiative is the UNETO central product database. This database is currently developed and is intended to contain all products of the ET sector (Uneto 1993, 1994). The objective of the database is to offer installers the possibility to search the database for all suppliers that supply products with certain technical specifications. These specifications are based on the ETIM classification.

The central database performs two functions: structured search for product information, and single source for updating master-data. The structured search function has been developed for purchasers that are searching a new product, to fulfill a specific function. The single source function has been developed mainly for the IT department, who can use the central database as a single source to update its company's product master data.

The following advantages are mentioned:
- No re-keying of information;
- Fast search of relevant suppliers;
- Single source for up to date product information.

The product database initiative is shown in Figure 3-3. Since the product database is a central reference point for all organizations in the sector, we will refer to this method as centralized distribution.

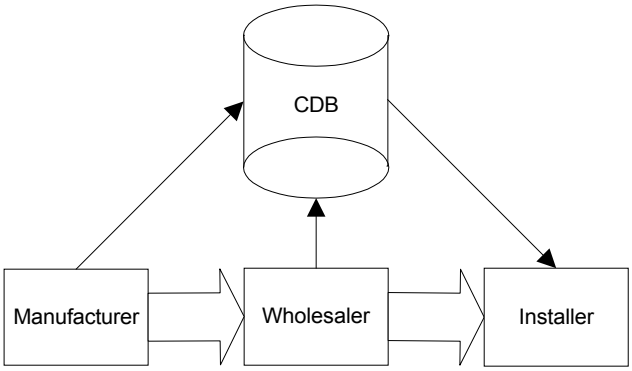


Figure 3-3: The central database initiative

EDI-article messages
Manufacturers and wholesalers in the ET business chain however proposed a different solution for the electronic exchange of product information, namely the use of EDI article messages. This means that a certain information supplier sends all relevant product information electronically via an EDI article message to the information customer. Because the message is send via the generic EDI standard format, this should guarantee automatic processing by the receiving application. Using the EDI article message also avoids re-keying of the information. Furthermore, the EDI article message is very flexible, since the EDI article message is always exchanged between any pair of organizations. This means that for each pair of organizations different implementation agreements can be established.

The EDI article message initiative is shown in Figure 3-4. Since the EDI article message is implemented between each pair of organizations, we will refer to this method as decentralized distribution.

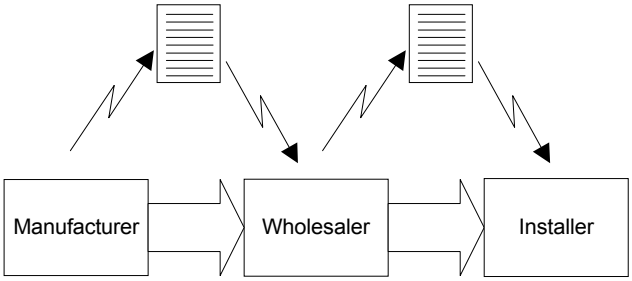


Figure 3-4: The EDI article message initiative

Problems
The main problem in the sector is that both installers and suppliers & wholesalers are arguing that their method should be used for product information distribution. From interviews with participants and related research we found several arguments against both methods.

Manufacturers and wholesalers are strongly opposed to a central database because this implies that installers not only receive (technical) information on their products, but also information on prices and availability of products. They fear that a central database creates a transparent market, where customers determine product choice only through price comparisons. An already existing example of such a situation can be found in the travel sector, where a travel agent can query the database for the cheapest ticket. A second argument is the fact that the information has to be supplied to a third party that is responsible for the collection and distribution of the information. This means that either a new party or an existing party will gain considerable power because it controls the entire flow of information in the business chain. Furthermore, the database contains valuable information such as sales information, price-development information, assortment information etc., which means that the party exploiting the database becomes an even more powerful player in the business chain. To summarize: The central database initiative means that manufacturers and wholesalers lose control over the product information.

The installers argue that the use of EDI article messages does not guarantee a complete market overview: this overview is depending on how many suppliers deliver information to the installers. Also, EDI messages do not guarantee that the information is actually delivered. The EDI article message only provides a standard format for sending product data. It does not provide a mechanism to ensure that the right information is delivered to the right place at the right time and in the right form. Instead, decentralized distribution with EDI article messages results in an arbitrariness of EDI messages sent to installers, who are overloaded with product information they do not need. To summarize: a standardized format to exchange article data will not prevent the data from being too late, not complete, or even completely false.

Other central database initiatives
Following the ET case, we studied over 10 different central database initiatives, which are presented in Table 3-2 (Bakkenist 1996, Jansen et al. 1997).

Year	Sector	Project name
	NETHERLANDS	
1994	Coachwork	Profiel
1980	Food	CUM
1996	Food	CBL database
1990	Major Food Consumers	GPI database
1993	Pharma	Productview
1995	Pharma	Information Warehouse
1995	Electrotechnical sector	ETIM database
1985	Music	Begotel
1996	Consumer electronics	Encodex
	INTERNATIONAL	
± 1990	Food	EAN product catalogue Australia
1996	Food	Swiss EAN code distribution system
± 1990	Food	SINFOSS, Germany

Table 3-2: Central database initiatives

Evaluating these initiatives, we found that almost all of them suffered from lack of support. From interviews we learned that there were two important reasons. First, political reasons such as *hidden agendas* and *disruption of the balance of power,* and second, *the large number of data fields*.

We found that competitors had hidden agendas because they are not happy about sharing information. In the ETIM initiative for the Electrotechnical sector, manufacturers were unwilling to cooperate because they feared market transparency. Although this was not the case, it delayed the development of the database for at least two years. A large Dutch retailer, who even refused to participate in a central Dutch product database that registered merely no more than the EAN article number, provides another example. It feared that its largest competitor would be able to search the database for all of its home brand articles. Although this information could also be retrieved through walking into one of its retail outlets, this possibility was enough for the retailer not to participate.

Furthermore we found that information disrupts the balance of power. Since the information in the database has to be administered, this creates a new participant in the sector. Since this participant has access to all data, it gains considerable power. This was clear in the Begotel case. When industry participation was falling, Begotel considered selling the data as marketing data to other parties. Clearly, this was not acceptable to industry participants (Vermeer 1995).

Finally, we found that unsuccessful initiatives had one thing in common: They contained many data fields (Bakkenist 1996 pp. 19). Sometimes over a hundred different fields were required per participant. This is in strong contrast with a few successful examples such as the Australian EAN product catalogue that has only 10 data fields. For data suppliers the problem is that they have to administer all these data fields. Since many fields are not relevant for them, it is not clear how they have to supply the necessary information. For example, the clothing size is very

relevant to a clothing manufacturer, but completely irrelevant to an office equipment supplier. Although this example may be obvious, in many instances this is not the case. For example, another important data field, such as the number of units per box, is very confusing to the same clothing manufacturer, as well as to fresh food suppliers (where weight is important).

To data receivers the problem is information overload. They receive an enormous amount of data fields they are not interested in. On the other hand, the information that is crucial to them often contains empty fields or is not reliable because of the data administration problems of the data suppliers.

3.2.4 Conclusions from cases

Based on the three cases we discussed in the previous three sections, we draw the following two conclusions:

1. The problems with scanning and EDI result from insufficient quality of product data across different links in the supply chain. Both the food case and the pharma case show that product data quality problems exist and are serious.
2. In practice, there are three approaches for solving the data quality problem:

   - More staff in the data administration department;
   - Electronic distribution of product data via EDI article messages;
   - Electronic distribution of product data via a central product database.

   However, despite their good intentions, these solutions do not work.

We will discus each of these solutions shortly and explain why they do not work.

More staff
In our study of the food sector we found that several large food retailers solved the problem through hiring more data-entry personnel. Their job is to scan catalogues and product descriptions of suppliers and pursue them for accurate product data. To facilitate this task, retailers are forcing their suppliers to fill out retailer specific product forms. This results in even more administrative personnel, only this time at the supplier's side. The question is, whether we do not overshoot the mark. The retailers implemented EDI to eliminate staff, which they are now rehiring to enter the data that supports EDI. This was clearly not the objective when EDI was introduced.

EDI article messages
EDI article messages are electronic product messages to send price information (PRICAT) and product information (PRODAT) in a standard form. The main problem is that using EDI article messages does not solve the information receiver problem. Information receivers are often flooded with information they did not ask for, while the information they need is not on time, not complete, or just wrong. We will clarify this with an example involving business cards. Similar to electronic article messages, business cards contain information in a standard format. However, as many have experienced, after some time most cards they possess contain outdated, mostly invalid information. This happens because most people do not send updates, simply

because they do not remember who they gave their cards to. Therefore, opposite to what is generally assumed, the existence of a standard in itself is not enough to guarantee synchronized data.

Central database
From the ET case and the other central database initiatives, we found two important reasons that caused resistance to central databases. Firstly, political reasons such as *hidden agendas* and *disruption of the balance of power* and secondly, *the large number of data fields*.

3.3 Problem analysis

We further analyzed the conclusions from the investigative cases using a Current Reality Tree (Goldratt 1997) to determine what the core problems are and hence which requirements a solution has to fulfill. The results of this analysis are described below. The analysis builds on the results of an early problem analysis described in the food case (Vermeer 1996:1) and a qualitative problem analysis that we conducted in the EAN-DAS case, where we also used Goldratt's techniques (see Section 8.5.5). For a detailed overview of the Current Reality Trees (CRTs) we used in the analysis, we refer to appendix A.

3.3.1 The problem of insufficient product data quality

The first CRT shows that insufficient product data quality is the result of two core problems:

1. There are no central agreements about the semantics of the product data, and
2. There is no distribution structure to synchronize the product data between the databases of multiple suppliers and customers. A distribution structure is defined as a set of procedures and functions that arranges the delivery of the right data to the right customer at the right time in a network of multiple suppliers and customers.

The fact that there are no central agreements about the semantics of product data means that customers must interpret the meaning of the data they receive. This means that the receiver either assigns his own meaning to the data or contacts the supplier. If the receiver assigns his own meaning, interpretation errors will occur, which leads to wrong data and hence insufficient product data quality. Otherwise, if the receiver has to contact the supplier, this normally results in a lot of extra work, which means that information updates are processed too late and hence also results in insufficient product data quality.

The fact that there is no distribution structure means that there are no clear procedures that arrange how data should be delivered, when and to whom. Therefore, to assure that customers receive their product updates, suppliers send their updates to everyone they know. Furthermore, since suppliers do not know precisely what customers want because there are no central agreements, suppliers will send what they *think* the customers want. This is normally less than what a specific customer needs, because the supplier will try to minimize the amount of work to collect the information. Thus, customers are overloaded with information they do not need (because many suppliers send information from products the customer does not sell), while the information they need is often not available (because the information is too little for their specific purpose). To solve this problem, receivers have to sort and select the information, and

retrieve all information that is still missing. Since this costs a lot of extra work, updates are processed too late, which results in insufficient product data quality.

3.3.2 Three different solutions do not solve the problem

The second CRT explains why the three solutions do not solve the product data quality problem. For one, all three solutions do not solve the core problem of making agreements about the semantics of the data. Secondly, regarding the problem of the distribution structure, the solution of EDI messages does not provide any mechanism to get product data at the right time to the right place. The solution of more staff in the administration department solves the distribution problem only partly, depending on the extra staff at the customer to pull the data through the chain. The central database solution does provide a distribution structure, because suppliers have to deliver the information to the database where customers can 'shop' to retrieve the right information. However, in the central database solution both suppliers and customers require several additional safeguards such as restricted access and no third party owner.

3.4 Conclusions of the explorative research

The objective of the explorative research was to explore what the cause for the problems with scanning and EDI was, to collect evidence of the seriousness of the problem and to determine what the core problems of insufficient product data quality are.

Based on the results of the explorative cases and the problem analysis, we draw the following conclusions:

1. The problems with scanning and EDI result from insufficient data quality of product data across different links in the supply chain. Both the food case and the pharma case show that product data quality problems exist and are serious.
2. In practice, there are three approaches for solving the data quality problem:

   - More staff in the data administration department;
   - Electronic distribution of product data via EDI article messages;
   - Electronic distribution of product data via a central product database.

   However, despite their good intentions, these solutions do not work.

3. The problem of insufficient product data quality is the result of two core problems:

   - There are no central agreements on the semantics of the product data, and
   - There is no data distribution structure to synchronize the product data between the databases of multiple suppliers and customers. A data distribution structure is defined as a set of procedures and functions that arranges the delivery of *updates* (such as product updates) to the right customer at the right time with the right quality in a network of multiple suppliers and customers.

Explorative Research

4. Data Quality and Context

4.1 Introduction

In our problem statement in Chapter 1 we discussed the problems of scanning and EDI in interorganizational business processes, and how we found that insufficient product data quality in the context of the communication was an important cause for these problems. The short EDI literature review, that we provided, shows that EDI results in a reduction of paper handling related activities, fewer data-entry errors, and elimination of the data entry function (Emmelhainz 1993, Hoogeweegen et al. 1996). Thus, EDI use positively influences the quality of the data, which results in more efficient business processes. This proposed relationship between EDI, data quality and the effect on the business process is shown in Figure 4-1.

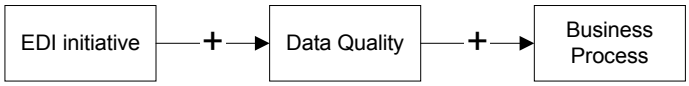


Figure 4-1: Relation between EDI-use, data quality and the effect on the business process

This relationship does not include the context of the EDI-enabled business process, and how this affects an EDI-enabled business process. Therefore we developed the following two research questions to explain the role of context in an EDI enabled business process:

1. What is the relation between context and data quality?
2. What is the role of data quality in an EDI-enabled business process?

In this chapter we will address both research questions as follows (see Figure 4-2).

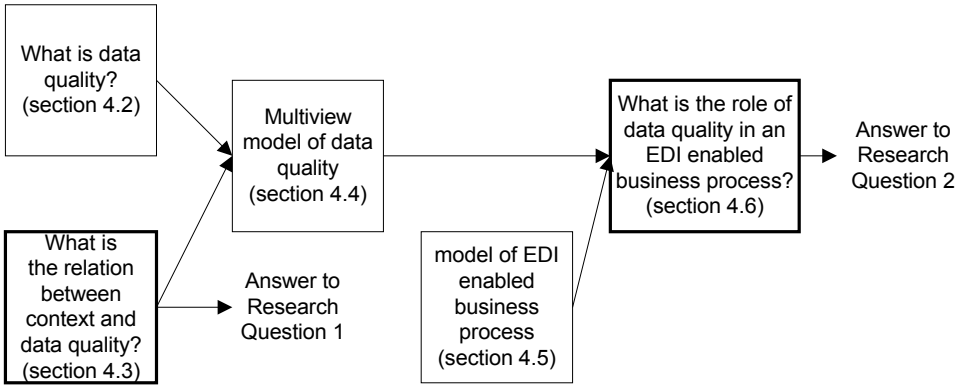


Figure 4-2: Chapter structure

First, in Section 4.2 we will start with a review of the data quality literature. From this literature we will learn there are two views of data quality, the reliability view, and the relevance view. However, the data quality literature does not offer an explanation for the relation between context and data quality. Therefore, in Section 4.3 we will extend our search to the communication literature, because communication theory is also concerned with data quality. From this literature we will develop a third view of data quality, which is the context view of data quality. Based on this view, we are able to answer the first research question, namely: what is the relation between context and data quality?

In Section 4.4, we will combine the quality views of the data quality and communication literature into a new model of data quality, which we will refer to as the multiview model of data quality. Next, in Section 4.5 we will provide a more precise description of the EDI communication process to include the context. Finally, in Section 4.6 we will combine the multiview model of data quality and the new model on the EDI communication process to answer the second research question, namely what is the role of data quality in an EDI enabled business process?

4.2 What is data quality?

Through the years, the concept of data quality has shifted from the reliability of data to the relevance of data. We will discuss both views and then combine them in a shared view of data quality.

4.2.1 Reliability view of data quality

The concept of data quality is first presented in the accounting and auditing community, where it is part of the system of internal control. Internal control comprises:

> *"The plan of organization and all of the coordinate methods and measures adopted within a business to safeguard its assets, check the accuracy and reliability of its accounting data, promote operational efficiency, and encourage adherence to prescribed managerial policies" (Committee on Auditing Procedure, 1949)*

In this traditional accounting view, data quality is primarily concerned with the reliability of the data, which means: the degree of agreement of account entries with what happened in reality (Starreveld 1976, 1994). Reliability consists of three dimensions (Starreveld 1994):

- *Correctness*, which relates to the degree that account entries conform to real events in reality.
- *Completeness*, which relates to the degree that an account balance contains all events that occurred in a certain time period.
- *Timeliness*, which relates to the degree that the time of entry of account entries is in conformance to the real time an event occurred.

The traditional view of data quality is shown in Figure 4-3. In this view, an accounting system processes data. The quality of this data is primarily defined as the degree of conformance to reality.

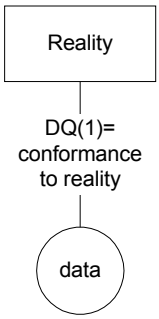


Figure 4-3: Reliability view of data quality

Same or similar views on data quality can be found in Weber (1999, pp. 851-877) and Cushing (1982 pp. 75).

4.2.2 Relevance view of data quality

Starting in the eighties, three new models of data quality appeared, which we will discuss below.

Relevance model
In 1980, the Financial Accounting Standards Board publishes the *hierarchy of accounting qualities* (FASB, 1980). This hierarchy introduces the *users* of accounting information who evaluate accounting information on its decision usefulness. Decision usefulness consists of two dimensions (Chandra et al. 1998):

- *Relevance*, which is defined as the capacity of information to make a difference in a decision by helping users to form predictions about the outcomes in the past, present and future or to confirm or correct prior expectations.
- *Reliability*, which is defined as the quality of information that assures that information is reasonably free from error and bias and faithfully reflects what it purports to represent.

Evaluating the relevance model, we see that the concept of relevance is added next to reliability.

'Fitness for use' model
In the beginning of the nineties, a similar approach is adopted in the information systems (IS) community. Several studies in the IS community showed that large organizational problems were the result of insufficient data quality. For instance, a study at the criminal justice department showed that 50-80% of computerized criminal records in the US were found to be inaccurate, incomplete or ambiguous (Laudon 1986). Another study reported that Dun & Bradstreet paid $350,000 to a construction company after it incorrectly reported that the company was bankrupt because a Dun & Bradstreet employee had entered inaccurate data into its credit database (Percy 1986). Redman found that tens of millions of dollars were spent on a system, whose only function was to verify the accuracy of each telephone bill (Redman 1995).

To solve these problems, Wang & Strong (1996) propose a different approach to data quality. Following the quality literature (Juran et al. 1974, Deming 1986) they propose to take the consumer viewpoint of 'fitness for use' in conceptualizing the underlying aspects of data quality. In contrast with the reliability view of data quality, the fitness for use approach argues that data quality cannot be assessed independent of the people who use data: the data consumers (Strong, Lee & Wang 1997).

However, since there are many users, each with their own definitions of data quality, it results in many dimensions of data quality (see for instance Zmud 1978, Taylor 1986, Fox, Levitin & Redman 1995, Wand & Wang 1996). Therefore, Wang & Strong developed a conceptual framework for data quality, based on a two-stage survey and sorting study. This framework is shown in Figure 4-4. The salient part of this study is that data attributes were collected from actual data consumers, instead of being defined theoretically or based on researcher's experience.

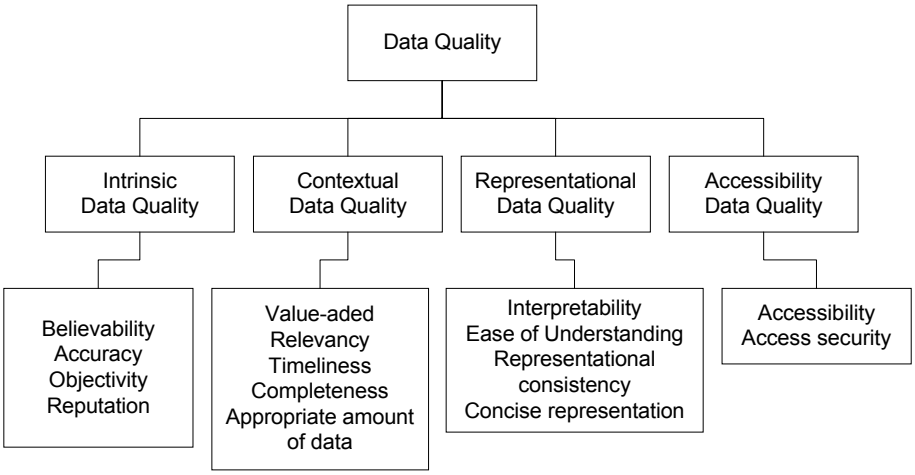


Figure 4-4: The conceptual framework of data quality (Wang & Strong 1996)

Evaluating the 'fitness for use' model of data quality, we see that in this view the user determines data quality.

Causal and teleological model
The causal and teleological models of data quality are introduced in the nineties in the Dutch accounting literature (van der Pijl, 1993). According to van der Pijl, two views on the quality of information exist: the causal model and the teleological model of information quality. In the causal model, information quality is the result of a chain of activities that together influence the quality of information. These activities can be grouped in two phases: the information system development phase and the information system operation phase. The importance of the causal model of information quality is that it is not possible to measure all aspects of the quality of information only from the information itself. The reliability of information is also depending on the measures that are taken in the IS development and operational phase.

In the teleological model the quality of information is determined by the objective for which the information is intended to be used. It is argued by van der Pijl that information depends on personal objectives, which in their turn (partly) depend on organizational objectives. The importance of the teleological model is that it introduces organizational objectives next to personal (e.g. user) objectives in the concept of data quality.

Evaluating van der Pijl's approach to data quality, we see that he provides two models that explain how the reliability and the relevance of data quality are affected.

Evaluation of all three approaches
Evaluating all three approaches, we find that they all include the users' perspective to explain data quality. Since data is used, the quality of the data largely depends on the specific use of the data, apart from the intrinsic quality of the data. This 'relevance' view of data quality is shown in Figure 4-5.

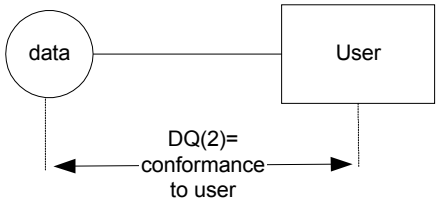


Figure 4-5: Relevance view of data quality

4.2.3 Conclusions

From the data quality literature we draw two conclusions. Firstly, according to the data quality literature, data quality consists of two views: the reliability view of data quality, and the relevance view of data quality. The reliability view defines data quality as conformance to reality. The relevance view defines data quality in terms of conformance to the users'

expectations (=fitness for use). Although both views offer a different perspective on data quality, they do not explain the role of context in data quality.
Secondly, using the reliability and relevance views on data quality, we are only partially able to explain the relationship between EDI and data quality, because EDI increases both the reliability and the relevance of the information in an EDI message. The reliability of the information is improved, because EDI leads to fewer data entry errors. Furthermore, the relevance of an EDI message is improved because EDI leads to a *complete* specification of the EDI information and to faster communication of the information. Since relevance is defined in terms of completeness and timeliness of information, EDI use increases the relevance of the information in the EDI message. Thus, EDI positively influences the reliability and the relevance of the information in an EDI message, and hence positively influences data quality.

4.3 The role of context in data quality

Since data quality theory does not explain the role of context in data quality, we will have to look in other scientific fields. Therefore, we chose the communication literature. First, because communication and information (or data) are closely related. Klooster et al. (1978, pp. 35) discuss several authors who argue that an information and a communication theory are distinct theories as well as that they should be considered as one single theory. Second, because EDI is a communication process, and the problems that resulted from the context were related to EDI. We will give an overview of the communication literature, after which we will show that this literature offers a third view of data quality: the context view of data quality.

4.3.1 Objective view of communication

In 1949, Shannon & Weaver introduced a mathematical theory of communication in which they provide a probabilistic treatment of the transmission of signals through noisy channels (Shannon and Weaver 1949). They give a symbolic representation of the communication *system*, which is shown in Figure 4-6.

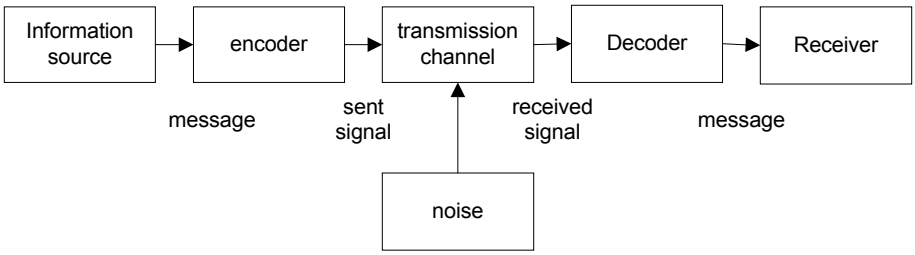


Figure 4-6: The communication system

In this system, communication is seen as the transmission of messages in the form of signals through a noisy channel. The objective of the communication system is to reconstruct the original message as good as possible at the receiver.

In the eighties, the International Standardization Organization introduces the Open System Interconnection Reference model, (the ISO/OSI model, see for instance (Tanenbaum 1989). This model provides a layered architecture for electronic interconnection over computer networks. Each layer provides a specific service, independently of the other layers. In the lower layers, the physical connection between computers is established, protocols handle the links between hubs in the network, error detection and correction is provided, and messages are routed through the network. On the higher levels of the architecture, the encoding and decoding of messages using a standard syntax is handled. The ISO/OSI model is an important extension of the communication theory, because the layered architecture creates transparency for each higher layer. This means that on each layer different protocols can be implemented, without having to worry about the impact of the specific protocol on the other layers.

An important characteristic of both theories is that both the content of the message (what does it mean?) and the objective of the message (What reaction is expected from the sender?) are not really considered. Both theories focus on transmitting the message without errors, assuming that the interpretation of the message is unequivocal, and hence objectively given.

4.3.2 Intersubjective view of communication

The semiotic framework

In 1992, based on the theory of Signs (Semiotics), Stamper introduces the Semiotics Framework (Stamper 1992, Huang 1998). This framework describes six levels of communication between two subjects (persons or organizations) that operate in a different context. When messages (for instance, EDI messages) are exchanged, agreements on the first five levels are necessary to guarantee successful communication (on the sixth level). A slightly adapted version of the Semiotic Framework is shown in Figure 4-7.

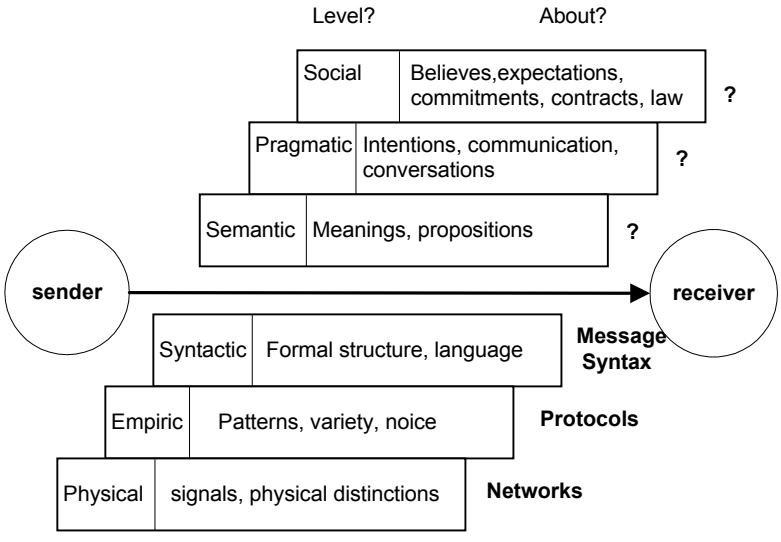


Figure 4-7: The semiotic framework (adapted from Huang 1998)

We will discuss each of the levels shortly:

- On the *first* level, the physical connection is established. This means that the physical devices and their connections are specified. For electronic communication, on this level agreements about the type of network connection that is used are established.
- On the *second* level, the communication of a signal over the physical connection is established. This means that a signal at the sender can be reproduced without error at the receiver. For electronic communication, on this level agreements about the network protocols are established.
- On the *third* level, agreements about the syntax of the signal (e.g. the message) are established. Syntax relates to the structure of a message: which data-elements can be used, and how and in which order are they displayed (for example: name, address, domicile). The syntactic level can be compared with a grammar dictionary that specifies which words are available in a language and how sentences are constructed.
- On the *fourth* level, agreements about the semantics of the signal (or message) are established. Semantics relate to the *meaning* of the data in the message. Here, the relation between data elements with other elements is specified and constraints are defined. This level can be compared with data dictionaries in Database Management Systems, where the conceptual data model is constructed and agreements about for instance the range of product numbers or the definition of turnover are established.
- On the *fifth* level, agreements about the *pragmatics* of the message are established. Pragmatics is concerned with the *intention* the sender has with sending the message. In electronic communication, these intentions are normally specified in procedures. For example, when company A sends an order, company B must understand that company A wants one of their products. Furthermore, they must understand that company A wants them to react through returning an order confirmation. Thus, on this level, agreements about procedures are established.
- On the *sixth* level, the signal should result in a change in the social world, e.g. a change in the believes, expectations, commitments, contracts, laws or culture of which the social world is constructed.

The importance of Stamper's Model
The importance of Stamper's communication model is not that he adds three new layers to the ISO/OSI model (the semantic, pragmatic and social level). The importance is that Stamper views information as intersubjectively defined, and not as an objective representation of reality.

To Stamper, information is a sign, which stands for a physical reality, *according to an interpretant* (see Figure 4-8).

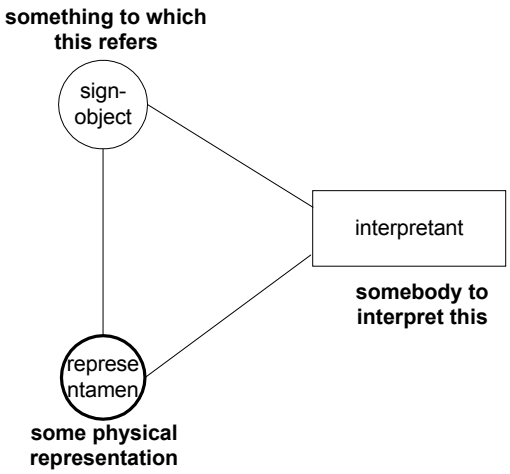


Figure 4-8: The theory of signs

An interpretant interprets a sign following the norms and rules of the social group the interpretant belongs to. Examples of such social groups are: a work department, an organization, the company soccer team, a dinner table group, etc. The norms of the social group are implicit rules that describe how things are (semantics) and how things should be done (pragmatics).

When two parties are communicating (which means that they exchange signs), the success of the communication depends on the degree that a shared norm system is established. Only if the communicating parties share the same meanings (semantics) and intentions (pragmatics), a sign can be faultlessly interpreted by each party. Thus, in this view, information (i.e. a sign) is not the representation of an objective reality, but depends directly on the shared norm system (which is intersubjectively defined). Because the shared norm system is established in the context of the communication process, *we will refer to the shared norm system as the context* (of the communication).

4.3.3 Context view of data quality

Using the intersubjective view of communication, we are now able to answer our first research question, which was: what is the relation between context and data quality?

Since, according to Stamper, information depends directly on the shared context, the quality of information will depend directly on the quality of the communication context. We will define this quality of the shared context as the context quality. Context quality depends directly on the degree that the contexts of the communicating parties are *aligned*. This new view of data quality is shown in Figure 4-9.

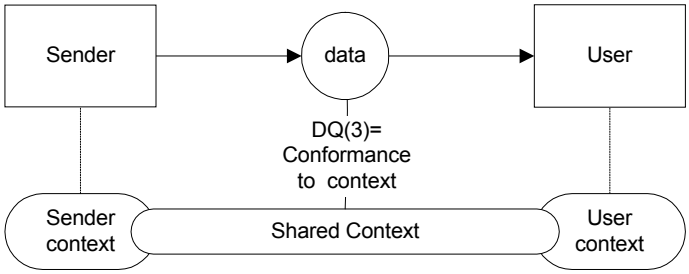


Figure 4-9: Context [1] view of data quality

As we can see from Figure 4-9, data quality is defined in relation to the shared context. Specifically, we will define data quality as the degree of conformance to the shared context.

Since we have introduced the concept of context quality, which was defined as the quality of the shared context of the communication process, and since data quality (of the EDI order) was defined as the degree of conformance to this shared context, data quality (of the EDI order) directly depends on the quality of the context. Hence, the answer to our first research question is:

> *The relation between context and data quality is that the quality of the (shared) context (which is defined as the context quality) positively influences the quality of the EDI order data in the EDI-communication process.*

The difference between data quality and context quality might be confusing. Therefore, we will illustrate this difference with an example. Suppose that the context quality is 70%, and the data quality is 100%. How is this possible? Did we not define data quality as conformance to the context, which would imply that data quality is also 70%? The answer starts with a good interpretation of the numbers. A context quality of 70% means that 70% of the product records between sender and receiver are equal. A data quality of 100% of the EDI orders means that 100% of the EDI orders are 'correct'. If we use the context view of data quality, 'correct' means that none of the EDI orders lead to interpretation problems, and therefore are correctly processed. In this case, the difference between 100% and 70% may be explained as follows. Seventy % context quality means that 30% of the product records do not match, perhaps because the receiver still uses outdated product codes. However, when the exchanged orders between sender and receiver are only about current products, then the degree of data quality of the EDI orders is 100%.

4.4 A multiview model of data quality

In Figure 4-10 we have combined the three views on data quality in a multiview model of data quality.

---

1 This concept of context quality should not be confused with the concept of contextual data quality of Wang & Strong. Our definition of context quality focuses on the degree of alignment of the sender's and the user's contexts, while contextual data quality focuses on the context of the user.

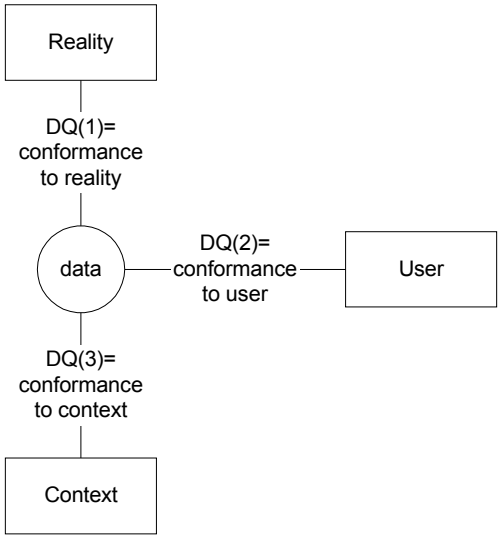


Figure 4-10: Multiview model of data quality

Figure 4-10 shows that three different perspectives on data quality exist, the reliability view, the relevance view, and the context view. The reliability view of data quality defines data quality in terms of conformance to reality. For instance, suppose that the data about a physical product describes its size as 5x5x5 cm, when the product really measures 5x5x5 cm, the data quality is according to the reliability view, 100%.

The relevance view of data quality defines data quality in terms of fitness for use. For instance, suppose that a user wants to place an order and that the user learns that the product measures 5x5x5 cm. When the user did not receive the price and the delivery time for that product, then the data quality is 0%, according to the relevance view. Although the information concerning the product is 100% correct.

The context view of data quality defines data quality in terms of conformance to the shared context. For instance, suppose that a user needs to know the measures of a product to place the product onto a shell in the supermarket, and suppose the user does not know that the product size is exclusive of packing. When the user learns that the product measures 5x5x5 cm, but then learns that it does not fit the reserved space on the shell, then according to the context view, data quality is 0%, because the context of the user and the supplier supplying the information is inconsistent.

When we compare the multiview model of data quality with the conceptual framework of data quality from Wang & Strong, we see that the four different dimensions of the data quality framework of Wang & Strong are similar to the three perspectives in the multiview model. The intrinsic quality dimension is similar to the reliability of the multiview model. The contextual data quality dimension is similar to the relevance view of the multiview model. The representational quality dimension is similar to the context view. Lastly, the accessibility dimension of data quality is also similar to the relevance view of data quality.

Now that we know the relationship between context and data quality, we are able to explain the role of data quality in an EDI-enabled business process. But before we do that, we will first provide a more precise description of an EDI-enabled business process.

4.5 An EDI enabled business process

Since we know that context plays an important role in a communication process, we can broaden the picture of the (EDI) communication process to include the context. We will do this by making a distinction between *transaction communication* and *data alignment*, as is shown in Figure 4-11.

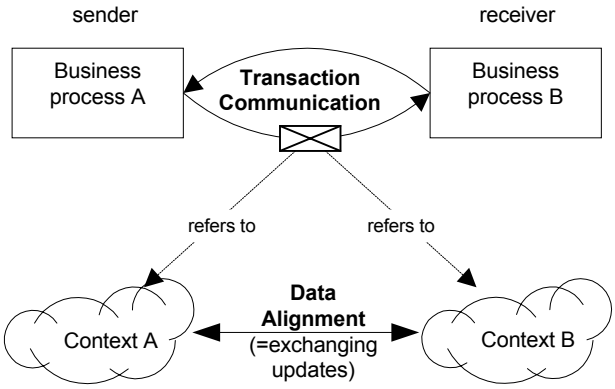


Figure 4-11: Transaction communication and data alignment

Normally, interorganizational business processes are coordinated by exchanging all kinds of messages, such as quotations, orders, manufacturing plans, shipping notifications and invoices. These messages are exchanged as part of the execution of a transaction between two or more organizations. We refer to this type of communication between interorganizational business processes therefore as *transaction communication.* An important characteristic of transaction communication is that whenever a message is received, the receiving process must react to it. Thus, transaction communication triggers the execution of another process. Since this implies that the state of the receiving process will change, we characterize transaction communication as *state dependent*.

The concept of data alignment was first introduced by Madnick (1995), who referred to it as context interchange. According to Madnick, a new problem for global information systems is surfacing, which he defines as the context problem. Madnick defines this problem as follows:

> *"Each source of information and potential receiver of that information may operate within a different context./…./ When the information is moved from one context to another, it may be misinterpreted."*

To solve this problem, Madnick introduced the concept of context interchange. This means that:

"As source or receiver contexts change, the necessary adjustments are made automatically allowing autonomous evolution of the individual systems."

These necessary adjustments are made through the communication of *updates* between senders' and users' contexts. Examples of these updates are information about products, services, clients and processes. This information is not exchanged to execute a transaction but to update each other's databases. These databases provide the *context* for successful transaction communication, since the content of the transactions *refers* to the information that is stored in the databases. For instance, a typical transaction exchange such as an order contains at least a customer identification number, a product number of the product to be delivered, the amount, and the price. Both the information about the customer (Who is it? Where does he live?) and the information about the product (What is it? How is it manufactured?) is not available in the transaction, but should be available in the information systems of both sender and receiver to ensure that the supplier can deliver the product. We will define the process of making agreements about mutual data and the exchange of updates between contexts to improve the shared context, as *data alignment*.

An important difference between transaction communication and data alignment is that context interchange takes place independent of the transaction, *without* changing the state of the receiving business process. The context information is stored in the receiver's database, where it may be useful in the future. Thus, data alignment is characterized as *state independent*.

The concept of data alignment plays an important role in solving the product data quality problem. We will address this issue further in Chapter 6.

4.6 Model of the role of data quality in an EDI business process

As we explained in the first paragraph of this chapter, our objective was to explain the role of data quality in an EDI enabled business process. Although we know that EDI use positively influences process performance, we are interested in the question how data quality, and especially context quality, fits into this model. Based on the overview of an EDI-enabled business process, and our multiview model of data quality, we can now explain the role of data quality in an EDI-enabled business process (see Figure 4-12).

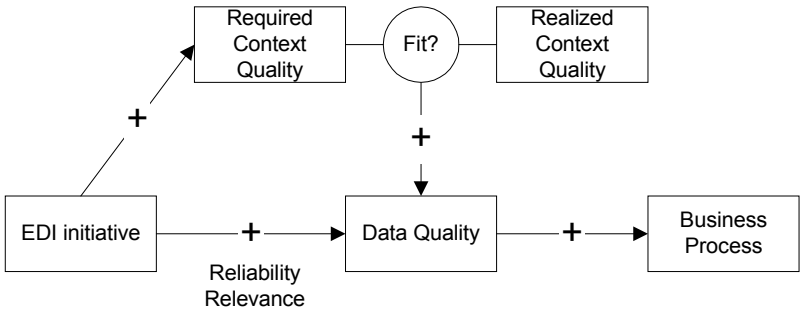


Figure 4-12: Model of the role of data quality in an EDI enabled business process

Figure 4-12 shows the same relationship between EDI, data quality and the effect on the business process as was shown in Figure 4-1. However, we expanded this model to include a direct and an indirect relation between EDI and data quality.

The direct relation between EDI and data quality is explained as follows. EDI positively influences the data quality of the information in an EDI message, because EDI increases both the reliability and the relevance of that information. The reliability of the information is improved because EDI leads to fewer data entry errors. Furthermore, the relevance of an EDI message is improved because EDI leads to a *complete* specification of the EDI information, as well as to faster communication of the information. Since relevance is defined in terms of completeness and timeliness of information, EDI use increases the relevance of the information in the EDI message. Thus, EDI positively improves the reliability and the relevance of the information in an EDI message, and hence positively influences data quality.

The indirect relation between EDI and data quality is explained as follows. An EDI initiative results in higher requirements for the quality of the context between sender and receiver, because EDI implementation eliminates human intervention in the transaction communication between business processes. As we learned from our first research question, the quality of the context positively influences the quality of the data in the EDI message. Without EDI in the business process, many errors in the transaction communication that are the result of insufficient context quality are detected and corrected through human interpretation. When EDI is implemented, this human correction mechanism is not in place. Hence, EDI results in a higher required context quality.

If the realized context quality matches the required context quality, then the high quality of the context positively influences the data quality of the EDI message, and therefore has a positive effect on the business process. However, when the context quality is not improved when EDI is implemented, the required context quality exceeds the realized context quality. Since context quality positively influences the data quality of the EDI message, the insufficient quality of the context will result in a decreased quality of the data in the EDI message, which will negatively influence the business process.

Although this model on the role of data quality in an EDI enabled business process provides a reasonable explanation, it has not been validated in practice. However, if we want to solve the problem of insufficient product data quality in an EDI enabled business process, we must be reasonably certain that our model is indeed correct. Since we are particularly interested in the effects of EDI use on the quality of the context (which we will operationalize as product data quality) and the resulting effects on the business process, we formulated the following proposition:

> If EDI is introduced, and the quality of the context is not improved, then this context quality will decrease the direct positive effects of EDI on the business process.

In the next chapter, we will conduct two case studies to verify this proposition about the effect of context quality on an EDI enabled business process.

5. Two Data Quality Cases

5.1 Introduction

In this chapter we will describe how we tested the proposition about the effect of context quality on an EDI enabled business process, which we formulated at the end of Chapter 4. To do this, we will first describe the overall research strategy for testing the proposition in Section 5.2. Since at the beginning of the test we did not know how to measure both context quality and impact on an EDI-enabled business process, we first conducted an explorative case where the main objective was to develop an assessment method. This case, which we will refer to as the SLIM case, is described in Section 5.3. The main result of this case, namely the assessment method, is described in Section 5.4. Based on the assessment method and the lessons learned from the SLIM case, we finally tested the model in the Schwartz case, which is described in Section 5.5. Finally, the conclusions of testing the proposition are presented in Section 5.6.

5.2 Research strategy

As a general research method for testing the proposition, we decided to use case studies. Yin (1994) provides three conditions for using a case study: (1) the type of research question posed (2) the extent of control an investigator has over actual behavioral events, and (3) the degree of focus on contemporary as opposed to historical events. According to Yin, in situations where the object of study can be controlled, an experiment is normally the best research method. Furthermore, when the object of study lies (far) in the past, archival analysis and Histories should be used. Since our object of study cannot be controlled, and since it lies not in the past, this leaves both the survey and the case study as preferred research methods. There are two reasons why we decided against the survey method:

1. Our objective is not only to find a measurable effect of the influence of context quality on an EDI-enabled process, but also to explain how, and under which conditions this effect appears. Following Yin, a How question is best studied using the case study method.
2. Although it is possible to measure process performance, it is virtually impossible to control for other influences other than context quality in a survey situation. In other words, how can we be sure that the measured impact on process performance is indeed the result of insufficient context quality and not poorly implemented systems, problems with user perceptions, worker motivation problems, etc. This problem is generally recognized in many IT impact studies (Brynjolfsson & Hitt 1998).

In addition, we decided to follow a strategy where we first conduct an explorative case and then a testing case. The main reason for doing this is that we need a detailed assessment method to measure both context quality and process performance. By first developing this assessment method in the first case, we could use the method for testing the proposition in the second case.

5.3 The SLIM case

5.3.1 Case setting

The case was conducted at a large Electrotechnical installer. The Electrotechnical installer is one of the three largest installation companies in the Netherlands with 1400 employees and a turnover that exceeds 150 million US dollars in 1995. In this case, we focused on the production department, which builds safety switches into switch-cupboards that are transported to a shipbuilding site. There the cupboards are installed into ships. About 35 employees work at the production department.

The case was originally conducted as part of a larger research project (The SLIM project, see SLIM-working group, 1997), where the researcher and two students focused on analyzing and solving the logistical and information problems of a complete supply chain (thus including supplier and wholesaler). The initial objective within this larger research project for this chain was to develop a first version of the Data Alignment through Logistics (DAL) method for a specific supply chain (see Chapter 8: the CBL case for this first version). However, during the chain analysis phase, it appeared that it was still unclear what exactly the impact of context quality on the installer's ordering process was and, hence, what the need was to solve this problem. Furthermore, it appeared that the poor ordering process performance resulted not only from insufficient context quality, but also because the ordering process was poorly designed. Since the latter problem was regarded more eminent, and because the company indicated that it regarded the context quality problem as a minor issue, it was decided to split the case. The student would focus on redesigning the ordering process. The research objective was altered towards measuring the impact of context quality on the ordering process. Both the student and the researcher carried out this part.

The shift was considered viable since it emphasizes the need to explain the role of context quality in interorganizational business processes. This need is also expressed in the first two research questions (see Chapter 1). Also, the case setting was considered to be still viable to the new objective, because the chain analysis provided some indications that context quality problems existed at least at the installer's:

"The throughput times of catalogs are long/……/This means that the quality of the product data in the systems of the installers is not very high. (Montfort 1997:1, pp. 13)".

5.3.2 Case design

For describing the design of the case, we used the criteria for case studies as defined in Chapter 2.4.

Type of study
Since this was the first time we researched the impact of context quality on an IT enabled business process, and since there is only one level of analysis, the case is classified as *exploratory* and *holistic (*for explanation of these terms, see Section 2.4).

Case objective and case questions
We defined the objective of this case as follows:

> *Measure the impact of context quality on the ordering process of the installer to determine a first version of an assessment method for the impact of context quality.*

Consistently with the research objective, we will operationalize context quality as *the quality of the product information* (e.g. product data quality) in the installer's database.

Based on this objective we formulated the following case questions:

1. Which factors determine product data quality?
2. How can we measure product data quality?
3. Which factors determine order process impact?
4. How can we measure order process impact?

Case choice criteria
Since at that time we did not focus specifically on EDI, but more generally on IT, we needed a company with indications of insufficient process performance and low product data quality in an IT enabled business process. We found strong indications of both low product data quality and insufficient process performance in the issue analysis at the installer. Furthermore, since the complete ordering cycle depends largely on the use of the information system (for ordering, warehousing and invoicing), we decided that the case was IT enabled.

Unit of analysis and resulting case structure
In the case, we identified two units of analysis that we examined as a whole: The article database, the ordering process and their interrelationship. The smallest unit of the ordering cycle we identified is a process step. The largest unit is the ordering process itself. The smallest unit of the database we identified is a data field. The largest unit is the complete article database for the safety switches. Both units are analyzed only for the safety switches. Thus, only the product data quality of the safety switches and the ordering performance of safety switch orders are measured. Although we have two units of analysis, we measure them on only one level. Therefore we classified the structure of this case as holistic.

Criteria for interpreting findings
Since this case is explorative, we will not test theory, and therefore we do not need criteria for interpreting our findings.

5.3.3 Case Methodology

We used the following procedures to obtain the results.

1. Measuring context quality. To measure context quality, we followed the data field measurement tests of Weber (1999) for measuring product data quality, which means that we compared the product data in the installer's database with the product data of the safety switch supplier. Especially, we chose to compare the article data on accuracy, timeliness and completeness of records. Here, completeness of records is defined as the degree that the installer's database contains all article records the data source contains and vice-versa. Accuracy was tested by comparing the actual contents of the product data fields of both supplier and installer. Timeliness was tested by estimating the obsoletion rate per month. Finally, record completeness was tested by comparing all article records of the supplier's product at the installer, not available at the supplier.
2. Defining the ordering process. To define the ordering process, we first interviewed all people in the installer's staff who are involved in the ordering process. Next, we together with the staff established the scope of the ordering process. The ordering process was defined as the moment an order entered the ordering department (from engineering) until the moment the order was sent to the supplier.
3. Measuring impact. Although many different performance measures exist, several authors agree on three performance measures for an ordering process:

    a. The labor time spent on processing an order;
    b. The throughput time of an order;
    c. The remaining number of errors in the process (the quality of the process), see Strong (1997), Mukhopadyay et al. (1995).

    Since the ordering staff indicated that throughput time did not play an important role, we decided to focus on labor time and number of remaining errors in the ordering process. Labor time was measured by following 10 safety switch orders through the complete ordering process. The remaining number of errors was measured by way of an invoice analysis, where we only counted errors that could be directly related to the ordering process (Hence, warehousing errors such as putting a product at the wrong location were excluded.

5.3.4 Case results

Assessment of context quality
Since the manufacturer was already sending price information on diskette to the installer, we used the information on these disks for the comparison (Montfort 1997:1, pp. 32). Table 5-1 gives an overview of the results of the context quality tests.

context Q category	context Q subcategory	context Q score
Completeness of records	Installer – manufacturer	74%
Accuracy	Article numbers	100%
	Descriptions	0%
	Prices	100%
	Price units	100%
Timeliness		99%

Table 5-1: Results context quality tests

First, comparing the article records from the installer with the manufacturer, it appeared that only 74% of the installer's records were available at the manufacturer. Further investigation showed that this was the result of articles that were obsolete according to the manufacturer, but that were still active in the database of the installer.

Second, we measured the accuracy through comparing the four fields of the price diskette. As we can see from Table 5-1, all fields matched, except for the article descriptions. The main reason was that all descriptions were modified because the manufacturer used 35 positions while the installer's description field only contained 23 positions, which meant that the descriptions were altered. Normally this will not cause problems, since the purchasers know the articles quite well. In case of doubt they consult the manufacturer's product catalogue.

Third, we measured timeliness through determining the obsoletion rate. It appeared that in the beginning of 1997, 19.6% of the products in the catalogue were obsolete and 15.3% were new, compared to the catalogue of 1995. This means that almost 35% of the product information had changed in two years. If we assume that the same mutation rate applies to the articles on the diskette that are sent each month, we can establish that at most 19.6% / 24 months ≈ 0.8% of the articles becomes obsolete per month.

Assessment of impact of context quality

To assess the impact of context quality on the ordering process, we first defined the ordering process at the installer (see Figure 5-1). As we can see from Figure 5-1, the ordering process is defined from the moment that a materials requirements list from engineering enters the ordering department, until the moment that the order is sent to the supplier.

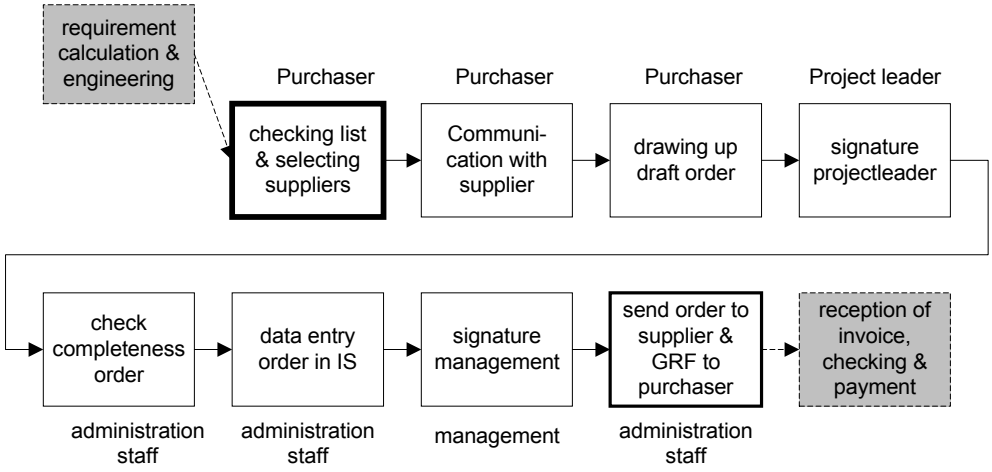


Figure 5-1: The ordering workflow of the installer

Time spent on human labor
To measure the time spent on human labor, we followed 10 orders of safety switches through the complete ordering workflow. The results are shown in Table 5-2 (see v. Montfort 1997:1, pp. 40).

	Steps in process	Who?	Time per order (minutes) per order	per line*	total
1	Selection & checking	Material Purchaser	2	0.5	4.5
2	Communication with supplier	Material Purchaser	2	0.5	4.5
3	Drawing up draft order	Material Purchaser	1	0.5	3.5
4	Signature project leader	Project Leader	3		3
5	Checking completeness order	Data Administration	2		2
6	Data entry & final checking order	Data Administration	3	2	13
7	Signature Management	Management	3		3
8	Sending formal order	Data Administration	4		4
	Total costs				**37.5**

* 5 lines per order on average

Table 5-2: Total labor time per order

Table 5-2 shows that the total labor time per order was 37.5 minutes.

Measurement of number or remaining errors
The analysis of all order lines in 1996 showed that 14% of all order lines needed a correction (From 235.417 order lines 31.412 lines were corrected, see van Montfort 1997:2, Appendix 4.5, pp. 2). Hence, the percentage of remaining errors after the ordering process is 14%.

However, to measure the remaining number of errors *resulting from context quality problems,* we used an invoice analysis, conducted in October 1996 (see Appendix B). For this analysis a sample of 86 orders with 335 order lines was taken out of 53,236 orders, which represented all orders of all 7 ordering departments of the installer. Because the ordering departments were all working very similar, using the same ordering software, we concluded that the results could be used for our purpose. The invoice analysis revealed 31 different causes for all types of errors. We analyzed the percentage of causes that were the result of insufficient *product* data quality in the database (see appendix B). We found that 14% of the causes were the result of insufficient product data quality. Hence, the estimated percentage of errors after the ordering process as a result of insufficient product data quality is 14% * 14% = 1.96%.

5.3.5 Interpretation of results

The overview of the data assessment test (see Table 5-1) shows the context quality is moderate. Although the low scores on article descriptions are not relevant to placing an order (since the ordering system will not use the article descriptions to process an order), the medium scores on record completeness (74%) could impact the ordering process. However, we expect this impact to be moderate.

The impact on the ordering workflow showed that the time spent on labor is long, namely 37.5 minutes. In contrast, the wholesaler in the mini-chain required only a few minutes to process a complete order. We could explain this long processing time from the moderate data quality score of the product data. However, there are two problems with doing just that:

1. We do not know for sure that the long processing time of 37.5 minutes is the result of low context quality or other effects, because we have not connected the activities in the ordering process directly to the quality of the product data.
2. If the moderate context quality explains the long processing time, we would, following our other assumptions, also expect a large number of remaining data quality errors. However, an estimated 2% of data quality errors in invoice lines, compared to an industry average of 2% (for *all* invoice lines), is not exceptionally high.

Thus, based on these results, we could not explain the long processing time with the relatively low remaining error rate (due to insufficient data quality) from the measured moderate context quality. Therefore, we decided to re-examine both context quality and the ordering workflow to explain the results.

5.3.6 Re-examination of results

Re-examination of context quality
From a detailed problem analysis of the ordering process, we learned two important things. Firstly, further interviews with material purchasers and a study of the order form showed that the *delivery time* appeared to be missing on the form. Therefore, the material purchaser always had to make a phone call to ask for the exact delivery times. Secondly, we found that users did not *trust* the system, because of former implementation problems. (see van Montfort 1997:1, pp. 20). The result of both problems was that the project purchasers hardly used the article database of the ordering system to process an order. An analysis of the orders that were placed

at the supplier showed that only 1% of all order lines with our supplier were processed automatically, which means that the article database was used in processing the order (van Montfort 1997-1, pp. 30). In the other cases, a phantom article number was used to process the order. From the feedback of these results to the material purchasers we learned the fact that they always had to call the supplier and that they had problems with using the system resulting in the manual processing of the orders. The consequence is that we are not able to make a direct connection between the quality of the product data (= the context quality) to the ordering process, because the database is hardly used.

Re-examination of process impact
The problem with measuring the impact on the ordering process was that we could not explain the long order processing time of 37.5 minutes together with the error rate of only 2%, where we expected, on the basis of consistency, a much higher error rate. Therefore, we decided to take a closer look at the ordering process. A further analysis of the steps in the ordering workflow and the time spent per step showed that a considerable amount of time was spent on checking activities. Since checking activities is a form of error correction, but then inside the process, we concluded that the direct impact of insufficient context quality should be assessed from the change in the labor time spent on checking activities. Only when we measure the time spent on preventive activities, we measure the ordering time *that is the direct result of context quality problems* (otherwise we would not have to attempt preventing measures). However, in this case we were not able to calculate the exact influence on time spent on checking activities, because from the data we collected, we could only calculate the total amount of time spent on checking activities. We did not know how much time was spent on which type of checking activities.

5.3.7 Case conclusions

The objective of the case was to measure the impact of context quality on the ordering process of the installer to determine a first version of a context quality impact assessment method. The results in the previous section lead to the following two conclusions:

1. Weber's field and record comparison tests provide a sufficient operational basis to measure the context quality of the installer's product database. However, measuring context quality in this way, proved not to be sufficient for establishing the impact on the ordering process, because it appeared that the database was hardly used. Hence, a necessary check to establish the context quality of the product database is to measure the degree that the ordering process actually uses the product database.
2. Although both labor time and remaining number of errors (as a measure for EDI order data quality) appear to be good measures to assess order process performance, they proved to be incorrect measures for measuring the impact of insufficient context quality on order process performance. The main problem with both measures was that they could not be linked directly to insufficient context quality, and therefore could be the result of other influences. However, through measuring labor time as the time *spent on checking activities* and through measuring the remaining number of errors as *the error rate of invoices that are the result of errors in the product data*, we are able to make a direct connection between context quality and its impact on the ordering process.

5.4 A context quality impact assessment method

Based on the results of the SLIM case, we defined a first version of a method to assess the impact of context quality on an ordering process. This method consists of the following steps:

1. Assess the context quality through comparing the product databases of sender and receiver using Weber's field and record data quality tests.
2. Assess the degree with which the ordering process actually uses the receiver's database, as a necessary condition certifying that the measured context quality affects the data quality of the EDI orders in the ordering process.
3. Define the different steps in the ordering process and establish the beginning and the end of the ordering process.
4. Measure the total labor time of the ordering process as the time spent on checking activities (due to context quality errors).
5. Measure the number of remaining errors (=EDI order data quality) as the error rate of invoices that are the result of errors in the product data.

We will now use this assessment method to test our proposition about the effect of context quality on an EDI enabled business process as described in Section 4.6.

5.5 The Schwartz Case

5.5.1 Case setting

The case was conducted at E. Schwartz B.V, which is a Dutch wholesaler that supplies ball bearings and other technical materials to the industrial market. E. Schwartz B.V. is a typical medium sized company with 40 employees and a yearly turnover of 13 million US dollars. In the Dutch technical materials sector, where EDI is virtually non-existent, E. Schwartz B.V. successfully implemented EDI with two large suppliers and several customers.

5.5.2 Case design

In describing the design of the case, we used the criteria for case studies as defined in Section 2.4.

Type of study
Since the case design is based on two levels, and since the objective it to test theory, we classified this case as *explanatory* and *embedded*.

Case objective
In this case we wanted to test the proposition about the effect of context quality on an EDI enabled business process. Hence, the objective of the case was defined as follows:

> *Test our proposition, which explains that the realized context quality will decrease the direct positive effects of EDI order process performance.*

This objective was further restricted as follows:

1. We will use the assessment method we defined in the SLIM case to measure the impact of context quality.
2. Similar to the SLIM case, we will operationalize context quality as the quality of the product data that is used in the EDI orders.

Case choice criteria
Since the case required a company that used EDI, we added this as the main choice criterion. Furthermore, we decided to look for a company with similar characteristics as in the SLIM case. That is, an article database that is used in the ordering process, a repeatable ordering process and long term relationship with suppliers. E. Schwartz B.V. complied with all of those criteria.

Case design
The main issue in this case was how to design the case to test our proposition. The best way would be to measure the product data quality at a company and then determine what the process performance with and without EDI would be under varying product data quality (or context quality) levels. However, this is rather difficult, since we cannot vary the context quality level within one company. Therefore, we chose to examine two nearly identical ordering processes, which differ in only one aspect, namely the use of EDI in the ordering process. This resulted in a design that consists of two mini-ordering cycles between E. Schwartz B.V. and an EDI and non-EDI customer. Therefore, the case design could be characterized as a quasi experiment, where all factors are kept equal except for the use of EDI. The advantages of this design are that it enables us to test:

- Differences in order processing time as a result of using EDI, and hence an extra check that the positive effects of using EDI applied in this case.
- Differences in prevention effort and the remaining error rate to confirm or falsify our proposition.

A second important design aspect of the case concerns the borders of the ordering system. To assure a valid measurement of the impact of product data quality, we chose to include the customer of E. Schwartz B.V. within the ordering cycle (and hence the case) because the measured context quality is a result of the combined product data quality of both E. Schwartz B.V. and the customer. Errors on both sides result in the measured context quality. Suppose that we measure, through a comparison of product data records, a context quality between E. Schwartz B.V. and the customer of only 20% and that we would focus only on the customer (as we did in the SLIM case). In that case, this would result in no impact on both order processing time and remaining number of errors at the customer, although E. Schwartz B.V. is confronted with an error rate of 80% in the received order lines (if we assume that the articles in the product database of the customer are evenly used in the order lines). This would mean that we would measure a context quality of only 20%, while the impact on the ordering process (at the customer) is negligible. The same argument holds the other way around, when we only measure the impact of insufficient context quality on E. Schwartz B.V. Because in that case, E. Schwartz B.V. would not know how many errors in the order were already corrected before they receive the order. An example is an order of a product that the customers' purchasing department already knows about, but which information was not processed into the customers' article

database, resulting in a matching error. To include both the sales order department at E. Schwartz B.V. and the purchasing department at the customer, these problems are prevented.

Criteria for interpreting findings
These criteria should indicate under which circumstances we are willing to falsify our proposition about the impact of context quality. This would be the case, if the assessment method would not find any evidence of changes in the amount of prevention and/or correction, although there is a difference in the quality of the shared contexts of both the EDI and non-EDI ordering processes. Therefore, we identified the following criterion:

> *If we do not find a substantial relative difference in processing time due to more prevention effort and/or a substantial increase in the number of corrective errors after the order is processed, while the context quality of the EDI enabled ordering process is equal or less than the quality context of the non-EDI enabled business process, our proposition is rejected.*

5.5.3 Case methodology

The case methodology was based on the assessment method as described in Section 5.4. Therefore, we will shortly review the steps while specifying how and where we collected the specific data[1].

Steps 1 and 2: Assess context quality and actual use
The context quality was assessed in the same way as we did in the SLIM case. Both E. Schwartz B.V. and the EDI and non-EDI customers supplied a download of their respective product databases on floppy disk. Here, the article selection criterion was that all articles were selected containing references to the other party (for instance through references of the article to pricing conditions with the other party). This selection criterion was valid, since both E. Schwartz B.V. and their two customers could not process orders (manually or automatically) without a reference to a pricing condition. Actual use was tested only at E. Schwartz B.V. Here we measured the rate of customer order lines actually using article numbers that were available in the Schwartz product database, so that the order lines could indeed be automatically processed (or semi-manually in case of the non-EDI customer).

Step 3: Defining the ordering process
The ordering process was defined together with the order administration staff, which resulted in an overview of the ordering process consisting of 6 and 7 sub-steps respectively (see Vermeer 1998:1, pp. 24-26). Here, we explicitly separated sub-steps that consisted of preventive work from other work, so that measurement of the time spent on preventive actions was possible. Furthermore, we focused on defining the process as small as possible, starting at the moment that an order was generated at the customer, until the moment that the order was accepted and stored in the order database of E. Schwartz B.V.

[1] For a complete overview of the case action plan, we refer to the Schwartz report (Vermeer 1998:1).

Step 4 and 5: measuring labor time, preventive actions and remaining error rate
The complete order processing times (in terms of amount of labor) were measured through following 10 orders, consisting of 173 order lines (=5% of all order lines annually) and 55 order lines (= 3% of all order lines annually) for a period of three weeks in 1997. For each order in the sample, the administration staff filled in how many minutes they needed to process the complete order per administration sub-step. Labor time spent on preventive activities was measured through summarizing the sub-steps that contained preventive actions. Finally, the remaining error rate (which is a measure for EDI order data quality) was measured through an invoice analysis of all invoices in 1997 of both the EDI and non-EDI customers. Since each invoice error at E. Schwartz B.V. contains a full description of what kind of error it was, we were able to select those invoices that measured errors as a result of insufficient context quality.

5.5.4 Case results

Selection of the two customers
We selected two customers for the case, a cigarette manufacturer and a ship handler. The cigarette manufacturer is the customer who uses EDI to process its orders. It orders about 4000 different articles (such as maintenance and spare parts) using a re-orderpoint strategy. This means that orders are completely electronically generated. The ship handler is the customer who is not-EDI enabled. It orders about 2000 different articles that it supplies to ships that enter the harbor and have to reload. Since every purchasing order of the ship handler is preceded by a quotation, all information for placing the order is already available in the ordering process. Therefore, the ship handler has started to implement EDI in its ordering process, which should result in direct orders that are driven by the quotation process.

Assessment of context quality
The results of measuring both the quality of the product data and the actual use are displayed in Table 5-3. We will comment on these results shortly.

Context Q category	Context Q subcategory	Schwartz & EDI customer	Schwartz & non-EDI customer
Accuracy	Article numbers	0%	0%
	Net price differences	5.8%	6.8%
Record completeness		18.0%	4.0%
Actual non-use		<9.8%	<0.3%
Total % of quality probl.		23.9%	10.9%
Context Quality		**76.1%**	**89.1%**

Table 5-3: Data quality for the ordering process with 2 Schwartz customers

For the comparison of field tests, we compared *article number customer, description, description E. Schwartz B.V., article number E. Schwartz B.V., gross price, discount percentage, packing unit, price unit, minimum order size and net price.* From these fields, we first compared article number customer with article number E. Schwartz B.V. Here, none were different (or contained duplicates), see Table 5-3. With respect to the description fields, we found considerable differences. However, since those fields are not mandatory for order processing, we excluded them from the test. Finally, we combined the remaining fields into a net price check where we considered differences due to price unit errors and gross price errors

(the non-EDI customer did not use a net price, but calculated the net price through applying a discount percentage to the gross price). Table 5-3shows 5.8% of the net prices of the EDI customer were different, compared to 6.8% of the net prices of the non-EDI customer.

For the comparison of record test, we measured record completeness similar to the SLIM case. Table 5-3 shows that 18% of the article numbers in the database of the EDI customer were not available in the Schwartz database. This was considerably more than the 4% in the database of the non-EDI customer. An important reason for this difference in data quality is a procedural error in the communication of an assortment change (see Vermeer 1998:1). The cigarette manufacturer had notified E. Schwartz B.V. that it cleaned up its inventory and therefore would reduce a considerable part of the assortment it received. Therefore, E. Schwartz B.V. had removed the links between these articles and the cigarette manufacturer in its system. However, since the cigarette manufacturer had not done so, this resulted in a large difference in the completeness of records measure. This already had caused problems, because several orders could not be processed.

Finally, we measured the actual use of the product database, which showed that 9.8% of the order lines with the EDI customer were processed with product numbers that were not available in the database of E. Schwartz B.V. This could indicate that the article database was not used to process the order (which is possible, if the order was manually processed). However, further research revealed that many of these order lines contained product codes that were removed during the year, but possibly available, when the order was placed. For the non-EDI customer we found that only 0.3% of such order lines existed.

Table 5-3 shows that in total 23.9% of the shared article numbers with the EDI customer contained quality problems. This means that the overall context quality with the EDI customer is 76.1%. In a similar way, the overall context quality with the non-EDI customer is 89.1%. Thus, the context quality with the non-EDI customer appeared to be better than with the non-EDI customer.

Assessment of impact on the ordering process
Next, we assessed the impact on labor time by way of comparing the differences in processing time and prevention time between the EDI and non-EDI customer. The results are shown in Table 5-4.

EDI customer		**EDI**	**Checking**	**Total**
EDI customer	Generate order advise and checking	2	26	28
Schwartz	Manual transport from PC to AS/400	3	0	3
Schwartz	Repairing missing order lines in order	0	7	7
Schwartz	Checking & repairing prices	0	11	11
	Total time per order line (sec)	**5**	**44**	**49**
Non EDI customer		**Non-EDI**	**Checking**	**Total**
Non-EDI customer	Manual entry quotation in order system	30	0	30
Schwartz	Answer phone & checking order	1	22	23
Schwartz	Data-entry Sales	22	0	22
	Total time per order line (sec)	**53**	**22**	**75**
Effects				
	Positive EDI effect	-64%		
	Negative extra prevention effect		+30%	
	Total effect			-34%

Table 5-4: Positive effect of EDI and negative effect of preventive checking[1]

Table 5-4 shows a processing time difference of 26 seconds (=34%) between the EDI enabled and traditional ordering processes. As we can see from Table 5-4, the positive effect of using EDI leads to a decrease of 64% in processing time. However, since extra checking activities are necessary, an increase of 30% is measured in the EDI enabled process.

Finally, we assessed the impact through detecting differences in error correction due to product data errors. The results of this analysis are shown in Table 5-5.

EDI customer			
	Coupling error customer	0.29%	
	Wrong net price	0.21%	
	Return of inventory	0.25%	
	Coupling error Schwartz	0.14%	
	Data-entry error Schwartz	0.07%	
	Total % errors in ordering process		**0.97%**
	Errors in other processes	0.57%	
Non-EDI customer			
	Wrong product ordered	0.05%	
	Wrong net price	0.05%	
	Wrong delivery address	0.16%	
	Data-entry error	0.11%	
	Total % errors in ordering process		**0.38%**
	Errors in other processes	0.33%	

Table 5-5: Analysis of remaining errors

[1] For illustrative reasons, some administraton steps in this table were combined.

Typical errors with the EDI enabled customer were *Coupling errors,* which are the result of article number matching problems between the article database of the customer and E. Schwartz B.V. W*rong net prices,* because of price differences. *Return of inventory errors,* which resulted because the customer stored two equal products under different private product numbers. These errors are especially expensive. *Data entry errors,* when the Sales department of E. Schwartz B.V. accidentally miscorrected some prices. The typical errors for the non-EDI customer speak for themselves.

Table 5-4 shows first of all that the number of errors in both ordering processes is lower than 1% of all order lines, which is low compared to a normal industry average of 2%. Furthermore, the number of data quality errors in the orders without EDI is more than twice as low as the number of data quality errors in the orders with EDI. The main reason is that net prices, coupling errors at the customer and return of inventory errors are not detected through preventive checking. However, these errors do not occur in the non-EDI process, since in that process the buyer and the seller talk with each other on the phone. We learned from interviews that in this conversation, problems with prices and mismatches in product numbers are detected before they can really occur.

5.5.5 Case conclusions

The comparison of the EDI-customer with the non-EDI customer leads to the following results:

1. The measured context quality of the EDI customer was 76%, while the context quality of the non-EDI customer was 89%. Hence, context quality of the EDI process did not exceed the context quality of the non-EDI process.
2. The total effect of EDI on processing time was –34% (decrease in processing time). The positive effect of using EDI was –64%. However, the negative effect of extra prevention activities was +30%.
3. The number of errors in the non-EDI process was 0.38% compared with 0.97% in the EDI process.

Since the case results show that the context quality of the EDI process did not exceed the context quality of the non-EDI processes, and since the case results show an increase in time spent on preventive actions, and an increase in the remaining error rate (which is a measure for the data quality of the EDI orders), we conclude that our proposition is confirmed in this case.

5.6 Conclusions

The objective of this chapter was to test our proposition, which was defined as follows:

> If EDI is introduced, and the quality of the context is not improved, then this context quality will decrease the direct positive effects of EDI on the business process.

To test our proposition, we used the case study method, mainly because we had to be certain that the impact on process performance was indeed the result of insufficient context quality.

This check could only be performed satisfactory through a deep understanding of the ordering process itself.

Since the Schwartz case results show that the context quality of the EDI process did not exceed the context quality of the non-EDI processes and since the case results show an increase in time spent on preventive actions, and an increase in the remaining error rate (which is a measure for the data quality of the EDI orders), we conclude that our proposition is confirmed in this case.

Because the EDI ordering process that was studied in the Schwartz case is similar to many other industrial situations, we believe that these results will also apply in these other situations. However, further testing will be necessary to improve external validity.

6. Data Integration and Distribution

6.1 Introduction

In Chapter 1 we explained that there are two issues we will address in this thesis. The first issue addresses the role of data quality in an EDI enabled business process, and especially the effect of context quality on an EDI enabled process. The second issue addresses how to solve the data alignment problem through making agreements on an interorganizational network level and through developing a data distribution structure. The data quality issue was discussed in Chapters 4 and 5. In the following four chapters, we will address the data alignment issue.

In this chapter, we will discuss the relevant literature for solving data quality problems in an interorganizational context through data alignment. From Chapter 3, we learned that there are two causes for insufficient data alignment between organizations:

- The lack of central agreements on the semantics of product data;
- The lack of a distribution structure to synchronize the product data between the databases of multiple suppliers and customers.

The first cause, namely the availability of central agreements or the lack thereof, is extensively discussed in the database literature, where it is referred to as data integration. Data integration generally means the standardization of data definitions and structures through the use of a common conceptual schema (Heimbinger and McLeod 1985, Litwin et al. 1990). When data is integrated, every user knows how to interpret the data, thus preventing data quality problems.

We will discuss the second cause, namely the lack of a distribution structure to synchronize data, from the literature on interorganizational systems (IOS). This literature traditionally builds

on EDI, as the leading technology to distribute electronic data between computer systems of different organizations. However, new advances in software technology (such as object orientation) and the general advance of Internet technology, result in new data distribution approaches, which will also be discussed.

In Section 6.2 we will present an abstract problem description that explains why data alignment presents a problem in multi-database situations. Next, in Section 6.3 and 6.4 we will introduce several data integration- and data distribution approaches respectively and we will explain how they work using the abstract problem description. Finally, in Section 6.5 we will evaluate all these approaches, after which the conclusions of this literature review will follow.

6.2 Abstract problem description

In a single database situation, real world facts are stored in a single database. This database basically consists of two parts: a database schema, describing the data in the database and the data itself (Date 1990 pp. 38, Elmasri & Navathe 1994, pp.23-28). The schema describes the structure of the data. It describes the real world entities the database recognizes and its attributes. Furthermore, the schema defines which entities are related to each other, what the types of their relationships are and what constraints apply. The data in the database consists of the actual values of the facts that are presented to the database. The single database situation is shown in Figure 6-1.

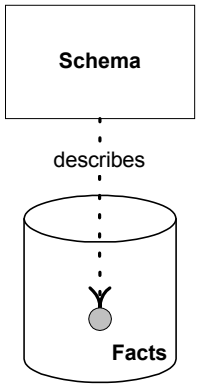


Figure 6-1: Single database situation

The relation between the schema and the data is as follows: When a new fact is presented to the database, the actual values describing the fact are entered into the database using the database schema as a reference model. This means that the schema is used to check whether the entered values conform to the structure and the constraints, as is described in the schema. When a fact is retrieved from the database, the schema is used to formulate the question to retrieve the values that describe the fact. Thus, the schema plays an important role whenever the actual data is manipulated.

In a multiple database situation, the same fact may be distributed over different locations, described by different schemas at each location. Using the schema / data distinction, we may represent the multiple database situation as shown in Figure 6-2.

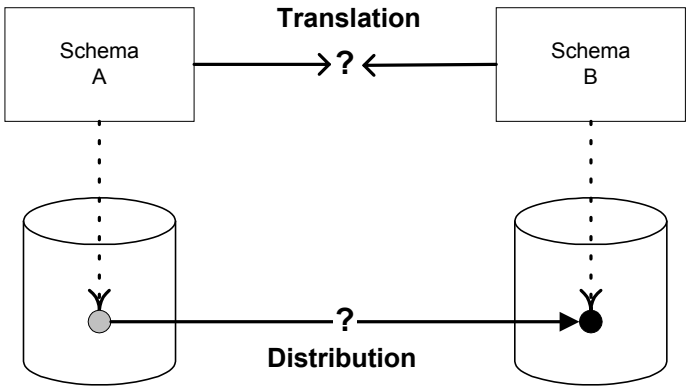


Figure 6-2: The multiple database situation

As we can see from Figure 6-2, two problems arise in a multiple database situation: a translation problem and a distribution problem.

The *translation* problem arises because the same fact may be differently structured at different locations. Therefore, schema translation is necessary to map the structure of the source schema to the structure of the manufacturer's schema. This results in a mapping schema between the source schema and the receiver's schema that is used every time a fact in the source database is updated.

The *distribution* problem arises because each fact update is first translated and then transported over a network to a limited set of users, where it is finally interpreted and stored in the receiver's database. During translation and interpretation, mapping errors may occur, which results in loss of data quality. During transportation, the data may get delayed, damaged, or delivered to the wrong recipient, resulting in inconsistencies among different locations.

Thus, if we want to analyze different data alignment methods, we have to address at least two questions:

1. How does the method solve the translation problem between the schemas of different network participants, which means that data definitions and rules are aligned?
2. How does the method solve the distribution of fact updates between the databases of different network participants, which results in updates that are delivered at the right time, at the right place, with the right quality?

6.3 Data integration approaches

In this section, we will discuss the first set of data alignment methods that are based on data integration. We will start with a discussion of the problem of data integration. It is important to understand how different data integration approaches deal with this problem. Next, we will introduce a taxonomy of multi-database systems to clarify which types of multi-database systems represent different interorganizational situations. Finally, we will discuss three data-integration approaches and we will discuss how they work using the abstract problem description.

6.3.1 The problem of data integration

Goodhue et al. (1992) provide a comprehensive model that explains the problem of data integration (see Figure 6-3).

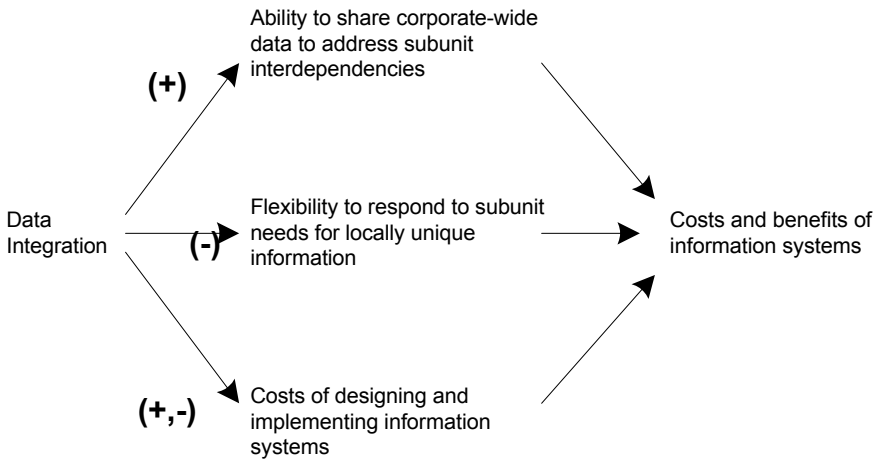


Figure 6-3: The data integration problem (Goodhue et al. 1992)

In their paper, they question the implicit assumption that data integration always results in net benefits to the organization. Using the Information Processing theory of Galbraith (1978) and later Tushman and Nadler (1978), they first identify the sources of uncertainty that drive information processing needs. These sources are: (1) Interdependence among subunits; (2) Complex or non-routine subunit tasks; and (3) Unstable subunit task environment. Next, they show how different degrees of data integration affect these sources of uncertainty. They argue that data integration will have a positive effect in organizational situations where subunits are highly interdependent. Data integration leads to improved coordination and less costs, because no ambiguous messages between subunits are exchanged. However, data integration has a negative effect on situations where subunit tasks are non-routine or the environment is unstable. Data integration requires that all subunits use the same, standardized agreements on data (the

same data model). This decreases the ability of subunits to meet their specific information needs, and therefore it lowers the level of local flexibility of subunits.
Finally, they argue that data integration may affect the cost of designing and implementing information systems either way. Normally, a more expensive initial design leads to fewer costs for subsequent modifications. However, *as the number and heterogeneity of subunit information needs increase, the difficulty of arriving at acceptable design compromises increases and therefore the initial design cost will increase more than linear.* This same effect will appear in later modifications, thereby increasing the long-term costs.

The importance of Goodhue's model is that it explains why data integration is already a problem in an organization, let alone in an interorganizational context. The use of common field definitions and codes[1] (that is: a common data model) means that no ambiguous messages are exchanged between interdependent locations across the network. This will prevent the numerous problems in computer applications, which we described in Chapter 1. On the other hand, the implementation of a common data model across many interdependent network participants is practically impossible. First, because the construction of such a data model from all the participants' models would take many years of work. Second, because such a common data model will virtually destroy the flexibility of all participants' organizations to address the needs for locally unique information. Third, because such complex data models are practically non-maintainable (Pels 1988).

6.3.2 Taxonomy of multi-database systems

In the database literature, the problem of data (base) integration between multiple databases is referred to as *multi-database*, *heterogeneous* or *federated* (Heimbinger & McLeod 1985) database systems. Sheth & Larson (1990) present a taxonomy of these systems which is presented in Figure 6-4 on page 82.

The taxonomy shows that a database system may be either based on a single or multiple database management systems (DBMS). Representatives of Database Systems (DBS) with a single DBMS are the *centralized* DBS and the *distributed* DBS. A centralized DBS consists of a single centralized DBMS managing a single database. A distributed DBS consists of a single DBMS managing multiple databases.

A Multi-database system (MDBS) supports operations on multiple component DBSs. Each component DBS is managed by (perhaps a different) component DBMS. A component DBS may be centralized or distributed and may reside on the same or on multiple computers, connected by a communication subsystem. Sheth & Larson make a distinction between the homogeneity / heterogeneity aspect and the federated / non-federated aspect of MDBSs. An MDBS is called *homogeneous* if the DBMSs of all component DBSs are the same; otherwise it is called *heterogeneous.* A MDBS is called *federated* if the component DBS are autonomous, yet participate in a federation to allow partial and controlled sharing of data. Otherwise the MDBS is called *non-federated.*

[1] This is the definition of data integration according to Goodhue et al..

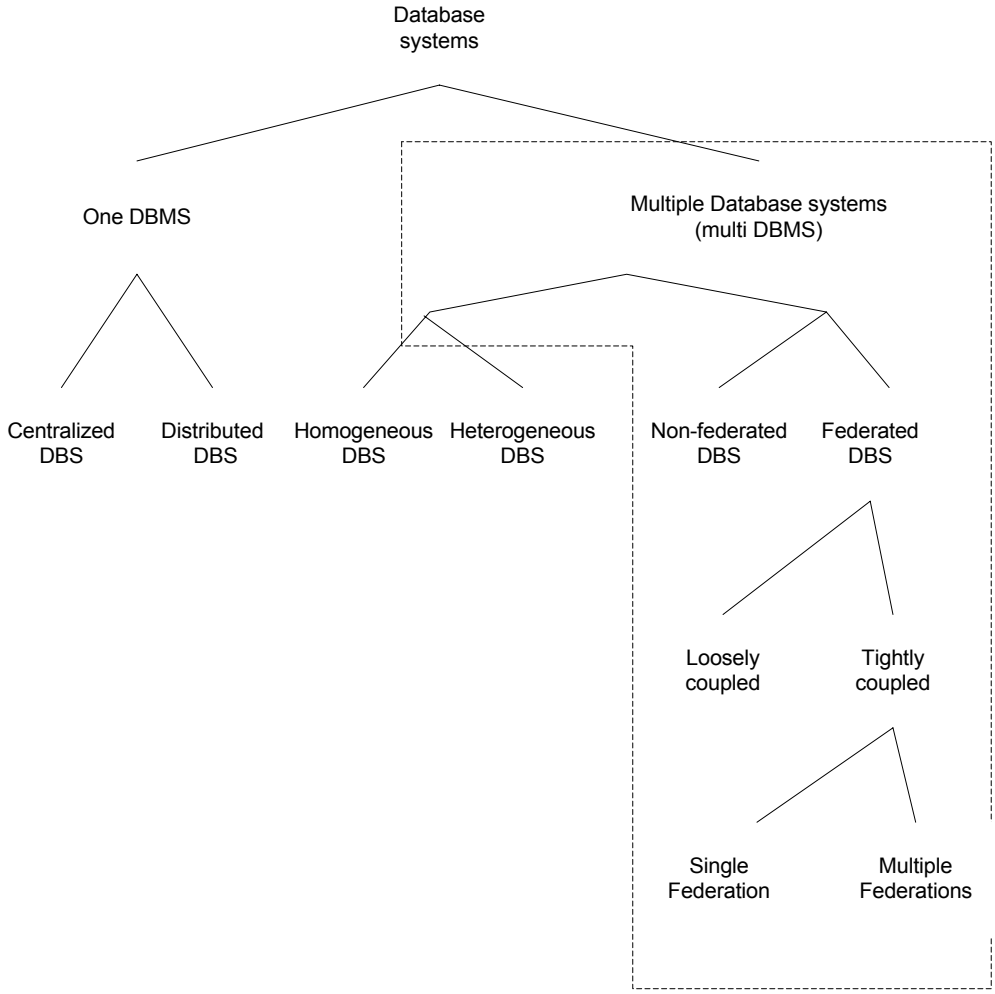


Figure 6-4: Taxonomy of database systems (adapted from Sheth & Larson (1990) [1]).

Federated Database Systems (FDBS) can be categorized as loosely or tightly coupled. An FDBS is *loosely coupled* if it is the users' responsibility to create and maintain the federation and there is no control enforced by the federated system and its administrators. A federation is *tightly coupled* if the federation and its administrators have the responsibility for creating and maintaining the federation and actively control the access to component DBSs. Finally, Sheth & Larson make a further distinction between federations having a single federated schema and multiple federated schemas.

[1] This taxonomy was adapted to include other types of database systems as Sheth & Larson only describe in the text of the article. The orginal taxonomy is shown inside the dashed line.

Goh et al. (1994) also make a distinction between a tight- and loose coupling approach for achieving logical connectivity (that is, meaningful data exchange) between heterogeneous systems. They argue that a tight coupling approach means that conflicts between multiple database systems are reconciled a priori in one or more (federated) schemas. In this framework, users are only allowed to interact with one or more federated schemas, which mediate access to the underlying component databases. In a loosely coupling approach, users interact with constituent databases directly using a multi-database manipulation language, instead of being constrained to querying shared schemas exclusively.

If we compare the definitions of tightly and loosely coupled systems of Goh. et al. with the definitions of the taxonomy of Sheth & Larson, we see that they are synonymous, if we define the use of a multi-database manipulation language as some form of minimum federated schema. For the remainder of this thesis, we will define tightly and loosely coupled systems based on this assumption. Thus, we define both tightly and loosely coupled systems as heterogeneous, federated multi-database systems, using either one federated schema (tightly coupled systems) or a multi-database manipulation language (loosely coupled systems) to translate data between different schemas.

We will now examine how data integration is established within the tightly and loosely coupled systems.

6.3.3 Tight coupling approach

The tight coupling approach basically consists of three steps. In the first step, the local database schemas are translated into component schemas expressed in the Common Data Model (CDM). A CDM describes the local schema in a single database language. In the second step, the component schemas are integrated into one or more federated schemas. The integration of the component schemas in one federated schema is normally established through view or schema integration (Batini et al. 1986). This procedure compares the different component schemas, through identifying naming conflicts and structural conflicts. Naming conflicts arise when the same name is used for different concepts (*homonyms*) or when different names are used for the same concept (*synonyms*). Structural conflicts arise when similar concepts are differently modeled. For instance, an airplane may be modeled as a single entity with a few attributes like passenger capacity, maximum air speed and maximum flight distance, or as a large complex of entities relating different types of wings, wheels and fuselages in admissible airplane configurations. When schemas are compared and differences are detected, the difference must be resolved after which the schemas can be integrated in one federated schema. In the third step, the transformations between local, component and federated schemas are constructed. This means that the mappings between the different schemas are generated together with an appropriate distribution or allocation schema. This schema contains information about the distribution of the data among different locations. Each time data is sent from one location to others, the mappings assure that the data is translated to the right context while the distribution schema assures that the data is sent to the right locations.

In terms of the abstract problem description, the *translation* problem in the tight coupling approach is solved through the definition of a single schema through view integration. This single, federated schema is used in the translation of messages between different locations,

resulting in unambiguous message exchange. Basically, this means that the tight coupling approach solves the translation problem through integrating two contexts into one.

The *distribution* problem is solved through view updating (Batini et al. 1992), also referred to as update synchronization (Ricardo 1990, pp. 511). View updates are used to synchronize multiple copies over different locations, thus providing fast access at multiple sites to the same remote data. The view-updating task is easy in the tight coupling approach, since the distribution schema describes where the copies are stored, and the mappings between the schemas translate the updates to the right format when the data is requested for use.

Since the tight coupling approach enables both unambiguous message exchange and fast access to the same remote data, it is specifically appropriate in situations where high interdependencies between processes exist. However, only when a few locations are involved, the large effort of developing the federated schema through view integration can be justified.

6.3.4 Loose coupling approach

The loose coupling approach uses a multi-database manipulation language to translate the data between different databases. This means that queries, expressed in the language of the first database system, are transformed into equivalent queries, expressed in the language of the second database system, through replacing ‘words’ in terms of the first language with equivalent ‘words' in the second language. The result is that a query expressed in one language is literally translated into a query in the second language.

In terms of the abstract problem description, the loose coupling approach leaves both the *translation* and *distribution* problem to the user. With respect to translation, the user must understand the semantics of the location where he retrieves his data to formulate a valid query. With respect to distribution, the loose coupling approach requires the user to inform about possible updates first, before a transaction can be processed. For instance, when the loose coupling approach is used to order a product, the user must first inform about what products are available at what prices, before the transaction can be processed. In contrast, in the tight coupling approach, the user already knows which products are available at which price, because this information was already provided through the view update mechanism.

Because the loose coupling approach leaves both the translation, and distribution problem to the user, it is specifically appropriate for situations where many different database systems are connected with each other where each location occasionally needs information from another location.

6.3.5 Context mediation approach

Goh. et al. introduce *context mediation* as a new solution for heterogeneous database integration that fits between the tight- and loose coupling approaches. The architecture of their solution in a simple source-receiver system is shown in Figure 6-5.

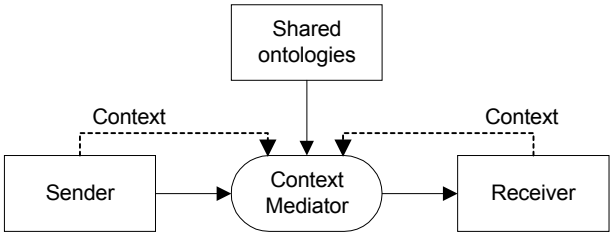


Figure 6-5: Architecture Context Mediation

The context mediation (CM) approach does not solve semantic conflicts a priori through semantic schema integration. On the contrary, only when data is actually transferred from one system to the other, the context mediator detects and resolves semantic conflicts. To illustrate how this works, Goh. et al. describe an example from the financial services community (Goh. et al. 1999). In the example they query two databases that each report the profit of different companies, the first database in the currency of the country with scale factor 1, the second in US dollars with scale factor 1000. In the CM approach, the context characteristics of the two databases are described in their contexts. For the example this means that the context for the first database defines financial data in the currency of the country, with scale factor 1, whereas the context of the second database defines the data as in US dollars, with scale factor 1000. When a query in the context domain of the first database is issued addressing both databases, the context mediator splits the queries for both databases, taking into account the context differences between the databases. This means that the necessary transformations between currencies and scale factors are automatically performed and reported back.

To make these transformations, the context mediator is based on a *shared ontology* of the financial service community. This shared ontology (or domain model as Goh. al. describe it in the 1999 paper) basically describes how this community views the structure of financial information of different companies in different countries. Using this shared structure (or shared schema), Goh. et al. (1994, 1999) are able to map the schemas of the local databases on each other and to translate different currencies and scale factors.

In terms of the abstract problem definition, the *translation* problem in the CM approach is solved through the definition of the shared ontology. If we look at this shared ontology as a unified schema description of the financial service community, we may conclude that the CM approach is based on a specific type of federated schema. Community experts define the central schema (or ontology), while local database administrators define the mappings between local and central schemas. This means that the translation problem is not solved through integrating the different contexts of different local users into one context through a very labor-intensive process, as is the case in the tight coupling approach. Rather, the translation problem is solved through many individual users, who define and maintain mappings between their local schemas and the central schema (the ontology). Compared to the tight coupling approach, the CM approach provides much more flexibility, because changes in local schemas are normally immediately processed by the local user. The downside of the CM approach is that the quality of the translations depends heavily on the individual interpretations of many local users. This may result in many translation errors during transport, because many interpretations are made of the same central schema, resulting in the need to check for data quality. In contrast, if one

central authority defines all mappings between local schemas and the central schema (resulting in one context), no translation errors will arise, because there is only one interpretation of all mappings.

With respect to the *distribution* problem, the CM approach does not solve the fact update problem but leaves this to the user. Thus, the user must contact the right data sources himself to retrieve the required fact updates before a transaction can be processed (as is the case in the loose coupling approach). Hence, we may conclude that the CM approach lies between the tight and loose coupling approaches. It resembles the tight coupling approach in that it provides unambiguous message exchange. On the other hand, it resembles the loose coupling approach in that it is flexible in connecting to other locations to retrieve remote data typically for one time use.

Since the CM approach does not solve the distribution problem, we conclude that it is most appropriate for interorganizational situations where many different databases are connected for incidental information interchange. Therefore, the CM approach is mostly an improvement of the loose coupling approach.

All three approaches that we discussed in this section will be evaluated in Section 6.5.

6.4 Data distribution approaches

In this section, we will discuss the second set of data alignment methods, which are based on electronic exchange and/or distribution of data. We will start with traditional EDI, after which we will introduce the direction of EDI development (Open-EDI) and some concrete results of this development (BIM & BSR). Next we will discuss new methods, based on developments in software technology (OO-EDI based on UML, CORBA/DCOM), and Internet technology (XML, and XML-EDI). Finally, we will discuss some methods from related fields (e.g. PDI & STEP). For each approach we will provide a short description, after which we will describe them in terms of the abstract problem description.

6.4.1 Traditional EDI

For the development of Interorganizational Systems (IOS), EDI has been a key enabling factor (Cash & Konsysnki 1985, Venkatraman 1994). EDI is defined as the automated, electronic exchange of structured and normalized messages between computers of different organizations (Vlist 1991).

Basically, EDI works as follows (Vlist 1991). If two organizations want to exchange business documents electronically (such as orders, invoices or a product information update), they first translate the business document from their in-house format into a standard UN/EDIFACT message using an EDI translation program. UN/EDIFACT is the worldwide leading standard syntax for EDI, consisting of standard data elements and segments, which are combined into a set of United Nations Standard Messages (UNSMs). UN/EDIFACT messages are developed in a worldwide standardization process coordinated by the UN/ECE WP4[1]. After the business

[1] United Nations Economic Committee Europe, Work Party 4

document is translated, the sender posts the message in his mailbox for delivery at a Value Added Network provider (VAN). The VAN owns a (physical) computer network with mailboxes and protocols for transportation and delivery of data (e.g. the X.400 protocols). The VAN sends the message from the sender and delivers it in the mailbox of the receiver. Finally, the receiver retrieves new messages, translates the UN/EDIFACT messages to his own in-house format and processes the information electronically.

In terms of the abstract problem description, the *translation* problem is solved through the use of a mediating independent business syntax, comparable with the multi-database manipulation language in the loose coupling approach. This means that human intervention is still required to understand the semantics of the message. However, because the meaning of a message between two parties can be established at the start of each EDI project, and because orders and invoices are not very complex, automatic exchange of similar messages is possible for future exchanges. This enables each pair of communication partners to define the content of their exchange first and then implement it automatically, resulting in high costs for each implementation, but low costs for the actual exchange of EDI messages.

The *distribution* problem of distributing fact updates is partly solved through two specific messages in EDI, namely the PRICAT and PRODAT messages. These messages can be used to exchange product and price information between different contexts, providing a standard syntax for translation and interpretation (see also Chapter 3). However, the actual transport of these messages is still performed manually, because for each new update, the administration staff in the sales department still determines which updates should be sent to whom and at what time. Furthermore, for each received update, the administration staff in the purchasing department still checks each update manually, before they are entered into the new system. In summary, update distribution with EDI is mainly used as a smart electronic faxing device.

6.4.2 Open-EDI

Because the project cost of setting up each new EDI link is very high (because of making agreements about the semantics of the exchanged data, and setting up procedures for exchange etc.), EDI was not widely adopted in many industries. Therefore, the International Standards Organization (ISO) developed concepts and models for *Open-EDI*. The objective of Open-EDI is to lower the barriers of using EDI through:

> *"Introducing standard business scenarios and support services. Once a business scenario is agreed upon, if implementations conform to the Open-EDI standards, there is no need for prior agreements between the trading partners, other than the decision to engage in the Open-EDI transaction in compliance with the business scenario". (ISO 1996).*

In order to provide standards required for the inter-working of organizations through interconnected IT systems, SC 30 developed the Open-EDI Reference Model. The Open-EDI environment the Open-EDI Reference Model is part of, is shown in Figure 6-6.

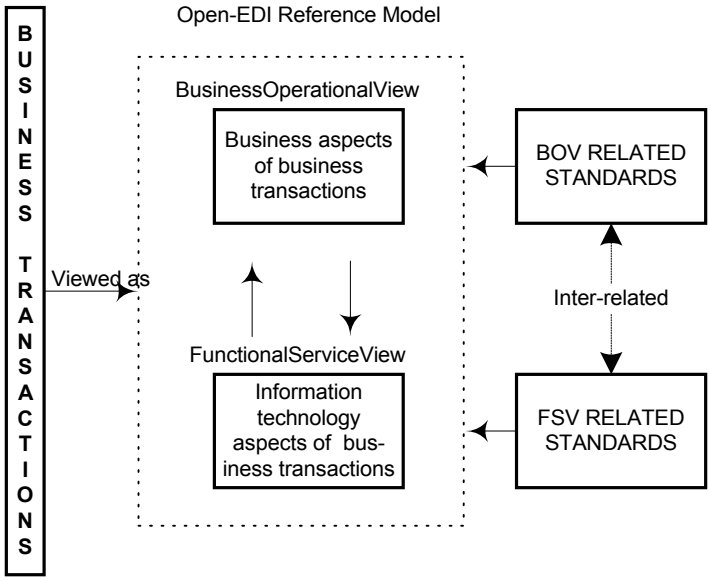


Figure 6-6: Open-EDI environment

The Open-EDI Reference Model uses two views to describe the relevant aspects of business transactions:

- The Business Operational View (BOV);
- The Functional Service View (FSV).

The BOV related standards are employed by business users who understand the operating aspects of a business domain or organization. The intention is that these user groups develop Open-EDI scenarios. These scenarios describe how *Information Parcels* are exchanged between different *roles.* A role models the external behavior of an organization. An Information Parcel is the formal description of the semantics of the information that is exchanged by Open-EDI parties playing roles in an Open-EDI scenario. Information Parcels consist of Semantic Components, which are units of information that are unambiguously defined. Compared to traditional EDI, the roles are the different business applications, the Information Parcels the UNSMs and the Semantic Components the data elements and segments in the UNSMs. Thus, once BOV related standards are in place user groups will use them to produce agreed upon models (e.g. Open-EDI scenarios), which represent their business transactions. These agreed upon models are candidates for registration and standardization.

The FSV related standards address the supporting services meeting the mechanistic needs of Open-EDI. It focuses on aspects such as syntax, control reconciliation, name/address resolution, security mechanism handling, message handling, file transfer, transaction processing, network management and data management. The FSV related standards are used by IT experts. The IT experts are those within an organization who use this technology to design and/or build IT systems, which support the business needs. These experts produce the generic,

FSV compliant IT products and services (Open-EDI systems), which can potentially support the execution of any Open-EDI transaction.

We can explain the contribution of Open-EDI using the abstract problem description. The Open-EDI scenarios (as defined and standardized in the Business Operational View) provide not only a syntax, but also a standardized schema, and a description of the actual transaction exchanges between two communicating parties. This means that Open-EDI scenarios do not only provide a syntax, but also the semantics and pragmatics of the complete communication situation. If we compare the data integration approaches with Open-EDI, then Open-EDI focuses on both transaction communication *and* data alignment, where the data integration approaches only focus on data alignment (For the difference between transaction communication and data alignment, the reader is referred to Section 4.5).

With respect to the *translation* problem, Open-EDI scenarios use a centrally defined (i.e. standardized) schema, where translation is accomplished through local mappings between local and central schema (similar to the CM approach). With respect to the distribution problem, Open-EDI scenarios as of yet do not define specific messages for distributing update exchanges between different contexts.

We make 2 remarks with respect to Open-EDI:

- The Open-EDI reference model only provides the general direction for solving interoperability problems between different computer systems. The specific standards and components that actually solve the translation and distribution problems are not yet developed.
- The concepts presented in the Open-EDI reference model, do mention the semantics of the communication (in terms of “semantic components” and ”information parcels”), but the emphasis is on the pragmatics of the communication. It is not clear how the semantics of the communication is addressed in Open-EDI, other than the recommendation to use a semantic modeling tool (such as Entity Relation modeling, or IDEF).

6.4.3 Business Information Modeling

An important step towards achieving Open-EDI was the work of the Business Information Modeling (BIM) group of UN/EDIFACT, which developed a number of modeling techniques to enhance the EDI message development process. These techniques are incorporated in the BIM framework, which is shown in Figure 6-7.

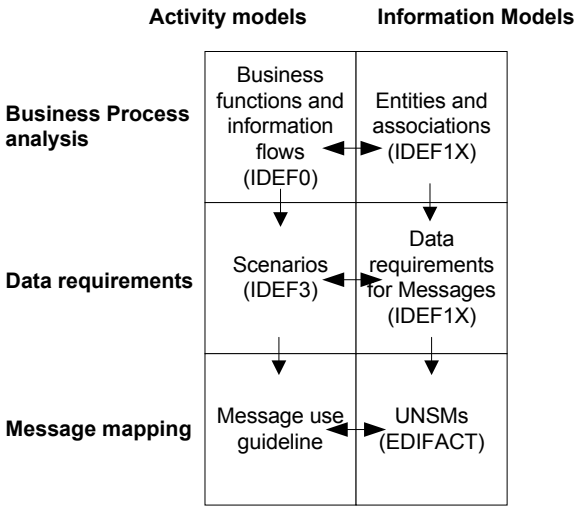


Figure 6-7: BIM framework (adapted from Huang 1998 pp. 36)

As we can see from Figure 6-7, the BIM framework consists of three phases, the business process analysis phase, the data requirements phase and the message-mapping phase, which are standard design steps in Information Systems Design activities. Furthermore, two modeling perspectives are used in each phase. These two perspectives are a dynamic perspective that describes the sequence of activities (the process model) and a static perspective that describes entities and their relations (i.e. the data model). In 1997 a special group of the EDI standardization committee, AC.1, chose the Integration Definition Language (IDEF) as an integrated suite of business modeling techniques to model these two perspectives in the first two phases.

The contribution of the BIM approach is that it provides a concrete set of methods (namely the IDEF modeling methods) to define the Business Operational View of the Open-EDI reference model. Using these methods, the UN/EDIFACT Message Development Groups (MDGs) are able to define both the semantics and the pragmatics of the communication process.

In terms of the abstract problem description, the *translation* problem is solved through the definition of a central schema using IDEF1X by a centralized standardization agency. Actual translation between local contexts and the central schema is accomplished through local mappings between local and central schema (similar to the CM approach). With respect to the *distribution* problem, the Open-EDI scenarios that were modeled using the BIM approach as of yet do not define specific messages for distributing update exchanges between different contexts.

6.4.4 Basic Semantic Repository

An important development in conjunction with Open-EDI in the beginning of the nineties is the Basic Semantic Repository (BSR). Although UN/EDIFACT is the leading EDI syntax, several other syntaxes exist (such as X.12, the American EDI syntax), that are currently migrating

towards UN/EDIFACT. Furthermore, the UN/EDIFACT syntax is very complex, consisting of hundreds of data elements and thousands of codes that all have specific meanings and are difficult to distinguish from each other. Therefore, the Department of Computer Science at the University of Melbourne under supervision of Ken Steel designed and built a BSR. The BSR is defined as:

> *"A tool, that enables, where possible, any organization to use the SAME data element concept, so as to be able to exchange the SAME data between systems and KNOW it is the SAME ". (BSR PROJ 1995).*

Basically, the BSR consists of Basic Semantic Units (e.g. Buyer.name), which are unique descriptions of a certain concept that is concisely defined. Furthermore, the BSR provides bridges between these BSUs and similar concepts in other EDI directories (in different languages), using bridges. The result is some kind of dictionary, where each BSU is unambiguously defined and is related to all possible concepts in other EDI directories in the world.

The experimental Australian Enhanced BSR and the user guide (Steel 1996:1) were placed on line on the Internet in January 1996. Since the BSR was closely linked to Open-EDI, the ISO and UN/ECE adopted the BSR approach from the Australian project. However, since:

> *"The original concept of the BSR, even with the Australian Enhancements is now seen to be addressing only a minor part of the requirements of such a facility… it has now been superseded. The experimental facility was decommissioned in February, 1998" (Steel 1996:2).*

The UN/ECE also pulled out of the BSR project.

In terms of the abstract problem description, the BSR was trying to solve the *translation* problem through developing a unified schema for business communication, using view integration. This approach was abandoned, most probably because it is virtually impossible to integrate the contexts of all business users worldwide. The distribution problem was not addressed in the BSR approach.

6.4.5 Object Oriented EDI based on UML and CORBA/DCOM

OO-EDI based on UML
In response to the slow adoption of EDI (see AC.1 1996) the UN/EDIFACT Steering Committee (ECG) developed a new strategy consisting of three tracks (UN/ECE 1996). The first track focuses on the continuation of mainstream EDI message development. The second track encourages development of simpler UN/EDIFACT messages relevant to the needs of SMEs. Finally, in the third track, the use of the Object Oriented approach to the design of future messages is researched. The Techniques and Methodologies Work Group (TMWG) of CEFACT[1] coordinates this research. CEFACT is the re-engineered UN organization for the administration of UN/EDIFACT standards development (TMWG 1998:1).

The use of Object Oriented Technology (OOT) in EDI was chosen because object orientation is now widely adopted within the software development community. Thus, if all new information systems are based on OOT, then EDI should also use object oriented principles to support

[1] CEntre For facilitation of procedures and practices in Administration, Commerce and Transport.

interorganizational communication. For a detailed description of OOT, the advantages of OOT and the construction and use of object classes we refer to Taylor (1998). Many principles for OOT for EDI are summarized in The Reference Guide for The Next Generation of EDI (TMWG 1998:1).

To support the OO-EDI development process, TMWG recently replaced the IDEF modeling suite with the Unified Modeling Language (TMWG 1998:2). UML was constructed from three influential OO-IS design methods, namely the Booch method (Booch 1991), the Object Modeling Technique, OMT (Lockheed et al. 1996), and the Object Oriented Software Engineering method, OOSE (Jacobson 1995), and was adopted by the Object Management Group (OMG) in 1997. UML is basically an integrated suite of business diagrams that model systems in an object oriented way. Important diagrams in UML are the use-case diagram, the class diagram, the collaboration diagram, the interaction diagram, the state diagram, the activity diagram, the package diagram, the deployment diagram and the component diagram. For more information on UML, we refer to Fowler & Scott (1999) or Ericsson and Penker (1998). Furthermore, the Reference Guide of TMWG (TMWG 1998:1) provides a detailed description of the use of UML for OO-EDI modeling applied to an example of a buyer executing a catalogue order with a seller.

In terms of the abstract problem description, using UML for defining Open-EDI scenarios is similar to using the IDEF methods from the BIM approach. This means that the translation problem in OO-EDI is solved through the definition of a central schema in UML. The distribution problem is not addressed within OO-EDI.

CORBA / DCOM
The Common Object Request Broker Architecture (CORBA) is a distributed system technology (Verwijmeren 1998, pp. 186). CORBA was developed by the Object Management Group (OMG), who is also responsible for the UML standard. CORBA specifies middleware that creates a transparent information-processing platform to which interacting system components can be linked and can then communicate with each other. CORBA is specifically designed for interoperability between heterogeneous, object oriented information systems.

The CORBA architecture consists of an Object Request Broker (ORBs), The Interface Definition Language (IDL), IDL stubs and skeletons, a General Inter ORB protocol (GIOP) and CORBA services and facilities. IDL is a standard syntax that can be used to specify object interfaces. The function of IDL is to provide translation between method calls expressed in specific object oriented languages. The IDL stubs and skeletons are the linking pins between the client or server objects, and an ORB. Basically, stubs and skeletons are compiled runtime components of the IDL definitions of the client and server objects, which perform the actual translation of method calls between client objects and server objects. The ORB is the basic platform that manages the objects. Therefore, it performs functions such as creating instances of objects, and managing the interchange of information between objects. The GIOP is an inter ORB communication protocol for communication between many ORBs. Finally, the CORBA Services and Facilities are two special object categories included in CORBA, which further support the development and application of distributed systems. CORBA Services provide services at the infrastructure level, such as a service to manage the lifecycle of objects. CORBA Facilities provide services at the application level, for instance, the definition of different

content libraries (with definitions) for specific application domains, such as healthcare, banking, transport etc. DCOM is similar to CORBA, except that Microsoft developed it.

From a distance, CORBA and DCOM are very similar to traditional EDI, except that it provides interoperability between object oriented information systems, not traditional information systems. In this perspective, CORBA and DCOM could be described as 'EDI for objects'.

In terms of the abstract problem description, CORBA and DCOM solve the translation problem through the provision of a standard syntax, namely IDL. The update distribution problem is not addressed in CORBA or DCOM in the sense that no special method calls for exchanging fact updates are defined (although this could be easily implemented).

Combining OO-EDI & CORBA/DCOM

We will combine OO-EDI based on UML and CORBA/DCOM, because OO-EDI & UML both are distribution approaches that are complementary to each other. OO-EDI with UML provides a solution for the semantic and pragmatic aspects of communication between objects, while CORBA/DCOM provides the syntax and translation mechanisms for communication between objects.

With respect to the combination of OO-EDI with CORBA/DCOM we make two remarks:

- Using OO-EDI up to this moment means that UML is used for defining Open-EDI scenarios. From these scenarios UNSMs are generated via the normal standardization process. However, TMWG describes a possible near future, where OO-EDI means the development of concrete working software components using the OO-scenario definitions. This could be achieved through a new standardization process, where development groups work closely together with software developers. This would enhance the usability of EDI significantly, because **in normal EDI**, the user has to make the mappings between the local schema and the EDI syntax completely by himself (the central schema only exists as a paper document about the central agreements in the central schema). Furthermore, whenever changes in the central schema arise, the user must interpret all mappings again, based on the new paper document describing these changes. Because both the distribution of paper documents and the interpretation of these documents take considerable time, this may result in many errors during translation. **In OO-EDI**, the central schema is already embedded in working software, which means that the user only has to map between the entities from his local schema and the central schema in the working software. Normally, this process is considerably simplified through intelligent user interfaces that model the mappings between local and central entities. Furthermore, OO-EDI has specific advantages when changes to the central schema occur, because these changes are reflected in the centrally defined class definitions, which means that they are immediately available in the working communication software, provided that this software is online. However, since the *consequences* of these central changes still result in manual mapping activities for the local user, we will define the implementation of the translation solution in OO-EDI as semi-automatic.

- Although this semi-automatic implementation of central agreements might work well in theory, many problems remain in practice. In fact, the Norwegian Computing Center tested this concept in an Open-EDI prototype, based on UML, CORBA and Java (TMWG 1998:3). Basically, this was a test to evaluate the vision of TMWG of a standardization process that results ultimately in an implemented and working version of object oriented EDI-translation software, which in its runtime version incorporates the class definitions that were developed in the Open-EDI scenarios. However, the report on this project discloses several difficult problems. Specifically, the report shows that scenarios, which were defined in UML, still required a lot of manual programming before they resulted in working communication components. Therefore, the report concludes that:

  "Use of formal description techniques for top-down development of software has been touted for a long time, and sometimes the obsolescence of programming has been prophesied. In practice, this has not happened, and 'programming by specification' mostly remains a sweet dream".

  With respect to the translation problem, this means that in practice the gap between agreements specification and the actual implementation of agreements in working translation software, is still large (unlike XML/EDI as is described in Section 6.4.7).

6.4.6 XML & XML/EDI

Other important developments for the near future are XML and XML/EDI. Extended Markup Language, or XML, is a markup language that enables user defined, and therefore customized, structuring of electronic documents. We will explain this definition using an example of XML, which is described in Table 6-1.

DTD (at location http://www.xmlsource.nl/dtds/address.dtd)

```
<?xml version =”1.0”?>
<!DOCTYPE address [
        <!ELEMENT address (name, street, city, zip, country)>
        <!ELEMENT name (#PCDATA)>
        <!ELEMENT street (#PCDATA)>
        <!ELEMENT city (#PCDATA)>
        <!ELEMENT ZIP (#PCDATA)>
        <!ELEMENT country (#PCDATA)>
]>
```

Document

```
<?xml version =”1.0”?>
<!DOCTYPE ADDRESS SYSTEM “http://www.xmlsource.nl/dtds/address.dtd”>
<address>
<name> TUE </name>
<street> Den Dolech 2 </straat>
<city> Eindhoven </city>
<zip> 5600 MB </zip>
<country> Netherlands </country>
</address>
```

Table 6-1: Example of XML (adapted from Goossenaerts 1999)

The example shows the definition of the address of the University of Eindhoven in XML (We adapted this example from Goossenaerts 1999). In XML, both the meaning of the address and the actual address are defined together. The meaning of an address (e.g. name, street, city, zip code, country) is defined in the Document Type Definition (the upper part of the example). This is done by specifying the elements of an address, and the type of each element. The meaning of an actual address (e.g. TUE, Den Dolech 2, Eindhoven, 5600 MB, Netherlands) is described in the lower part of the example, using tag-pairs of elements that were defined in the DTD (e.g. <name> TUE </name>). Because an XML document normally contains a DTD or a reference to a DTD (as shown in the example), in XML the meaning of a document, and the document itself are defined together.

The basic idea behind XML/EDI, is to define electronic business documents for electronic data interchange using XML, because XML can perform much the same function as UN/EDIFACT (TMWG/XML task group 1999, pp. 5). Several suggestions have been made to auto-generate XML DTDs directly from EDI directories (EEMA EDI Working Group 1998). However, TMWG recommends not to do this, but to focus on ensuring that a standardization process is also enforced in new XML-EDI directories to prevent the proliferation of multiple, largely similar DTDs. Recently, UN/CEFACT and OASIS (the standardization group of the W3C, which is the development group of internet technologies) joined forces in ebXML (http://www.ebXML.org) to establish such a standardized EDI/XML directory for electronic business XML.

The technology of XML has three important advantages:

- Firstly, the gap between specifying agreements about the meaning of a message and the implementation of this specification in a translation package, is strongly reduced. XML enables the library with standard agreements about the meaning of an XML message (expressed in the DTD), to be used online in order to process an XML message. In contrast, traditional EDI requires that the library with agreements about standard EDI messages (which define the meaning of an EDI message) are manually entered into the EDI translation system, either by the EDI translation software vendor, or by the company EDI software team. Thus, XML provides not only a way to specify agreements to solve the translation problem, but it also enables direct implementation of this specification in the translation software because the XML parser, that translates the XML message, uses the DTD to perform this translation. Although this characteristic of XML greatly reduces the manual activity for incorporating changes in agreements into the translation software, it does not eliminate manual activities. Especially when changes occur in the relations between entities in the central conceptual schema, this requires that new mappings between the local schema and the central schema, expressed in the DTD, are defined, which remains a manual activity.
- Secondly, the gap between the place where agreements are established, and where they are used, is further reduced. XML enables each user community to use the standard message syntax (using the DTD of a standard EDI message), and **extend** it to include user group specific agreements (which in traditional EDI are defined in the Message Implementation Guidelines). This results in a much higher degree of semantics that is included in the DTD; that is, if we follow Stamper's assumption that the meaning of an (sign) object depends on the interpreter (see Section 4.3.2). Following this line of

reasoning, the proliferation of multiple, largely similar DTDs should not be prevented but encouraged. In our opinion, the standardization committees should focus on the syntax of business communication (developing message DTD/s), while user groups should extend these DTDs to make their agreements more meaningful.

- Thirdly, XML messages can be interpreted both by computers (as discussed above) and human beings simultaneously, thus without editing or altering the content of the XML message. Specifically, human interpretation is possible through the use of XSL (which is a style sheet manipulation language) that makes the message immediately readable in all browsers used today.

In terms of the abstract problem description, XML-EDI solves the *translation* problem through the definition of multiple, user defined schemas, which are formed and *maintained* by specific user groups. Mappings between local and central schemas are still manually defined. Although changes to the central, user defined schema are automatically available in the translation software, because the local parser in the translation software always refers to central agreements via a URL link. The result of this solution is semantically rich schemas (in the definition of Stamper), because each user group controls the contents to a large extent. In practice, this means that the syntax for communication in a specific sector (for instance, the food sector) is defined centrally, while the semantic agreements for each sector are implemented in additional DTDs for each country.

Similar to the BIM approach, OO-EDI and CORBA/DCOM, in XML/EDI the *distribution* problem of updates is not addressed.

6.4.7 PDI & STEP

In the technical industry community a standard was developed for the exchange of technical product data, namely Product Data Interchange (PDI). PDI was originally a collection of neutral file formats enabling the interchange of complex drawings between different computer drawing applications. However, as integrated product engineering methods developed, the requirements of the product engineering community also increased. Especially the introduction of Generic Product Models generated a new requirement for PDI. Generic Product Models are semantic descriptions of complex products, of which views can be constructed for specific disciplines (e.g. a manufacturing, purchasing and accounting view of the same product, which are constructed from the same Generic Product Model). In 1984, the development of a new standard for Product Data Interchange was started, named STandard for Exchange of Product model data (STEP). STEP is an ISO standard, which is described in a collection of parts of ISO 10303 (Owen 1993).

STEP consists of seven classes, of which the most important ones are the Integrated Resources and the Application Protocols. The *Integrated Resources* provide an assimilated set of information models, which are the resources from which application protocols are built. The *Application Protocols* provide the comprehensive requirements for implementations by defining the application domain (the context). This is achieved by constructing the application protocol from an application-specific interpretation of the context-independent entities present in the integrated resources. The information in the integrated resources and the application protocols is described using standard *Description Methods*. The most important method in

STEP is EXPRESS and EXPRESS-G, which is a semantically rich data modeling language, also suited for constraint specification.

In terms of the abstract problem description, STEP focuses only on the *translation* problem, largely in the same way as EDI with the BIM approach. STEP defines a central schema using EXPRESS and EXPRESS-G. Local users are responsible for mapping between local schemas and the central schema. The distribution problem is not addressed within STEP.

Today, STEP is a widely used standard for the exchange of technical product data in the industrial sector. Despite its wide acceptance, STEP and more specifically EXPRESS-G have not been adapted by the trade community, which still uses UN/EDIFACT. Recently, the TMWG rejected EXPRESS-G as a candidate for data modeling, but preferred the OO-data modeling techniques in UML.

6.5 Evaluation of data integration and distribution approaches

Finally, in this section we will provide an evaluation of the data integration and distribution approaches based on the discussion in the previous two sections, using the abstract problem description of Section 6.2. First, we will work out the translation and distribution problem in the abstract problem description in a classification model, to compare the different approaches in Section 6.5.1. Next we will use this model to evaluate the data integration approaches in Section 6.5.2, and the data distribution approaches in Section 6.5.3.

6.5.1 Classification model

With respect to the translation problem, we will focus on four issues:

1. What is the result of solving the translation problem? In other words, what type of shared schema is developed? We have four choices:
   - Central syntax. This means that shared agreements on the syntax level of communication are established.
   - Central schema. This means that a central shared schema is constructed from multiple local schemas.
   - Central ontology. This means that a central shared schema is developed, not from multiple local schemas, but independently of local schemas through a central agency.
   - Multiple, user-defined ontologies. This means that *multiple* shared schemas are developed, one for each user group that wants to make agreements for use in their specific group. Since such schemas are closer to the users, and since according to Stamper semantics or meaning is defined in relation to the shared context *of a specific social group,* we define user defined ontologies as semantically very rich.

2. How is the shared schema developed? We have three choices:
   - View integration. This means that multiple local schemas are integrated into one schema, through resolving schema conflicts.
   - User mapping between local and central schema. This means that local users define the mappings between local and shared schema, based on their own interpretation of the shared schema.

- 'Flat' mapping. This means that messages in the local syntax are literally (or word for word) mapped to the central syntax. For instance, the literal mapping of 'sla me' from Dutch to English would translate to 'lettuce me'.

3. What is the degree of semantic richness? We make a distinction between three levels of semantic richness:
   - Low. Translation using mapping between different syntaxes, is considered to have low semantic richness, because the relations and constraints between constructs in a message are not defined in a syntax.
   - Medium. Translation using mappings between a local and a centrally defined schema is considered to have medium semantic richness, because it enables mapping between constructs, but it requires only one 'true' central schema, which is developed by one central agency. Since semantics (according to Stamper) is defined in relation to the shared context *of a specific social group,* one centrally defined schema for a group that contains hundreds of subgroups is less semantically rich than multiple shared schemas for each subgroup.
   - High. Translation using mappings between a local and multiple user defined and maintained schemas is considered to be semantically rich.

4. How are central agreements implemented in working translation software? We have two choices:
   - Manually. With manually we mean that the user interprets the central agreements from a paper document, and manually maps between local and central schema.
   - Semi-automatic. With semi-automatic we mean that the central agreements are already available in the working communication software, and that the user only has to map from the data entities in the local schema to the data entities in the central schema. The advantage of semi-automatic is that changes in the central schema are immediately available in the online working software, thus enabling better data quality checking.

With respect to the distribution problem, we will focus on two issues:

1. How does the approach get the fact updates at the right time at the right place? We have four choices:

   - View updating. View updating means that data is physically located at multiple sites for reasons of risk avoidance and query performance. Whenever an update occurs, the data at multiple sites is immediately updated, using the translation definitions that are defined in each site's view. Strictly speaking, view updating is no data distribution mechanism, because data distribution is defined as the distribution of fact updates between *multiple user contexts*. Since view updating is only used when schemas are integrated, this means that only one and not multiple contexts exist.
   - Retrieve with transaction. Retrieve with transaction means that fact updates are retrieved just before a transaction is executed. This means that a purchasing application first retrieves the latest products and prices of the seller before it orders a product. This in contrast with distributing updates to all users, the moment they arise.
   - Manually with PRICAT/PRODAT. This means that fact updates are distributed the moment a fact update arises, using the PRICAT and/or PRODAT message in EDI. However, we consider this as manual distribution, because for each new update, the

administration staff in the sales department still determines which updates should be sent to whom and at what time. Furthermore, for each received update the administration staff in the purchasing department still checks each update manually before they are entered into the new system.
- Not addressed (X). This means that getting fact updates at the right time at the right place is not addressed in the specific approach.

2. How does the approach ensure that the fact update has the right quality? We have three choices:
   - No issue. This is the case when the view updating mechanism is used. Since the update does not cross the context boundary, interpretation errors cannot occur, resulting in a 100% data quality. This is the effect of the definition of data quality. Since we define the central schema as the reference for the local schemas to measure data quality, and since the central schema is equal to the local schemas when schema integration has been completed, data quality is 100%[1].
   - Manually. This means that fact updates are checked manually to eliminate interpretation errors as much as possible. Manual inspection is only necessary when mappings between two different contexts are used.
   - Not addressed (X). This means that getting fact updates at their destination with the right quality is not addressed in the specific approach.

We will use this classification to evaluate the data integration, and data distribution approaches in the next to sections.

6.5.2 Evaluation data integration approaches

Table 6-2 gives an overview of the different data integration approaches in terms of the classification model.

[1] One could compare this with the obvious error of calling a cat a dog. If someone maps the word dog on a cat, this results in a mapping error in our terminology. However, when the whole world refers to a cat with the word dog, than (by definition) a cat is a dog, although this is normally an obvious error.

Problem		Tight coupling	CM approach	Loose coupling
Translation	Result?	Central schema with ER	Central ontology with ER	Centrally defined multi-database language (=central syntax)
	How?	View integration	User mapping between local & central schema	'Flat' mapping between local and central syntax
	degree of richness?	Medium	Medium	Low
	Implementation?	Manually	Manually	Manually
Update distribution	Right time, right place?	View updating	Retrieve with transaction	Retrieve with transaction
	Right quality?	No issue	X	Manual checking

Table 6-2: Data integration approaches

Translation problem
With respect to the translation problem, the loose coupling approach is less sophisticated than the tight coupling approach and the CM approach, since the tight coupling approach uses only a standard syntax to translate between different contexts. The user is still needed to interpret the context of the communication partner, before a transaction can be executed, or a query can be formulated.

The tight coupling approach solves the translation problem differently from the CM approach. The tight coupling approach requires that a labor-intensive integration process is performed, which results in a new, completely integrated context from two old contexts. The advantage is that translation errors resulting in loss of data quality cannot arise. In contrast, the CM approach requires all local users to implement the mappings themselves. This results in much less labor than view integration. However, this comes at the cost of more data quality errors. Since each local user may interpret the central agreements differently, or make mapping errors, translation between contexts will result in loss of data quality. It will depend on the complexity of the mapping, the number of contexts to be integrated, the number of model changes and the number of fact updates, when the tight coupling approach or the CM approach will be the best solution.

Distribution problem
The problem of distribution of fact updates is addressed in all three approaches, although the tight coupling approach deals with it most adequately. The tight coupling approach uses view updating to replicate data between physically dispersed locations. Data quality is no issue, because interpretation does (by definition) not occur within one context. Thus, if the cost of view integration justifies the effort, the distribution problem of fact updates is adequately solved in the tight coupling approach.

The other two approaches use a retrieve when required approach for getting the fact update at the right time at the right place. With respect to the issue of getting the fact update with the

right data quality, the loose coupling approach leaves this problem to the user (through manual checking). The CM approach does not address this issue.

6.5.3 Evaluation data distribution approaches

Table 6-3 on gives an overview of the different data distribution approaches in terms of the classification model.

Translation problem
With respect to the translation problem, the table shows that all approaches use a form of mapping to a central ontology, largely similar to the CM approach of the data integration approaches. This generally means that a large degree of flexibility can be achieved, compared to view integration (especially in large interorganizational networks). EDI scores low on solving the translation problem, because it only uses a central syntax. The other approaches also provide some form of semantic modeling of the data. Only XML/EDI scores high. The main reasons for that is that XML/EDI solves the translation problem through the development of multiple, user defined shared schemas, which can be actively maintained by each user group. Using our definition of semantics, this means that these types of schemas are semantically very rich. Furthermore, XML/EDI provides for semi-automatic implementation of agreements in translation software.

Problem		Traditional EDI	Open-EDI				PDI/STEP & VAN
			BIM-EDI	BSR approach	OO-EDI &CORBA/ DCOM	XML-EDI & internet EDI	
Translation	Result?	Central syntax in EDIFACT	Central ontology in IDEF	Central schema	Central ontology in UML	*User* defined and maintained ontology in XML	Central ontology in Express-G
	How?	'Flat' mapping between local and central syntax	User mapping between local and central schema	View integration	User mapping between local and central schema	User mapping between local and central schema	User mapping between local and central schema
	Degree of richness?	Low	Medium	Medium	Medium	High	Medium
	Implementation?	Manually	Manually	Manually	Semi-automatic in theory, but manual in practice	Semi-automatic	Manually
Update distribution	Right time, right place?	Manually with pricat/prodat	X	X	X	X	X
	Right quality?	Manual checking	X	X	X	X	X

Table 6-3: Data distribution approaches

Distribution problem
With respect to the distribution problem, only traditional EDI provides a solution for distributing fact updates. With respect to getting the data at the right time at the right place, the sales department of the sender manually determines which update is sent to whom at what time, using the PRICAT, and PRODAT messages. With respect to data quality, the purchasing department at the receiver still checks all updates manually before they are entered in the system. All other approaches do not address the distribution problem. Hence, the distribution problem is not adequately addressed in the data distribution approaches.

6.6 Conclusions

In this chapter we presented an extensive overview of data integration and distribution approaches as possible solutions for the data alignment problem. Furthermore, we constructed a classification model to evaluate which method or combination of methods is sufficient to solve the translation and distribution parts of the data alignment problem.

Based on the evaluation in Section 6.5, we draw the following three conclusions:

1. In large interorganizational situations, the best way to solve the *translation* problem is to define one central ontology per specific user group, where local users define their own mappings between local and central schema. In case of a few schemas, and less complexity of each individual schema, the view integration method of the tight coupling approach is preferred, because after integration only one context exists, which means that the data alignment issue is eliminated completely. In case of many local schemas, the central ontology approach is preferred, because it provides for much more flexibility than the view integration method. However, this comes at the cost of having to check for data quality, because many interpretations of the central schema result in mapping errors. The central syntax solution is not adequate, because it still results in many translation errors.

2. From all the approaches we studied, the XML-EDI approach provides the best solution for the translation problem, because it is semantically rich, and because it provides for semi-automatic implementation of central agreements in working EDI translation packages.

3. In large interorganizational situations, no effective distribution mechanism for fact updates is available. The data distribution problem for fact updates is only solved adequately when the data integration solution is selected for data alignment, which is only the case in small interorganizational networks (consisting of 2-3 participants). However, from the approaches we discussed, only the CM approach provides a solution for getting the updates on time at the right place through the retrieve-with-transaction mechanism. With respect to data quality, only traditional EDI checks fact updates, but only manually. All the other approaches do not provide a mechanism for data distribution at all.

We will use these conclusions to design a new data alignment method in the next chapter.

7. Data Alignment through Logistics

7.1 Introduction

In the last chapter, we concluded that current data integration and distribution methods do not meet the requirements for sophisticated data alignment. Although several approaches solve different parts of the translation problem, none of the approaches are sufficient for solving the data distribution problem as a whole. Therefore, in this chapter we will introduce Data Alignment through Logistics, a data alignment method for interorganizational business networks. Realizing that the real challenge in data alignment is to distribute data from a source database to many receivers while maintaining a certain quality level, the idea behind this method is to apply principles from the field of logistics to the data alignment problem.

In Section 7.2, we will start with an overview of the field of logistics, after which we will translate the concepts of logistics to the problem of fact update distribution in the information world. Based on this translation, we will present two important conclusions about designing fact update distribution networks. In Section 7.3, we will use the results of this analysis together with the results from Chapter 6 (how to solve the translation problem) and Chapter 2 (general requirements for design methods) to formulate the precise requirements for a data alignment method we want to develop. Next, we will introduce Data Alignment through Logistics (DAL) in section 7.4, which is our approach solving the problem of data alignment. We will also discuss how DAL fulfills our requirements. Since some requirements cannot be fulfilled beforehand, we will end this chapter with a description of the first version of DAL in Section 7.5. This version of DAL is the basis for developing and testing this method further in two practical cases discussed in Chapter 8. Finally, in Section 7.6 we will compare DAL with the other approaches for data alignment from Chapter 6.

7.2 Logistics in Information

7.2.1 The field of Logistics

The field of logistics is concerned with the production and distribution of goods from raw materials to finished goods available to the consumer. The focus of logistics is on the supply chain. A supply chain consists of two parts: the physical structure of the supply chain and the control structure of the chain (see Figure 7-1).

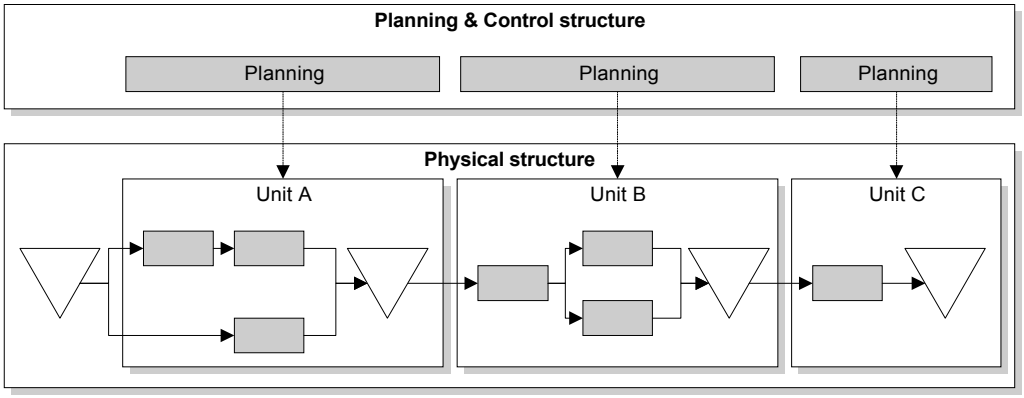


Figure 7-1: The Logistical structure of the supply chain

Physical structure
The *physical structure* of the supply chain consists of processes that are linked to each other, sometimes separated by stocking points. In Figure 7-1 the gray boxes represent the processes, and the triangles the stocking points.

A process uses materials (input) and produces goods (output) during a certain time period, consuming one or more resources. A stocking point is a controlled point in the supply chain where finished goods from a process are stored, until further processing. With controlled we mean that an explicit production order from the planning function is required before processing proceeds (unlike uncontrolled stocking points, where the next production process may retrieve goods from stock whenever these are required). Finally, Figure 7-1 shows that different organizational units exist in the chain, where each unit controls a part of the supply chain. Organizational units may be within company boundaries, or across company boundaries. Examples of organizational units within companies' boundaries are: pre-production department, production department, and distribution center. Examples across company boundaries are: supplier, wholesaler and retailer.

Control structure
The *control structure* of the supply chain is concerned with its coordination. The control structure describes per organizational unit how information between the planning functions of different organizational units is exchanged. In each organizational unit, the control structure

describes which plans are developed on each level of the planning hierarchy, what information each plan uses to construct the plan, and how different plans are linked to each other. For instance, in a weekly production schedule typically the following information can be used: demand information about received orders for the next week, current inventory levels, current work in process figures, a forecast of available machine capacity for next week, and control parameters such as minimum batch sizes, maximum department occupancy rate, maximum inventory levels etc. Based on this schedule, the orders for a specific part of the supply chain are then released during the next week. The control parameters play an important role in the design of the control structure.

Objective of logistics
Hutchinson (1987) defines the objective of logistics as "having the right quantity of the right item in the right place at the right time". Van Goor et al. (1989) describe the objective of logistics as "The effective and efficient propulsion of good flows between manufacturer and customers, such that goods arrive at the customers on the right place at the right time".

We will define the objective of logistics as follows: The objective of logistics is to design and operate a supply chain such that the costs of production and distribution are balanced with throughput time and the level of service to the customer. Throughput time is defined as the time that is needed to manufacture and deliver products to the customer. The level of service to the customer is normally defined in terms of the percentage of products delivered at the right time at the right place with the right quality.

Logistical design decisions
When designing logistical systems, a logistical designer takes a number of design decisions with respect to the physical and control structure of the logistical system. We will give an overview of the important issues for designing both structures.

Design decisions for the physical structure
In designing the physical structure, two important design choices are made:

1. Location of activities and stocking points.
2. Location of the Customer Order Decoupling Point. (CODP).

In the first decision the structure of the network is determined in terms of sequence of activities, and number of parallel capacities (number of production plants, or number of warehouses on supply chain level), and location of stocking points such that the costs of production and distribution are minimized under the constraint that the maximum lead-time, as required by consumers, is maintained. When redesigning supply chains, this first decision is mainly concerned with the redistribution of activities over the organizational units in the supply chain. An example is pushing the final customization process (like finishing a computer) from the central production plant to national sales organizations.

Once the basic structure of the network is in place, the next decision is the location of the Customer Order Decoupling Point (CODP). The CODP decouples planning based production from production to order. Planning based production uses forecasts to process finished goods, which means that the processed goods are stocked, until they are sold. Production to order means that an actual customer order is the trigger for producing the finished good. The stocking

point in the supply chain where planned production stops, and actual production to order starts is defined as the CODP. Figure 7-2 gives an overview of different locations for the CODP in the supply chain.

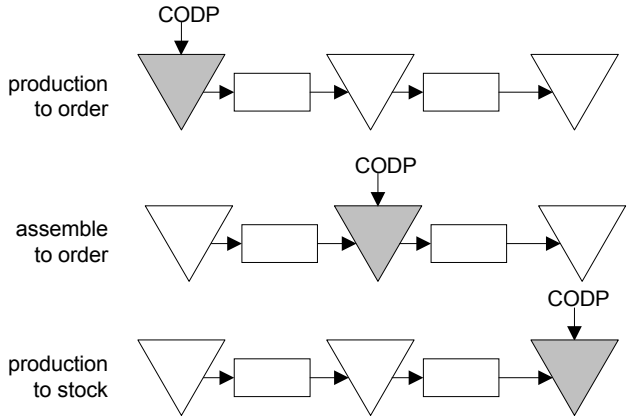


Figure 7-2: Different locations for the CODP

When the CODP is at the right of the supply chain, it means that all production is based on forecasts. Hence, all finished products are kept in stock until delivery to the customer. The advantage is that immediate delivery is possible. However, since the variety of end products is usually high, this results in a high inventory holding cost. When the CODP is at the left of the chain, this means that each order from a customer triggers the complete production process. This means that stocks in the rest of the supply chain are normally low (only inventory due to batch sizes), which results in a lower inventory holding cost. However, this comes at the cost of having customers wait for their orders. Normally this results in compressing the throughput time of the customer order so that it complies with the maximum lead-time the customer is willing to accept. This may result in much less than optimal batch sizes in the production process, and hence inefficient production, resulting in a higher production cost. Finally, the CODP may also be located in the middle of the supply chain, which means that planned production takes place for the half product, after which orders are assembled to order. An important advantage of this location is that the structure of the bill of materials shows that the smallest variety in products is normally at the half product level, which results in less inventory costs than in the production to stock situation. However, the best location for the CODP is very dependent on the exact supply chain situation.

After choosing the CODP for the supply network, the structure of the supply network is finished. As both design decisions show, the design of the structure focuses on the static aspects of the supply or distribution network. The next step is the design of the control structure.

Design decisions for the control structure
For designing the control structure, three important design choices are made:

1. Separation of overall control from local control. The logistic planning function focuses on controlling volume fluctuations in demand and the overall throughput time. The local

planning function focuses mainly on maintaining productivity in the department, while maintaining throughput time. The design choice is about where the boundaries of local control should be placed (e.g. which departments should be grouped or separated from a logistical point of view).

2. Controlling the bottleneck. The next decision is about determining which location in the supply chain is the bottleneck. The bottleneck location is defined as the most expensive process in the supply chain (in terms of resource consumption). The idea about bottleneck control is to situate the planning process on the bottleneck stage, because the slowest link in the production line determines the throughput of the overall chain. Through linking the planning of all other production stages to the bottleneck stage, an efficient and effective planning process is established.
3. Determination of control parameters. Finally, the different values for the control parameters are set. Figure 7-3 gives an overview of the different parameters that normally play a role in this decision.

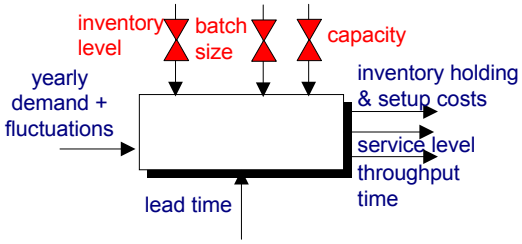


Figure 7-3: Control parameters

As Figure 7-3 shows, the level of inventory, the batch size, and production capacity are important control parameters that are used to balance logistic costs with throughput time, while achieving an acceptable service level. Although these parameters are important control parameters, other parameters may also play a role. For instance, on departmental level an important control parameter is the normative occupancy rate per department.

While we evaluate the physical and control structure, we see that the objective of both structures is different. In the design of the physical structure, the focus is on the *static structure* of the network, balancing the production and distribution costs of different arrangements. In the design of the control structure, the static structure is perceived as given. The focus is on controlling *the dynamics* of the network, balancing operational logistic costs such as the inventory holding cost and the setup cost with different levels of customer service.

7.2.2 Applying logistics to the information world

As we explained at the start of this chapter, the fact update distribution problem seems very similar to physical logistics, because fact updates, just as products must be delivered at the right time, at the right place, with the right quality. Therefore, we will look at the distribution of fact updates from a logistical point of view to see if design principles and choices from the world of logistics can be used to solve the fact update distribution problem. We will start with translating the concept of products from the logistical world to the information world and more specific to

our problem situation. Next, we will discuss similarities and differences between physical products and their counterparts in our problem situation. Finally, we will determine which concepts from designing logistical supply chains can be used for our problem, and which not.

Translating physical products to our problem situation
If we compare the world of logistics with the information world, products in the logistical world would be translated into information in the information world. The problem with information is that it exists in all kinds of forms. For transactions, the 'product' would be the invoice or order. For management control, the 'product' would be the business report such as the forecast, or sales report. For our problem, the 'product' is the fact update. We will define a fact update as the *change* in the state of affairs in the social world. Examples are changes in existing persons, products, locations etc. In summary, products from the logistical world translate into fact updates in our data alignment world.

Similarities and differences
A number of similarities exist between physical products and fact updates, a fact update, just as a physical product:

- can be identified (a new address of a person, a brochure of a new product, etc.);
- has certain characteristics (size, format, etc);
- has a certain quality (accuracy, timeliness, or completeness of the information in the fact update);
- has a certain value for a specific customer, depending on how it will be *used*;
- has an owner who may want to determine what happens with it, and who is held responsible for its quality;
- has a lead-time, meaning that the update should be processed into the receiver's database, *before* it is used in one of the applications of the receiver (Here, lead-time is defined as the maximum time that is available before the product or fact update is used in any application of the receiver);
- may deteriorate as a function of time. Just as physical products, information may become outdated.

However, in general, information is said to have certain characteristics that make it different from products. Namely, the production of information is cheap, because information can be copied endlessly and in virtually no time. Also, both time and costs of transporting and storing information are close to zero, but not zero.

We will examine whether these notions about information also apply to fact updates.

- Assumption: Production time and cost of fact updates are close to zero. A fact update consists of only a few kilobytes of information, and the number of fact updates per user is relatively small, e.g. a few hundred updates annually at the most. Hence, both the duplication time and the costs of fact updates are very low. However, this does not mean that the *production* cost of fact updates is low. Normally, the production of a fact update means that a change in the social world must be identified, interpreted, and entered into a database. This process is a manual process and normally takes considerable time (especially when the interpretation of a fact is disputed). It also uses considerable resources. Hence, the assumption is probably false for fact updates.

- Assumption: Transportation time of fact updates is close to zero. Even if this is true, the transportation time is only relevant in relation to the lead-time that consumers will allow for a specific update. For fact updates, the lead-time of a fact update is generally much longer than the transportation time, because information travels much faster than the physical counterpart of the fact update does. When a network like the Internet is used, and communication links are checked, the transportation time of a fact update is normally a few minutes at the latest, compared to hours or days for physical products. Consequently, the assumption is likely true for fact updates.
- Assumption: transportation cost of fact updates is close to zero. A fact update consists of only a few kilobytes, and each user sends only a few hundred updates annually. This is in contrast with other types of data exchanges, such as streaming video, which consists of hundreds of megabytes per transaction. An overrated estimate for a general user would be that, say 1000 updates annually of 10 Kb each = 10 Mb annually have to be transported. This amount of data could be easily transported over the Internet without any costs via a free Internet provider. Therefore, the assumption is probably true for fact updates.
- Assumption: the cost of handling, and storing fact updates is close to zero. Handling of physical products is similar to electronically receiving, and storing fact updates in a database. Although the volume of fact updates annually is relatively low (only 10 Mb), the costs of handling and storing are much higher, because a facility for handling, and storing fact updates must be operated. This means that the costs of IT personnel for database administration, and technical administration must be incorporated in the total costs. Because of this, the assumption seems to be false for the handling and storage costs of fact updates.

Comparing logistical network design with fact update network design
In Section 7.2.1, we discussed several design choices that play a central role in designing the physical and control structure of logistic networks. We will now examine if, and how these design choices play a role in fact update network design.

Static structure fact update distribution network
With respect to designing the static structure, the first important decision is the arrangement of processes, and stocking points in the logistic network. In their article about designing physical distribution networks, Mourits & Evers (1995) discuss the different design parameters that play a role in this decision. The objective of the network *design* problem is to determine how many warehouses are needed in a network, and where they should be geographically placed, such that *warehousing*, *transportation* and *inventory* holding costs are minimized, under the restrictions that a certain assortment of products is delivered to the right customer, within specific maximum time limits. For the fact update problem, the amount and geographic location of warehouses would translate to the amount, and geographic location of product databases in the network of suppliers and receivers. However, both these decision variables play no significant role in fact update distribution, because geographic location of product databases is not a major issue (under the assumption that the lead-time for delivering product updates worldwide is much higher than the transportation time. As we discussed before, this is a reasonable assumption for fact update distribution). Therefore, these specific decision variables are not useful for designing the static structure of the fact update distribution network, although some of them, such as the availability of a central database, or the operational costs of such a warehouse, may play a role.

With respect to the second decision for designing the static structure of a distribution network, namely the location of the CODP (see page 105), this concept is transferable to the fact update distribution problem. Figure 7-4 gives an example of 3 possible locations of the CODP for a product brochure.

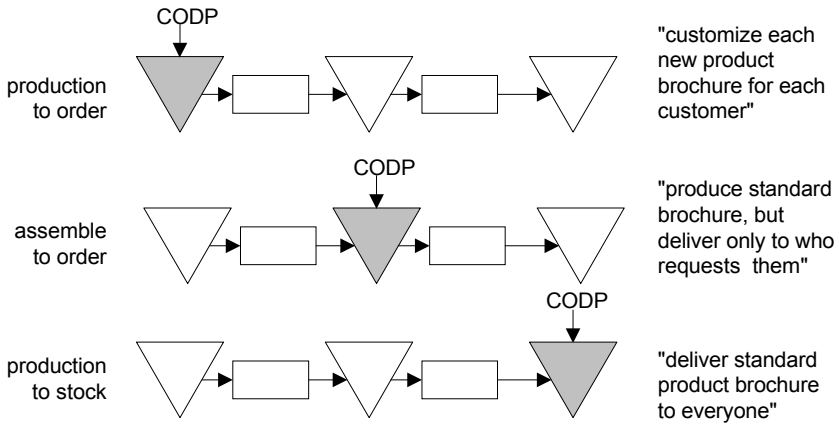


Figure 7-4: CODP for product brochure

In summary, for designing the static structure of the fact update distribution network, the arrangement of the network in terms of processes and the location of the CODP is still an issue, although the exact parameters from the logistical situation are not suitable.

Control structure fact update distribution network
The first question is what exactly determines the dynamics in the fact update distribution network. In the logistical control structure, the logistic cost is balanced with throughput time and service level (in terms of timeliness). Although timeliness certainly plays a role in fact update distribution, the next two chapters will show the importance of data quality as a performance measure for the fact update distribution network. In that case, the objective of the control structure of the fact update distribution network would be to dynamically balance the data quality checking cost with the remaining data quality level. If this is the case, this would mean that the emphasis shifts from time in logistics to data quality in fact update distribution.

If balancing the data quality checking cost with the remaining data quality level would be the objective of the control function, the three design choices for the control structure are affected as follows:

- With respect to the first decision, namely separation of logistic and local control, this would translate into separating localized data quality management from overall or network wide data quality management. This would certainly be necessary, since data quality errors in the beginning of the complete chain will have major impacts further up in the chain. Therefore, through monitoring and controlling data quality at the start of the chain, the quality of the overall chain could be effectively managed. This in

contrast with localized data quality management that should focus on realizing a certain, constant data quality level. The relation between the two levels would be that the constant level of the localized data quality management function would be set by the overall data quality management function.

- With respect to the second decision, namely controlling the bottleneck, this would translate into finding the location in the fact update chain, where the quality checking function is most expensive. That would be the starting point for controlling the overall data quality in the chain.
- With respect to the third decision, namely setting the control parameters, this would be the same in fact update distribution. The problem is that we do not know what exactly the control parameters are. These parameters should first be established.

In summary, for designing the dynamic or control structure of the fact update distribution network, the same type of decision is likely to be relevant. The major problem is that we do not know which parameters are relevant to the control structure of the fact update distribution network.

7.2.3 Conclusions

From this review of comparing logistics and fact updates, we draw two conclusions with respect to fact update distribution:

1. Many concepts from the logistics world seem applicable for designing our fact update distribution network. Especially the concept of the supply chain with both a static and a control (or dynamic) structure is applicable for fact update distribution. Furthermore, the type of design decisions that are relevant to each structure for physical distribution also seem relevant for fact update distribution. However, the exact parameters for both the static and dynamic structures are not similar to physical distribution, and should therefore be examined for fact update distribution.
2. The design of the logistics network showed that the static structure is created first, after which the control structure is designed since this reduces the complexity of the design problem. Therefore it seems reasonable to start with the question, which parameters determine the design of the *static* structure of a fact update distribution network in solving the fact update distribution problem. Only then the control part of the fact update distribution network should be addressed.

7.3 Requirements for a data alignment method

Using the original requirements from Chapter 1, our evaluation model from Chapter 6, and the conclusion about using logistics for fact update distribution in this chapter, we developed the following detailed requirements for our new data alignment method. These requirements are as follows.

For specifying agreements

a.	The method should specify agreements about the semantics of product data	Chapter 1, question 2
b.	The method should be based on a central ontology model per user group, which means mapping between local and central schema is left to the local users	Chapter 6, conclusion 1
c.	Preferably, the method should support community based agreement specification. This means that each specific user group must be able to define and extend their own central schema (as is available in XML)	Chapter 6, conclusion 2
d.	Preferably, the method should support semi-automated implementation of agreements (such as XML-EDI)	Chapter 6, conclusion 2

For the data distribution structure

e.	The method should specify a fact update distribution structure for aligning the product data across the network	Chapter 1, question 3
f.	Since the design parameters for a logistical network are not similar to the design parameters for a fact update distribution network, we should focus on examining which parameters determine the fact update distribution network	Chapter 7.2.3, conclusion 1
g.	The method should first focus on the parameters that determine the design of the *static* structure of a fact update distribution network. Only then, the coordination problem of the network should be addressed	Chapter 7.2.3 conclusion 2

General requirements for design methods

i.	A design method should have a problem representation	Chapter 2.3.1
j.	A design method should state objectives and constraints	Chapter 2.3.1
k.	A design method should describe different research paths for instance through describing different scenarios	Chapter 2.3.1
l.	A satisficing design method should have an explicit progress evaluation function; an optimizing method should have an optimization function	Chapter 2.3.1
m.	A design method should state how the results are evaluated	Chapter 2.3.1
n.	A design method should be hierarchic	Chapter 2.3.1

7.4 Data Alignment through Logistics

Based on the requirements for sophisticated data alignment, we developed a new method for data alignment specifically appropriate for large interorganizational business networks, which we named Data Alignment through Logistics (DAL). The DAL method aims to solve both the translation problem and the fact update distribution problem, using as much of the concepts in current integration and distribution methods as possible.

We will first give an overview of the DAL method, and then describe how the DAL method solves the translation and distribution problem below.

7.4.1 Overview DAL method

An overview of the DAL method is shown in Figure 7-5.

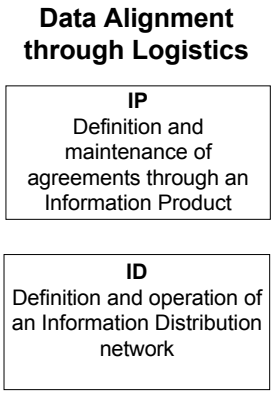


Figure 7-5: The DAL method

The method consists of two phases:

1. Definition and maintenance of the Information Product (IP). This means that a semantically rich schema is defined and maintained by a user community at sector level for the (product) information that needs to be exchanged.
2. Definition and operation of an Information Distribution (ID) structure. This means that based on the requirements of senders and users, a distribution structure is defined and operated that optimizes the costs and the resulting data quality of fact update distribution for the specific user community.

7.4.2 DAL and the translation problem

The translation problem in the DAL method is solved through the definition of an Information Product (IP). Similar to Wang (1998) we define an IP as:

> *A semantic data model, which is defined and maintained by a user community (i.e. the interorganizational network users), and which will be used for a specific purpose.*

With the term *semantic data model*, we mean that the agreements in the IP are modeled using a data-modeling tool. In this thesis, we used Entity Relationship modeling, which is also used for view integration. With the term *defined and maintained by a user community*, we guarantee that the IP reflects the up-to-date agreements of the sector community, which leads to a high

degree of semantic richness. With the properties of XML in mind, the use of XML for specifying these agreements would be obvious.

Furthermore, through specifying the *specific purpose of use* (or application) of the IP, we enable the definition of multiple shared schemas, even within a large user community. We prefer this, because in large user communities several subgroups normally exist. The idea behind using the objective of use as a subgroup selection mechanism is that the objective of use is a strong selection mechanism to find non-overlapping subgroups (marketing and logistics are different user groups, although they might use similar product data). This ensures that the views of different subgroups are captured adequately, resulting in a higher degree of semantic richness, while drastically reducing the amount of time needed for consensus. Thus, by viewing the central schema as a product, we provide focus in the definitions, so that a minimal set of mutual agreements is established.

If we compare the Information Product (IP) with the physical world, it would be translated to product *type*. In logistics we control the goods flow of products, which is in fact based on a product type, not on each individual product. Similarly, for our fact update distribution problem, we will want to control the flow of fact updates, which are based on a semantic model of facts. Thus, an IP is similar to a product type, and a fact update (or instantiated IP) to a specific product.

Examples of IPs for the food retail community are: Product Master data, Product Price information or Product Nutrition information. The IP *Product Master data* is used for logistical purposes and describes the product hierarchy, which relates consumer units (e.g. the smallest sellable unit in the retail store) to their trade units and their transport units (For instance, 40 chocolate drink cartons fit in one box. A pallet of chocolate drinks contains 9 boxes per layer with 4 layers). The Product Master data are used in many logistical processes, such as receiving goods in the warehouse, scanning products as the checkouts etc. The IP *Product Price information* is used by corporate purchasers for purchasing purposes and describes which price types exist (the consumer store price, standard selling price, the purchasing price), which bonus- and discount structures exist and how prices and bonus/discount structures are related. The IP *Product Nutrition information* is used, for instance, by dieticians in hospitals for diet composing purposes and describes how food ingredients, their expected effects and their way of preparation are related to each other.

As we explained before, an important characteristic of the definition of IPs in the DAL method is that the users themselves are responsible for making the mappings between the component data models of each user to the central community schema (the IP). In the tight coupling approach, the architects make this mapping. The main advantage is that this delegation of the mapping task reduces the complexity of this task. For instance, in the food retail sector in the Netherlands, about 1200 component supplier data models and 50 component retail data models have to be mapped to the central community schema. For a single architect, this task is virtually impossible to perform, but is relatively easy performed by the IT departments of the 1200 suppliers and 50 retailers. The reader is reminded that the central community schema is not prescriptive in that it forces the participants to adapt their component schemas to the central schema structure. Rather, the IT departments of the participants may structure their component schemas in whatever way they want. They only have to make sure that externally communicated information is structured according to the central community schema. In other

words, each participant may speak its own language at home, as long as they speak the community-defined language when they work together.

However, a major consequence of the *delegation* of the mapping task is that the mapping between the local data model and the IP may be wrongly implemented. Therefore, it is necessary to check whether update messages that are exchanged between companies comply with the centrally defined IP. This requires that a checking structure is set up within the specific community to check the conformance to the central IP.

7.4.3 DAL and the distribution problem

DAL specifically aims to solve the data distribution problem in a large interorganizational network. The distribution problem arises because updates of product information in source databases need to be translated, transported and finally checked and entered into a limited number of receiver's databases, which are actually users of the information. Thus, the right information needs to be at the right time at the right place. Within the DAL method, this problem is solved through the definition of a fact update distribution network that determines which receiver wants what information, if and how a central community schema (the IP) is maintained and if, where and how much data quality checking takes place.

The advantage of an update distribution network over using a view updating schema, such as in the tight coupling approach, is that a distribution network does not require that one system's architect knows precisely who needs what information on what time. Only then, the system architect can define the right queries to fill the view update schema. This means that distribution control is very centrally organized. In contrast, in a (logistical) distribution network, supply and demand coordinate the distribution. Because of this coordination mechanism, many local view-update schemas exist, which results in much more flexibility in addressing distribution changes. Thus, in this situation, distribution control is more decentralized.

7.5 The DAL method, first version

Although the general structure of the DAL method is clear, the question remains how both agreement definition, and implementation of a fact distribution structure takes place in the DAL method. In this section, we will develop a first version of the DAL method, using the knowledge from previous chapters and the requirements for the method to construct this first version.

7.5.1 Agreements definition

As discussed previously, for the definition of agreements to establish a shared schema, the DAL method will do this through the definition of Information Products (IPs) for different purposes in a (large) user community. These IPs will be modeled using Entity Relationship modeling since XML was not available when the cases were conducted. This leads to the following first step of the DAL method first version:

Steps in IP

Step 1: definition of the IP

In the first step we define the objective of using the IP, and the data model that specifies the IP.

7.5.2 Definition of a fact update distribution network

Now we have established that we want to develop a fact update distribution network as part of the DAL method, the next question is how we are going to do that. Following the requirements for the DAL method (see Section 7.3), we will start with designing the *static structure* of the fact update distribution network. Since we cannot use the parameters from physical distribution theory (conclusion 1 in Section 7.2.3), we will have to examine which parameters determine the static structure and which parameters control the network. In Section 7.4.3, we described the process of fact update distribution as the process of translating, transporting, and finally checking fact updates from a source database to a limited number of receiver databases. Based on this view of fact update distribution we are able to construct a first version of a model that determines the goals, inputs, outputs and design variables of the *static* structure of a fact update distribution system. This first model of the static structure of the fact update distribution network is shown in Figure 7-6.

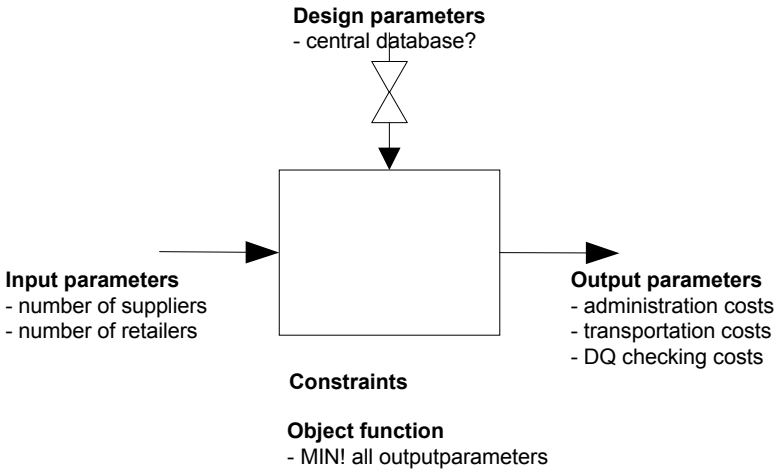


Figure 7-6: An update distribution structure

We will discuss the model shortly. The objective is to determine how the fact update distribution network must be designed to balance the costs of administering schemas and transporting new updates with the cost of checking the data in a network consisting of a limited set of suppliers and retailers under a number of constraints. One possible design parameter may

be the inclusion of a central database system, for example, a central data cross-docking station, to enable centralized data quality checking.

In defining the steps in the first version of the DAL method with respect to fact update distribution, we will use the above model of a fact update distribution network, and structure them according to Simon's steps for developing a design method. We will add one step for analyzing the current and desired situation. This leads to the next four steps for the DAL method, first version:

Steps in UD

Step 2: Qualitative description of the situation
In the second step, we first analyze the current and desired situation, so that we can represent the distribution problem in terms of *design* parameters that determine the static structure of the distribution problem.

Step 3: Abstract definition of the design problem
In the third step, we develop a representation of the distribution problem, consisting of input- and output parameters, network design parameters, objective and constraints.

Step 4: Scenario formulation
Based on the representation of the distribution problem, we define the problem space through formulating all possible scenarios.

Step 5: Evaluation
Finally, in step 5, we evaluate all scenarios, using the objective function.

7.6 The DAL method compared to other approaches

We will compare the DAL method (first version) we developed with the data integration and distribution approaches that we discussed in Chapter 6. To do this, we will use the parameters of Goodhue's model to evaluate them. These parameters are:

1. The cost of integration;
2. The flexibility to introduce local design changes;
3. The ability to share data. Here, we will define the ability to share data as the degree of semantic richness.

The three parameters above only address the problem of data integration. Therefore, we add one parameter from the perspective of the distribution problem, which is:

4. The amount of work for data quality checking.

We have used these parameters to evaluate the eight approaches for data integration and distribution, including the Data Alignment through Logistics approach. The results are shown in
Table 7-1.

Problem	Tight coupling	CM approach	Loose coupling	Traditional EDI	BIM	OO-edi, Corba/Dcom	XML-EDI	PDI/Step	DAL
Integration cost	--	+	++	-	+	+	+	+	+
Loc. design flexibility	--	+	++	+	+	+	+	+	+
Ability to share data	++	+	-	-	+	+	++	+	++
DQ checking cost	++	N/A	--	--	N/A	N/A	N/A	N/A	+

Legenda	++	Excellent	--	Very poor	N/A	not applicable
	+	Good	-	Poor		

Table 7-1: Evaluation

With respect to the *cost of integration*, the tight coupling approach is clearly the most expensive (score: very poor), because integrating the schemas of more than a few hundred participants is virtually impossible. Furthermore, the integration cost of traditional EDI is high (score: very poor), because for each new relationship, many implementation details have to be added in the translation program. Since the other approaches include many of the implementation details already in their central schemas, their integration costs are considerably lower. The integration costs of the loose coupling approach are the lowest, since no integration is needed whatsoever. In this approach, the human users solve all translation problems.

With respect to the *flexibility to introduce local design changes*, again the tight coupling approach scores very low (score: very poor), since every local design change has to be analyzed and approved at the central level. Because all other approaches use a central or community based schema, the local design changes are decoupled from the central schema. Thus, at the local level many design changes can be implemented without affecting the central schemas, resulting in a good flexibility score. The only consequence is that the mappings from the local schemas to the central or community schema must be re-established. Although traditional EDI does not use a central schema, the consequences of changes in mappings for the translation process are the same as for the other approaches. The loose coupling approach has the highest score, since schema changes do not affect the mechanism of the translation process in the loose coupling approach (except when the syntax of the schemas is changed).

With respect to the *ability to share data,* both the loose coupling approach and traditional EDI have the lowest score, because they do not support the semantic level of communication. Since the other approaches use a shared schema, their ability to share data is high, which results in a score of ‘good’. With respect to community based agreements, the tight coupling approach, XML, and the DAL method score ‘excellent’, because both XML-EDI, and DAL have the highest degree of semantic richness. This is also true for the tight coupling approach, because once the schema is created, it incorporates all user and sub-user group requirements. The question is whether this is possible.

With respect to the *data quality checking cost*, the loose coupling approach and EDI score very low (score is very poor), because all data is checked manually at the purchasing department. Because the DAL method enables automatic data quality checking (partly), the cost of data quality checking scored ‘good’. Finally, the tight coupling approach scores excellent, because

data quality checking is not required (since all data is already integrated). All other approaches do not address the data distribution problem and hence have a score of ‘N/A’.

In the next chapter, we will apply the DAL method first version in two case studies to further develop and test the applicability of the DAL method.

8. Two DAL Cases

8.1 Introduction

In the last two chapters we reviewed the theory on data alignment, after which we introduced the Data Alignment through Logistics (DAL) method. In this chapter we will further specify and test the DAL method in two cases where the method was applied. In Section 8.2 we will start with defining the research strategy, which is based on the evolving serial case study. Next, in Section 8.3, we will discuss the CBL case, where we applied the DAL method first version which we developed at the end of Chapter 7. In Section 8.4 we will introduce a second version of the DAL method, based on the reflection in the CBL case. In Section 8.5 we will then describe how we applied this second version in the EAN-DAS case. This chapter ends with conclusions in Section 8.6.

8.2 Research strategy

The problem of design research is that we basically have to satisfy two different objectives: (1) we need to solve a practical design problem, and (2) we want to ask ourselves new research questions we want to study in a new situation so that we can improve our method. As described in Section 2.3.2, we will follow the serial evolving case study strategy to study our new design method. The strength of this strategy is that during the case it focuses on solving the case objective, while after the case reflection is used to continually improve the method. In these two cases we chose to detail the evolving case study strategy one step further in two ways:

1. We will first describe the case setting and the practical case objective that solves the practical design problem, which will be discussed in the case conclusions. Next, we will provide focus to our reflection through asking ourselves *beforehand* what the research case objective is.

2. Since we want to develop a design method following the (classical) principles of scientific methodology, we need to test the method. However, this is contradictory to the Singerian inquiry strategy where the design method changes with each new implementation. We tried to solve this problem through splitting up each case in a testing part and an exploratory part. In the testing part, concepts in the design method that were already found are tested in a new case. In the exploratory part, through critical questions beforehand and through reflection, we further expand the method with each new implementation.

Both steps are incorporated in our case study designs.

8.3 The CBL case

8.3.1 Case setting

The case was conducted in the Dutch grocery sector. The direct motivation for the case was the request of the Dutch Retailer agency, CBL (Centraal Bureau voor Levensmiddelenhandel) to EAN Nederland to implement a central database system for the collection and distribution of its category data. EAN Nederland is the Dutch EAN organization. As such, EAN Nederland provides two services for its members: article coding and EDI message development. The CBL is the representation agency for the Dutch retailers. One of their activities is the administration of the CBL category system, a product classification system of the complete supermarket assortment. Since the task of collecting all EDI article codes to assign the appropriate CBL codes is so labor intensive, the CBL seeks other ways to facilitate this task. An important solution for the CBL would be if EAN Nederland would collect the article data in a central database. Such a database could improve the quality of EDI significantly. Since EAN Nederland is always seeking new ways to improve the service to their members, the CBL thinks that EAN Nederland should carry out this task.

However, EAN Nederland doubts whether a central database solution is the best way to support the distribution of the CBL category data. Therefore, a case study was set up with the following practical objective:

> *Investigate what would be the best way to collect and distribute the CBL category data.*

8.3.2 Case Design

For describing the design of the case we used the criteria for case studies as defined in Section 2.4.

Type of study
In this case we want to further develop the Data Alignment through Logistics (DAL) method. Therefore, we want to apply it in the food sector to design an optimal distribution structure for category data in a large interorganizational network. Since the DAL method first version, which we will use, is still in its infancy stage, we classify this case mainly as *exploratory*, although we will test the applicability of data modeling to define the IP. Furthermore, since our level of

analysis is the whole food sector where we only interview specific suppliers and retailers to gather information, we classify this case as *holistic*.

Case research objective and case questions
The case research objective is defined as follows:

> *Develop the contents of the DAL method, through using the DAL method first version in the grocery sector for designing an optimal distribution structure for category data.*

As we described at the end of Chapter 7, the DAL method basically consists of two phases: (1) Definition of the information product, and (2) Definition of an optimal distribution structure. Although we have a clear picture of how to solve the first part (which we will do through using a data modeling method), the content of the second part is still largely unclear. With an unclear content is meant that we already defined the steps for the second part of the method (step 2-5). However, it is not clear of what input, output and design parameters the distribution system consists of, which costs are relevant for evaluating the optimal situation, and how different distribution scenarios could be modeled.

We developed the following research questions for this case:

With respect to the first part of the DAL method
1. How important is the strict definition of the objective of the IP?
2. Is the data modeling approach effective in defining an IP, or should other alternatives be considered?

With respect tot the second part of the DAL method
Taking a systems view of the problem:

3. What parameters determine the input, the output and the structure of the category data distribution situation?
4. What is the objective of the optimal distribution situation and which constraints apply?
5. Which cost types can we identify in the object function?

Thus, this case consists of a testing and an explorative part. With respect to defining the IP, we will test our view that it is important to first define an objective and that Data Modeling is sufficient to describe the IP. With respect to determining an optimal distribution situation, we will explore which parameters describe the design situation and which costs do apply.

Case choice criteria
As selection criteria for this case we applied the criteria that we identified in the research objectives in Chapter 1. These objectives describe that the research will focus on product information distribution in the grocery sector, which will be studied on sector level. Since category information is part of product information, and since the problem addresses category information distribution on sector level in the Dutch grocery sector, we concluded that the CBL case satisfied our criteria.

Case design
Since we analyze the complete interorganizational network of all food suppliers and retailers on sector level in the Dutch grocery sector, one unit of analysis is on sector level. Although we will interview individual companies to collect data, the results are only used on sector level, not on organizational level. Hence, there is only one unit of analysis, which is defined on sector level. The case design is holistic, thus permitting different kinds of results, which do not have to be defined in detail beforehand. Clearly, this is an important characteristic of an exploratory case study.

Criteria for interpreting findings
Since this is mainly an exploratory case, we only identified criteria to interpret the first step of the DAL method. We defined the following two criteria:

- Does the steering committee which used the results of the method, acknowledge the usefulness of first defining the objective, or is this step unnecessary?
- In the eyes of the steering committee, is the data model of the IP we developed complete and does it reflect the information the committee wants?

8.3.3 Case methodology

Data was collected following the steps defined in the DAL method first version. We will discuss each step in the method shortly.

Step 1: Definition of the IP, and 2: Qualitative description of the current situation
For (1) defining the precise objective of the category information, and getting a qualitative description of (2) the current situation and (3) the desired situation, we interviewed CBL (the Dutch retailer agency), SMA (the Dutch Food Manufacturers agency), EAN Nederland (the Dutch EAN standardization organization) and two retailers using a structured questionnaire. In this questionnaire, the same questions concerning the above three subjects were asked to each of the interviewees. Furthermore, we conducted desk research of product offerings and annual reports of two Market Research Organizations (MROs) to investigate how they collected and used category information. Finally, we did some literature search to define the different objectives of classification systems.

Step 3: Abstract definition of the design problem
As described in Chapter 7, we asked ourselves: "which parameters define the distribution situation?" To determine these parameters, we developed several questions in the questionnaire concerning the optimal situation:

- We identified *input and output parameters* through asking all interviewees what the characteristics of the distribution situation were.
- We identified the *object function* and the *design parameters* through asking all interviewees what the preferred situation would be. First, this forced them to determine what their objectives were. Second, through determining the 'preferred' situation we could establish what the optimal situation was in their view (and thus what the object function was). After that, we discussed with EAN Nederland (as the

main stakeholder) which values each design parameter could have, because the range of values per design parameter determines the problem space.
- Finally, we developed *constraints* through asking what the minimum conditions were for the new situation.

Step 4: Scenario formulation
Since we identified design parameters and a limited number of values for each parameter, we were able to develop a limited number of scenarios ourselves.

Step 5: Evaluation
Since only four scenarios resulted from the scenario formulation phase, it was quite simple to make a quantitative evaluation of each of the scenarios. The final result was presented to the steering committee, which then did an evaluation.

8.3.4 Case results

In this section we will describe the results of applying the DAL method first version.

Step 1: Defining the information product

The CBL category system
The CBL category system is a product classification system of the complete Dutch supermarket assortment. The system consists of 19 sections, 97 groups and 374 subgroups (see Bakkenist 1996). The Dutch Retailer organization CBL (Centraal Bureau voor Levensmiddelenhandel) each year draws up the category system and has been doing this since 1992.

Step 1.1 Defining the objective of the IP
According to the first step of the DAL method, we should first determine the exact objective of using category information. Based on interviews with retailers and the CBL and desk research about classification systems, we initially found two objectives for category information:

- Retailers want standardized categories to support the assortment search process;
- Retailers want category information to evaluate the performance of their products in a certain category compared to their competitors. To accomplish this, they need a standardized category classification.

After a lot of discussion, we found an important difference between both objectives. For the first objective it is extremely important that the categories reflect the mutual understanding of all retailers about the average assortment in a supermarket. The question is, if this is necessary for a good search instrument. Defining a common search structure is extremely difficult and perhaps even impossible. Furthermore, it might not even be necessary to accomplish such a universal structure, because a global structure that is further detailed by each retailer might be a sufficient solution. Finally, it will take a lot of time to construct such a structure.
However, for the second objective, to evaluate each other's performance it is more important *that* a product is classified in a specific structure than *how* that product is classified in that structure (as is the case in the first objective).

After this difference became clear, the steering committee chose the second option as the main objective of the CBL classification. Because of the difference that was established, the initial important requirement that the classification should be a mutually agreed classification was dropped. Instead, the requirement that all articles *must* be classified was determined to be the basis of the new classification.

Step 1.2 Data model IP
After we established the exact objective of the CBL category system we could simply define the contents of the CBL category system. It was established that the CBL category system should include at least:

- The CBL classification codes;
- All EAN article codes that belong to each CBL category.

Furthermore, to improve the usability of the classification, the IP was extended with:

- The CBL classification description. This is important for the users of the classification;
- The EAN article description. This is important to support the classification process itself.

There are two remarks we want to make at this point. One is to notice that important information for searching, such as synonyms and keywords, is excluded from the definition. And the other, that because of the focus on the exact objective of the category information, the structure of the remaining information became extremely simple, consisting basically of one entity and four fields. Hence because of this a graphical data model was not further developed.

Step 2: Describing the current situation

From the interviews we learned about the current situation according to the different stakeholders and their perceptions of the future situation. To enable the reader to evaluate how we arrived at an abstract problem description (in step 3), we will first provide a detailed outline of the current and desired situation.

Current distribution situation
The suppliers send information about new products to the retailer using product brochures, promotion schedules or standard product forms[1]. When the retailer receives this information, he enters it in his article registration system. Since all retailers have their own article classification, they perform both the task of administration of the classification and the task of actual assignment of classification codes to the EAN codes by themselves. Thus, during the data-entry process of new articles, a retailer specific classification code is assigned to the new article. All article data are stored in the Product Master database (P/M file).

Each month, a number of retailers supply the scanning data of a selected number of outlets to the MROs. This means that the MRO receives all new articles as well. This data is entered into

[1] A standard product form is a form issued by the retailer, which contains a number of required information items that the supplier has to fill out.

the P/M file of the MRO, during which an MRO specific classification code is assigned to the articles. The MRO uses the data in the P/M file to generate market reports for the retailers (and suppliers). These reports contain sales- and turnover figures for each retailer per MRO category compared with the totals of the MRO categories. These reports are sent to the retailers in hard copy or on file. The retailers then use these figures to measure their performance compared to others and to support the assortment management process. Figure 8-1 gives an overview of the current situation.

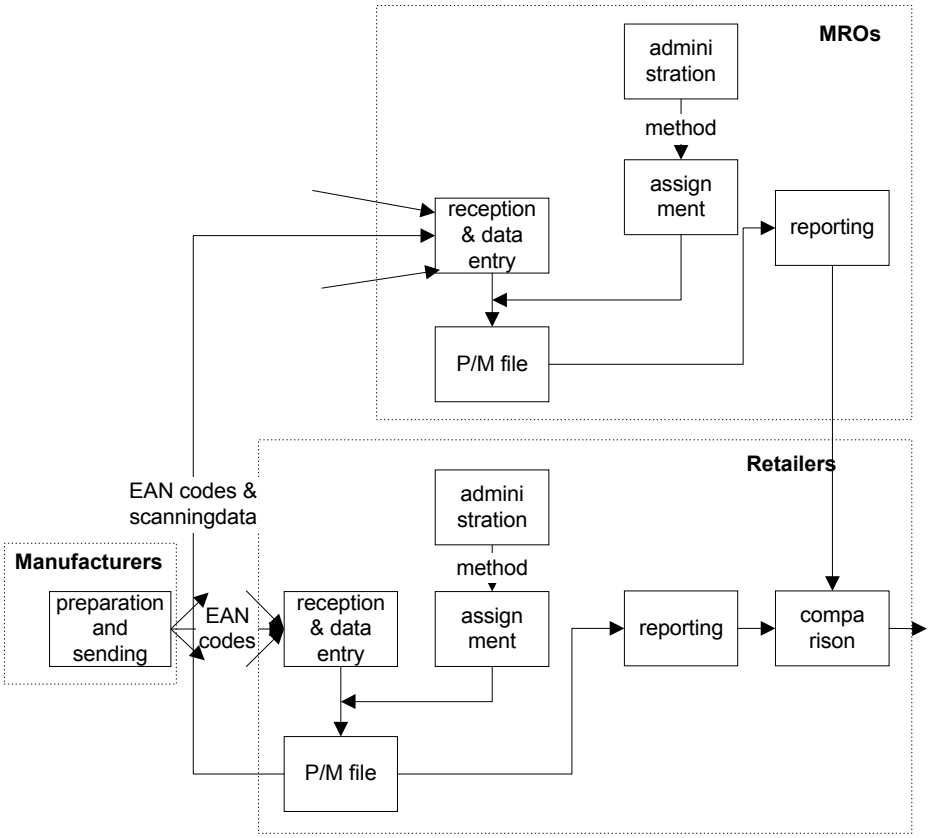


Figure 8-1: Current distribution situation

The CBL solutions for the desired situation
Since the current situation was not tolerable, the CBL already started to make a detailed CBL category system. It collected all EAN-codes from all its retailers, which it assigned to categories of the existing CBL classification. This category system was stored on a floppy disk and selectively distributed to the retailers (the retailers received only those EAN-codes they supplied. Every now and then it repeats this procedure. We will refer to this scenario as the CBL interim scenario. However, this interim scenario was not preferable to the CBL, since both the collection of EAN-codes and the assignment to categories involved a lot of manual labor. The CBL suggested another solution, based on a central database, administered by EAN

Nederland. In this central database scenario, the suppliers send all article data (not only the EAN codes, but also the article master data) to a central database. From there the article data is selectively distributed to the retailers. Using the database, the CBL would save much time in collecting the EAN article codes. Both scenarios will be discussed in greater detail in the scenario section.

Step 3: Abstract definition of the design problem

Step 3.1 Input and output parameters
As described in the case methodology, we first identified the characteristics of the problem situation to develop the in- and output parameters. In the food sector, there are approximately 5000 suppliers who supply their products to circa 35 retailers. Each week about 400 new articles are introduced, of which at least 70% are newer versions of existing articles. Each retailer has approximately 20,000 articles in his assortment. The complete retail assortment in the sector consists of approximately 100,000 articles. Concerning the costs, we found that a central database system costs at least 100,000 NLG annually. The costs of assigning CBL codes to articles are not exactly known. Finally, the probable number of assignment errors was estimated for the situation that the suppliers would be assigning the CBL codes to the EAN codes, instead of the CBL itself. This estimate is based on the experiences of the retail panel, which at the moment performs this activity. The two retailers (from the panel), and the CBL who we interviewed, estimated that 0.5% of all new articles are probably erroneously assigned. This number is extremely low, because the assignment task is very simple (at least 70% of the articles is based on an older version).

From this description, we identified the following input- and output parameters:

Input parameters
- The number of suppliers;
- The number of retailers;
- The number of MROs;
- The number of new articles per time period.

Output parameters
- The cost of assigning and distributing the updates
- The remaining number of assignment errors.

Step 3.2 Design parameters
The major problem was to identify the parameters that determine the structure of the distribution situation, so that we can develop a number of scenarios. The CBL offered the first important parameter, since it indicated that it preferred a central database solution, administered by EAN Nederland. Thus, the first parameter is centralized versus bilateral exchange of article data. This parameter played already an important role in the three investigative cases. To discover the other parameters, we examined the current situation more closely.

From the description of the current situation, we can see that the complete classification process basically consists of two functions, the *administration function* and *the assignment function*. Administration of the classification means that new classes are added or old classes deleted

based on the evolving set of articles of the classification. Thus, the administration function determines the classes to which the articles have to be assigned. In the assignment function new articles are assigned to the existing classes. To do this, an overview of classes and a description of their characteristics are needed. This overview should be provided by the administration function.

Since these functions can be split up and fulfilled by different stakeholders, the locations of both functions form two important design parameters for the distribution scenarios. Thus, we discovered three design parameters:

1. How to distribute the data? (Centralized versus decentralized distribution). With centralized distribution we mean that all data from the suppliers is sent to a central database, administered by EAN Nederland, and then distributed to the retailers. Decentralized distribution means that the suppliers send their data directly to the retailers.
2. Where to locate the administration function (CBL). For the administration function, only the CBL is a serious candidate. The suppliers do not appear to be an option, because they have no interest in the classification. The retailers are not either, since it would mean that each time a new article is presented, 35 retailers would have to decide on new or old categories. The only way it could work s through establishment of a small panel of representatives, which in this case is already the CBL. Also, the MROs form no good alternative, since there are at least three different MROs that already in the past could not agree on one category system. Therefore, we decided that only the CBL is a serious candidate for the administration function.
3. Where to locate the assignment function (CBL or suppliers). For the assignment function there are two serious alternatives, the CBL or the suppliers. The retailers and the MROs form no real alternative, because this would mean that each new article is more than once assigned to a category (at least 3 times in the case of the MROs and 35 times in the case of the retailers).

Step 3.3 Objective
Knowing the output parameters and the design parameters, we were able to define the objective of the optimal distribution situation. The objective was:

> to select that certain scenario, which proved to have the lowest costs of distribution, administration and assignment, and which has the lowest remaining number of assignment errors.

Step 3.4 Constraints
From the interviewees we learned the following additional requirements for the optimal distribution situation (which could prove to be constraints):

- Restricted access. The retailers demanded that not everybody receives the complete CBL category system because they feared market transparency.
- No deep classification. Both suppliers and MROs required that the classification does not have too many categories. Otherwise this could interfere with their much more detailed supplier categories.

- Neutral party. To prevent distortions in the balance of power, the suppliers required that a central database is set up and that it be administered by a neutral third party, such as EAN Nederland.

For the abstract problem description, none of these requirements is relevant. They must all be addressed during implementation.

Step 4: Scenarios
Since the first and third design parameter each have two values, and the second only one value, ultimately 4 scenarios are possible, which are shown in Figure 8-2.

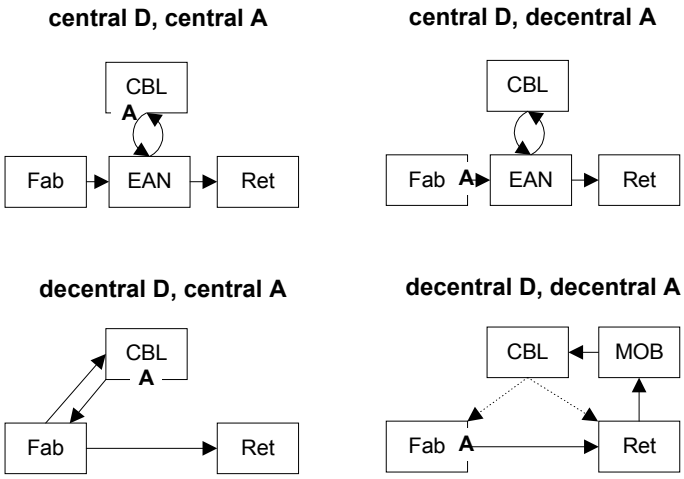


Figure 8-2: Four scenarios

Centralized distribution, centralized assignment
In the centralized distribution with centralized assignment scenario the supplier sends the article data to a central database administered by EAN-Nederland. The CBL regularly collects the new articles and assigns the right CBL codes to the new EAN codes. Finally, the new article data is selectively distributed to the retailer.

Centralized distribution, decentralized assignment
The centralized distribution with decentralized assignment scenario is largely the same as the previous scenario, except that the supplier assigns the CBL codes to the EAN codes. At the end of each day, the CBL checks the assignments and if necessary, changes them. Both the centralized distribution with centralized assignment and decentralized assignment are examples of the central database scenario the CBL suggested.

Decentralized distribution, centralized assignment
The decentralized distribution with centralized assignment scenario means that the suppliers send the EAN codes directly to the retailers, while the CBL assigns the CBL codes. The proposed CBL interim scenario is an instance of this scenario, although it is not suitable for our purpose. Since in this interim scenario the data is collected about once every two months (not

regularly), the MROs will be missing all new article numbers during these two months. Thus, this scenario falls outside the scope of our problem space.

The only way that centralized assignment is possible within our problem space is when the assignment takes place before distribution. This means that the suppliers first have to send all new EAN article codes to the CBL, which assigns the right CBL code. Then the assigned articles are sent back to the supplier who then sends the articles to the retailers.

Decentralized distribution, decentralized assignment
Finally, in the decentralized distribution with decentralized assignment the supplier assigns the CBL codes to the EAN code and then sends them directly to the retailer. At regular moments, the CBL collects all new article codes in the market. The fastest way to do this is via the MROs and not via the retailers, since there are 35 retailers and only 3 MROs. The CBL *checks* (not assigns) all article codes and sends the rectified codes to all participants, who then rectify their own databases.

Step 5: Evaluation
This case was mainly based on a qualitative evaluation, where we used some quantitative arguments. Here, we will only discuss the quantitative arguments we used to arrive at the best scenario from a quantitative viewpoint.

As defined in the object function, the administration cost, the assignment cost, the distribution cost and the remaining number of errors determine what the best scenario is. Since the administration cost is the same for all scenarios, we will exclude them from the evaluation. To evaluate the scenarios we will first quantify the different decision factors.

Assignment cost
From the interview with CBL, we know that the decentralized distribution with central assignment scenario practically means that the CBL receives all new articles on fax. There are 400 new articles per week meaning 80 new articles per working day. Answering 80 faxes a day is a fulltime job, with an estimated cost of 70,000 NLG annually. However, when the articles are solely checked (as is the case in the decentralized assignment scenarios) we can take advantage of the fact that the new category classification is required just once per month (because only then the MROs use the new numbers to generate their market reports). Thus, in the checking situation, the CBL will receive a list with of all last month's articles that have to be checked, which will consist of 400 articles per week * 52/12 = 1733 articles per month. With an estimated working speed of 10 articles per minute (because the checking task is very easy), we estimate that an average person requires 1733/(10*60) = 3 hours per month, which adds up to 34 hours annually for both decentralized assignment scenarios. With an average of 1600 working hours annually, this is equal to 34/1600 * 70,000 NLG = 1500 NLG annually. Finally, in the central distribution with central assignment scenario, the articles are not delivered per fax, but automatically. That situation is largely comparable to the decentralized assignment situations, because the CBL will be able to assign all articles once per month, after which the new assignments are distributed. However, this will require that the distribution of categories and the article master data are separated (because the retailers will not accept that they receive article master data only once per month).

Distribution cost
As part of the case, we also investigated other centralized database initiatives (which were the Switzerland central database, Sinfos, the GPI database and Productview, see Bakkenist Management Consultants 1996, pp. 22). Here, we learned that the exploitation costs are on average 200,000 NLG annually, excluding the costs of software development and maintenance, customer support and marketing and sales. Thus, centralized distribution will cost at least 200,000 NLG annually. On the other hand, decentralized distribution is free of charge as long as the category information is send as part of the article master data, which is currently the case.

Remaining assignment errors
All scenarios have an error rate of 0%, except for the decentralized distribution with decentralized assignment scenario, which has an error rate of 0.5%.

8.3.5 Case conclusions

The practical case objective was to determine the best way the CBL category data could be distributed. In Table 8-1 we have summarized the quantitative results of all four distribution scenarios we developed.

	Central D, Central A	Central D, Decentralized A
Assignment cost (NLG/yr)	1.500	1.500
Distribution cost (NLG/yr)	> 200,000	> 200,000
% Assignment errors	0%	0%
	Decentralized D, Central A	**Decentralized D, Decentralized A**
Assignment cost (NLG/yr)	70,000	1.500
Distribution cost (NLG/yr)	0	0
% Assignment errors	0%	0.50%

Table 8-1: Summary of quantitative results

The table shows that the decentralized distribution with decentralized assignment has the lowest costs. However, an error rate of 0.5 % remains. Since this would mean that the CBL has to send all participants (5000 suppliers and 35 retailers) an update of the 'right' classification at least each month, which leads to all kinds of updates in the databases of all participants, the steering committee rejected this scenario. Therefore, from this quantitative evaluation, the decentralized distribution with centralized assignment scenario remained as the preferred scenario.

However, during the evaluation session with the steering committee, it was concluded that the total costs of centralized distribution (>200,000 NLG) was no good estimate, because these costs would be 0 NLG, if the category data were distributed together with the normal article master data (as is currently the case in the decentralized distribution scenarios). This lead to the final conclusion that we first had to investigate what the best distribution method for article master data was.

8.3.6 Reflection

The research objectives of the case were to test the first step of the DAL method and to develop the DAL method further, specifically with respect to the parameters that define the distribution situation, the costs that are incorporated in the object function and the graphical representation of different alternatives. Based on the case results as discussed in the last section, we arrived at the following conclusions:

With respect to the first step of the DAL method (Question 1-2)

1. The case confirms the importance of strictly defining the objective of the information. Only because we separated the search objective from the standardization objective, the discussion in the steering committee about category information could be resolved. Afterwards the steering committee acknowledged the importance of making this distinction.
2. The effectiveness of using a data modeling approach to describe the IP could not be tested, because data modeling was not deemed necessary, since the resulting IP consisted of only 4 data fields.

With respect to the second part of the DAL method (Q 3-6)

3. Using a systems view, we were able to develop an abstract description of the problem situation, which is displayed in Figure 8-3.

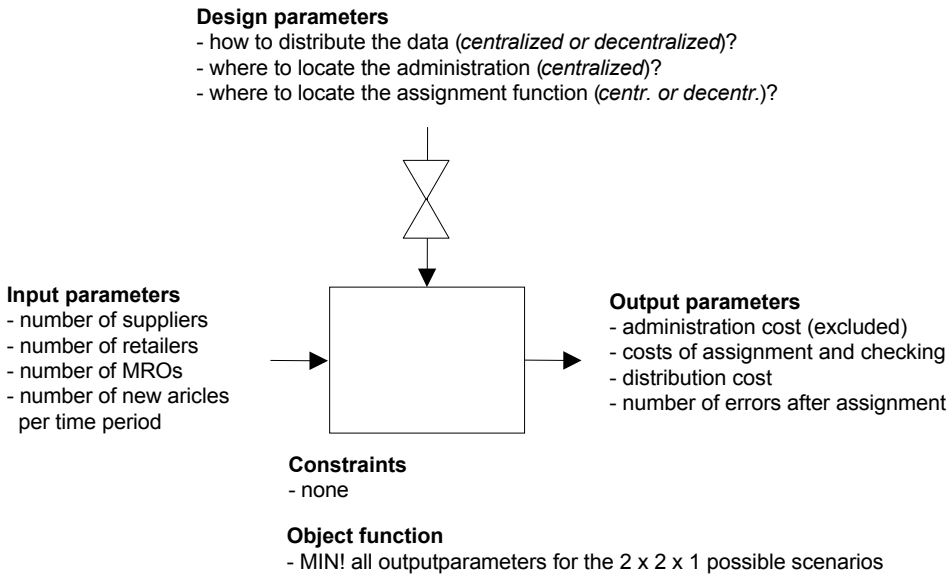


Figure 8-3: All parameters that define the distribution situation

This abstract problem description is important for the DAL method, because it extends our method with a detailed view of the design parameters that determine the design situation and the costs that are involved to evaluate the design situation.

4. Although the abstract problem description clarified the discussion about an optimal solution, an optimal solution could not be selected, because there was discussion about what the cost of central distribution is: 0 NLG (if the category data was distributed together with the article master data), or 200,000 NLG if not. Thus, the method did not lead to a choice of one of the alternatives.

8.4 Second version of the CBL case

Based on the results of the CBL case, we developed a second version of the DAL method. This version is similar to the first version of the DAL method, except that step 3 in the DAL method is extended with the parameters and object function as described in Figure 8-3. This extended version of the DAL method will be reapplied in the EAN-DAS case, where we will further explore the method and specifically test the abstract problem description that we found in the CBL case.

8.5 The EAN-DAS Case

8.5.1 Case setting

In the Netherlands, Efficient Consumer Response (ECR) is becoming increasingly important. To enable seamless electronic communication, which is a basic requirement for ECR, the data between the databases in the supply chain must be fully aligned. At the time the case study was conducted, EAN Nederland was developing a new system, the EAN-Data Alignment Service (EAN-DAS system). The objective of this system is to support full data alignment of article master data in the food sector. The EAN-DAS system is based on several concepts from this thesis, such as the concept of article information as information products that are distributed (or even cross-docked) through the interorganizational network.

The EAN-DAS system is a central database system that collects data from the suppliers, checks the data for errors and then distributes it to the retailers. Although the database is a central system, suppliers and retailers keep control over the contents of the database, and control over which data is distributed to whom. This is accomplished through a distribution mechanism that is based on filters. The supplier may determine which data is sent to whom. The retailer's filter uses assortment definitions that determine from which products the retailer receives which data fields.

An important question for both EAN Nederland and the companies that are involved in the EAN-DAS pilot was the one of added value of such a central system. For data alignment can also be achieved using the bilateral exchange of EDI article messages. There were different opinions in the steering committee about which solution would prove to be best. With the members of the EAN pilot group we agreed that we would investigate the added value of both scenarios *using the DAL method*. This leads to the following practical case objective:

Analyze which scenario is best, the centralized EAN-DAS, or decentralized bilateral exchange scenario compared to the current paper based situation and to each other, for the distribution of Article Master Data in the Dutch Grocery Sector.

8.5.2 Case design

To describe the design of the case, we used the criteria for case studies as defined in Section 2.4.

Type of study
Since the main objective of the study is to test the parameters in the abstract problem description of the CBL case, we classify this second case as mainly *explanatory*. However, other aspects of the DAL method are also further explored. Additionally, since we analyze the data on one level, (namely on *chain* level) and use the results to determine which scenario is preferable on *sector* level, we classify the case as embedded.

Case research objective and questions
The main objective of this case is defined as follows:

> Test the parameters in the abstract problem description we extracted from the CBL case through reapplying the DAL method, focusing on the functions that are performed in two extreme distribution scenarios.

This leads to the following main research questions:

1. Which parameters (and especially design parameters) determine the structure of the distribution situation?

We have several issues that we want to explore further:

2. The CBL case showed that the quantitative evaluation of alternatives was not sufficient because there were doubts about *how* to measure the costs of different distribution scenarios. Therefore, we want to explore further how we should measure different costs, such as distribution and assignment costs.
3. Although we developed a form of representation in the CBL case for the different scenarios in a very intuitive way, we asked ourselves if we could develop a more formalized form of graphical representation.

Case selection criteria
After we concluded in the CBL case that the optimal scenario for the distribution of category data depended on the optimal solution for article master data, the choice for the flow of article master data was a logical continuation. Furthermore, as is described in the case setting, there were several stakeholders who were interested in doing an in-depth study to determine the costs of several important functions for both extreme situations, which was the objective of our case.

Case design
Our main objective is to *test* if the parameters we found in the CBL case are indeed the relevant parameters that determine the structure of the abstract problem description. Since designing such a situation has not been done before, we cannot study similar situations from which to draw conclusions. Therefore, we chose to examine the actual distribution situation as closely as possible, which meant at least on individual chain level. Only when we look that close we can determine which functions are performed in the actual distribution process, and which functions will result from implementing either of the two extreme scenarios. From these functions we can then infer which parameters determine the design process, after which we can compare them with the parameters from the CBL case. Following this strategy means that we will analyze the case on individual chain level.

We then decided to study not one but two mini-chains. We had two reasons for doing this. First, through comparing the data alignment processes of both mini-chains, averaging the results and eliminating extreme outcomes, we would improve the internal validity of the case. Second, the participants demanded anonymity, which we could provide through hiding the results of the individual chains. We selected four participants (two retailers and two suppliers) for the two mini-chains, namely: Douwe Egberts Nederland B.V., CPC Benelux B.V., Trade Service Nederland B.V. (part of Schuitema) and Koopconsult (part of Dirk van den Broek Bedrijven B.V.). All four companies belong to the top of the Dutch food sector and are highly regarded in the sector, especially when it comes to EDI.

Finally, apart from testing the right parameters, we also wanted to determine which distribution scenario is best. Since we want to answer this question for the complete food sector, we will extrapolate the results of these mini-chains to sector level, to determine which distribution scenario on *sector level* will be preferable. Hence, we have two units of analysis, both mini-chain level and sector level.

Criteria for interpreting findings
For testing the parameters in the abstract problem description, we will require that at least the same (or similar) parameters are found in this case.

8.5.3 Extension of the case objective with qualitative evaluation

During the case, we learned that determining the contents of the scenarios was extremely difficult, mostly because a conflict existed between those who favored the EDI article message and those who favored the EAN-DAS system. Hence, much time was lost with explaining concepts (and thus the exact contents of each scenario) and falsifying several arguments. Some committee members did not even believe that there was a problem.

To deal with these discussions, we decided to make a *qualitative* evaluation of both scenarios simultaneously, which resulted in three new research questions:

1. What is the actual data quality in the two mini-chains?
2. What exactly are the causes for insufficient data quality?
3. In what way and for how much do both extreme scenarios solve the insufficient data quality problem?

8.5.4 Case methodology

Data was collected following the steps defined in the DAL method second version. We also added some extra steps to perform the qualitative evaluation. We will shortly discuss the way that we collected the data step by step.

TRANSLATION PROBLEM

Step 1: Definition of the IP
For defining the precise objective of the article master data, we interviewed the IT managers of the 4 companies in the mini-chain. The content was based on the Message Implementation Guideline (MIG) for article master data in the Dutch grocery sector.

QUANTITATIVE EVALUATION DISTRIBUTION PROBLEM

Step 2: Qualitative description of the current and desired situation
Most data was collected from structured interviews with the IT managers and the account or purchasing managers of the four companies in the mini chains. In the interviews we collected data on the following subjects:

- The objective of the article master data (part of step 1)
- The structure of the current data alignment process and the tasks that are performed (step 2)
- The different checks that are performed in the checking process (step 2)
- The characteristics of the distribution situation (step 3)

We measured the time spent on data alignment and by reason of that the cost of data alignment to assess the benefits of using both electronic forms of distribution as compared to the current paper based situation. Since these measurements are not relevant in comparing both extreme scenarios, we will not discuss them further in this case description. For an overview of the data alignment costs and the quantitative benefits of electronic communication, we refer to Vermeer (1998:2, pp. 24).

The desired situation was established together with the four participants by developing an imaginary project plan to implement each of the electronic distribution situations in detail. We used first the EDI-in-10-steps plan of EAN Nederland to determine which activities are necessary to implement an EDI article message. This plan is useful for both scenarios since the decentralized scenario is based on the use of the EDI article message between suppliers and retailers. We then consulted the IT manager of one supplier, who has an extensive experience with EDI and who worked in many EDI message development groups for EAN Nederland over the past 10 years. At the moment the case study was conducted, the IT manager participated in both the pilot project for centralized article data distribution via a central database and another pilot project for decentralized article data distribution using the EDI article message. Using these two sources we developed a plan of action for the implementation of both scenarios.

Step 3: Abstract definition of the design problem
We identified *input and output parameters* through asking all interviewees what the characteristics of the distribution situation were. The *design parameters* were identified through reviewing the steps in the project plans for both scenarios. The *object function* was also determined from the project plans of both scenarios. Specifically, we focused on determining all costs that were involved in realizing both scenarios, after which we reviewed these costs to determine the precise cost types. These cost types determine the object function of the distribution situation. Unfortunately, we were not able to include data quality, since the EAN-DAS system was not operational at the time of this study. Therefore, we could not determine how the resulting data quality changed using, for instance, central checking, For the case, we worked around this problem through determining the extra cost of implementing data quality checks locally in the decentralized scenario, assuming that those checks can deliver the same quality as in the case of central quality checking. Finally, because of all the discussion about the qualitative objectives of the two scenarios, we left the *constraints* out.

Step 4: Scenario formulation
As described in the case objective, we developed two extreme scenarios. Therefore, scenario formulation was simple. During scenario formulation we also worked on a better way to represent scenarios. To do this, we used a graphical notation that we adopted from physical distribution (Bertrand 1990). Finally, the quantitative evaluation was performed by comparing the total costs of both scenarios.

Step 5: Quantitative evaluation
In the quantitative evaluation, we quantified the costs and benefits of the two electronic distribution scenarios and compared them to each other. Specifically, we together with the four participants determined the costs of implementing the decentralized and the centralized distribution scenario, by estimating the costs of each step in the imaginary project plan.

QUALITATIVE EVALUATION DISTRIBUTION PROBLEM

In the qualitative evaluation we followed research question 5-7. This leads to the following three additional steps

Step 6: Measuring current data quality
We first measured the current data quality through matching the article databases of suppliers and retailers in the two mini-chains. Here we followed the measurement method that we developed in the two data quality cases.

Step 7+8: Finding out the causes of insufficient data quality & determining if and how both scenarios solve the insufficient data quality problem
In this part of the qualitative evaluation we wanted to relate the arguments of all stakeholders to each other to determine the causes of insufficient data quality and which scenario is best. To do this we used Goldratt's Critical Chain analysis methods (Goldratt 1997). These methods are based on the creation of cause-effect trees, where the underlying assumptions between each cause and effect are made explicit. To determine the root causes of insufficient data quality, we constructed several Current Reality Trees (CRTs). To determine if and how the two extreme scenarios solve these root causes, we constructed several Future Reality Trees (FRTs). The Undesired Effects (UDEs) that are the input for constructing the CRTs, were collected as part

of the structured interview. In these interviews together with the four participants we asked the interviewee what their perception of the problem was by describing the undesired effects. The arguments that are the input for constructing the FRTs were collected by asking what the advantages and disadvantages of both extreme scenarios were. This resulted in a list of four root causes and a number of qualitative arguments to explain the usefulness of both extreme scenarios, grounded in the stakeholders' views of the problem and their solutions.

8.5.5 Case results

Step 1: Defining the Information Product

Step 1.1: Objective
The article master data that is distributed in the *data alignment process* is used in the logistical replenishment process. Specifically, it is used to ensure that different automated applications such as the ordering system, the goods handling system, the automated stock allocation system, the cross docking system, the supermarket inventory system and the invoicing system, use the right article data to prevent errors in the logistical process. We specifically want to separate this objective from several purchasing and sales objectives, where article information is required to select products and/or to change the assortment. In that case, this information would require information about prices and conditions. Since the purchaser decides about new articles in the assortment, this type of information should be human oriented, containing images of the article, and compelling arguments to buy that article. Instead, the data alignment purpose of article master data requires that the information is computer oriented, which means that it should be in computer readable form. Based on this description, we define the objective of the article master data as follows:

> *To ensure that automated applications in the interorganizational supply chain use the same article data to prevent errors in the logistical replenishment process.*

Step 1.2: Contents
For defining the contents of the IP we used the MIG for the Dutch grocery sector. Here, the agreements about the structure and the meaning of the article data for the grocery sector, specifically to support the logistical replenishment process, are described. The data model of the IP is shown in Figure 8-4.

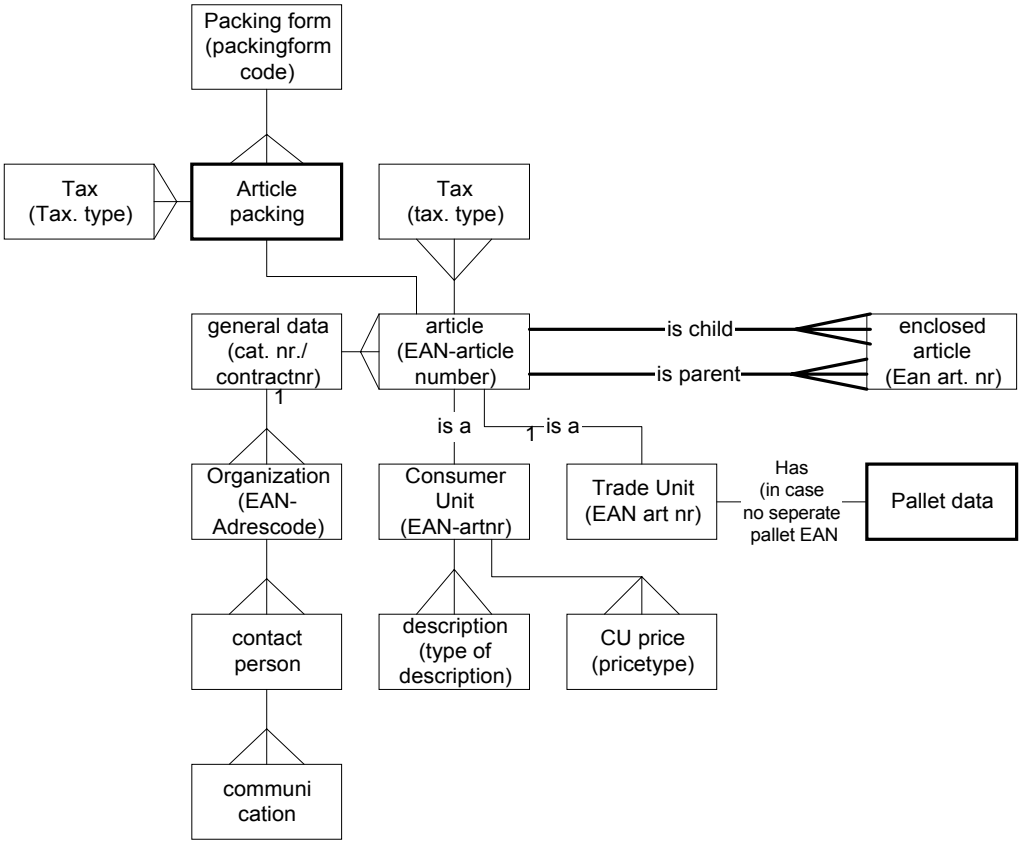


Figure 8-4: The Information Product for the article master data

The IP is no static description but continues to evolve. For instance, during the EAN-DAS project the participating retailers and suppliers were working on a new version of the MIG.

Step 2: Qualitative description of the current and desired situation

Step 2.1 the current situation

Figure 8-5 shows the current data alignment procedure that we found in the two mini chains.

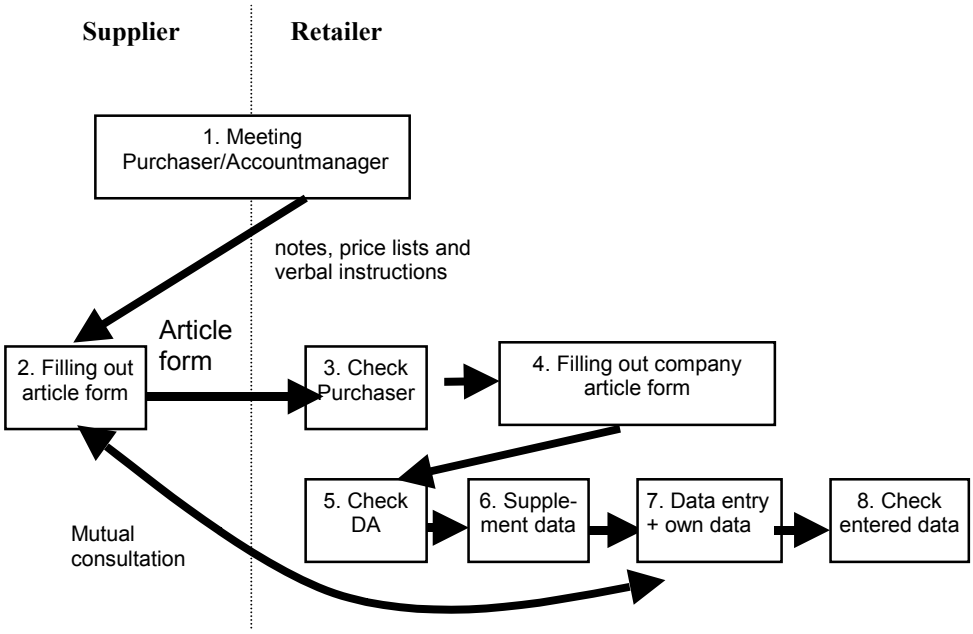


Figure 8-5: The data alignment procedure

The procedure starts with the regular meeting between the account manager of the supplier and the purchaser of the retailer (1). In this meeting they discuss new product introductions and promotions. Upon his return, the account manager orders his assistant to fill out the Standard Article Form of the retailer and send it back (2). If no Standard Article Form is available, the account manager normally will supply a brochure instead, which normally contains extra logistical product information. The purchaser receives the Standard Article Form and checks prices and conditions (3). Then he passes the document to the assistant purchaser. The assistant fills out the Standard Purchasing Form (4), which largely resembles the Standard Article Form. Next, the assistant purchaser sends this form to the Data Administration (DA) department. The DA department checks (5) the data on the Standard Purchasing Form. Normally several required data fields are missing, which means that the data has to be supplemented. The DA department (or the assistant purchaser) contacts the supplier to retrieve the missing information (6). Next, the data administrators enter the data in the article registration system (7). Normally, more than half of the entered information is company specific (private article codes, classification codes etc.). Finally, the next day the entered data is checked for data entry errors (8).

Next to describing the current data alignment procedure, we also investigated the current checking process in more detail. Checking the data plays an important part in the data alignment procedure, since both the purchasing assistant and the DA department perform this activity. Table 1 gives an overview of the checking activities currently performed.

Checking activity	Now?
1. Timeliness	
a. Number of updates that are processed within a certain time limit	No
2. Completeness	
b. Record completeness	No
c. Information Field completeness	Yes
3. Accuracy	
d. Physical samples (e.g. comparing physical products with their data)	Sometimes
e. Checking of meeting minutes	Sometimes
f. Comparison of two article data bases	No
4. Consistency	
g. Checking with logical rules	Yes
5. Interpretability	
h. Conformance to the IP (or Article Implementation Convention)	No

Table 8-2: Current checking activities

Table 8-2 shows that from the eight possible checks we identified, only two are actively performed, namely, checking information field completeness and checking with logical rules. However, the very important interpretability check is not performed.

Step 2.2 the desired situation
For describing the project plans of the two scenarios, we assumed five things:

- Since all four participants in the two mini chains already used EDI, we did not include the steps for implementing EDI message exchange.
- Since all four participants are interested in fully automated data alignment, we excluded the use of the EAN-DAS working station for the centralized scenario. This working station is a software program that was developed within the EAN-DAS project to include suppliers and retailers who are not EDI-enabled. Also for the decentralized scenario we assume that an automated connection between the applications and the EDI-gate is built.
- We further assume that the checks on information field completeness (2.c), consistency (4.g) and conformance to IP (5.h) are fully automated. Consequently, these are implemented in the checking applications of the EAN-DAS system or the retailer. However, we suppose that the standard checks on prices and conditions are still handled manually.
- Next, we assume that the three automated checks are present in both distribution scenarios. In the centralized scenario the EAN-DAS system performs this task. In the decentralized scenario both supplier and retailer have implemented checking software that performs this task.
- Finally, we assume that the resulting data quality in both scenarios will be the same, since the three checks are implemented both centrally and decentrally. Thus, in the comparison we do not include the impact of both scenarios on data quality.

Centralized distribution
An overview of the activities for the centralized scenario is shown in Table 8-3.

Step	Supplier/ retailer	Activity
Preparation Phase		
1	S	Subscribe to EAN-DAS
2	S	Collect all available information per EAN code in the corporate databases
3	S	Supplement all remaining information that is not available
4	S	Change account management application and develop a connection with the EDI gateway
5	S	Implementation of three checks (2.c, 4.g and 5.h) in own system
6	S	Send all article data records to central database
7	S	Test message exchange with central database
8	R	Subscribe to EAN-DAS
9	R	Create assortment definition
10	R	Create connection from EDI gateway to Purchasing application and Data Adm. Application
11	R	Match current corporate article database with data from the central database and correct it
12	R	Test message exchange with central database
Exchange Phase		
13	S	Preparation of update in account management application
14	R	As result from meeting: change assortment definition
15	S	Push a button to send article information to EAN-DAS
16	Central	Checking of articles (checks 2.c, 4.g and 5.h) and distribution to retailer
17	R	Send article updates from EDI gate to Purchasing application and Data Adm. application
18	R	Purchasing and Data Administration check the received data (checks 1.a, 2.b, 3.d, e, f)
19	R	Data entry of extra data (private classification codes for instance)
20	R	Check data entry

Table 8-3: Activities in the centralized scenario

The table shows a distinction between preparation activities and actual exchange activities. As we can see, many activities are related to software development and testing.

Decentralized distribution
An overview of the activities for the decentralized scenario is shown in Table 8-4 on page 144.

In this table, in the Preparation phase, we also made a distinction between one-time activities and supplier-retailer relation activities. The supplier-retailer relation activities have to be performed for each implementation between a supplier and a retailer.

	Supplier/ retailer	Description
Preparation Phase		
One time activities		
1	S	Development of address maintenance module
2	S	Development of assortment maintenance module
3	S	Collect all available information per EAN code in the corporate databases
4	S	Supplement all remaining information that is not available
5	S	Change account management application and develop connection with EDI gateway
6	R	Create connection from EDI gateway to the Purchasing application and Data Adm. appl.
7	R	Implementation of three checks (2.c, 4.g and 5.h) in own system
Activities per Supplier/Retailer relation		
8	R	Match the corporate article database with the database from the supplier and correct it
9	S/R	Define the article message implementation agreements
10	S/R	Define exchange procedures and directions for use
11	R	Enter the data in the assortment module
12	S/R	Test message exchange between supplier and retailer
Exchange Phase		
13	S	Preparation of update in account management application
14	S	Push a button to send article information to retailer
15	R	Send article updates from EDI gate to Purchasing and Data Administration application
16	R	Purch. and Data Adm.check the received data (checks 1.a, 2.b,c 3.d, e, f, 4.g and 5.h)
17	R	Data entry of extra data (private classification codes for instance)
18	R	Check data entry

Table 8-4: Activities in the decentralized scenario

Step 3: Abstract description of the design problem

Step 3.1 Input and output parameters
The characteristics of the distribution situation play an important role in the evaluation of the different distribution scenarios. We identified the following characteristics for the two mini-chains. On average there are 750 suppliers and 25 retailers per mini chain. The average assortment of the retailers contained approximately 12.400 articles. The retailer receives weekly on average 180 new article updates and 325 existing article updates. The number of assignment errors or checking errors is not known. However, the retailers demand that at least 95% of their article data records do not contain any errors.

Step 3.2 Design parameters
Based on the detailed project plans of both scenarios, we identified the functions determining the structure of the design situation, with the following values for each parameter (between brackets).

1. Administration function (centrally). The Information Product, which is described in the MIG, continually evolves, which causes changes to the IP. Similar to the CBL case, this results in an administrative task, which can only be reasonably performed at a central organization.
2. Assignment function (centrally or supplier). This function *enters* article information into the article database (of new and existing products), according to the structure described in the IP. This function can be performed at a central organization or at the supplier.

3. Checking function (centrally or retailer). Several (although not all) checking activities can be performed both centrally and locally. Especially, since we have defined a central IP description, we have to check the fields that are specified in the IP (check 2.c) and the conformance to the IP specification (check 5.h). Also various consistency checks can be implemented (check 4.g). This function may be performed at a central organization or at the retailer.
4. Implementation agreements and message exchange procedures administration (centrally or per supplier-retailer relation). As we saw in the scenario descriptions, we do not only have to make agreements about the structure of the article data (the IP), but also about exchange procedures and message usage. These agreements can be made and administered centrally for all participants or they can be made locally for each supplier-retailer relation.
5. Message testing (centrally or per supplier-retailer relation). Message testing is a necessary activity when electronic messages are implemented. This activity can be performed centrally (which means that all suppliers and retailers test the message exchange with the central organization) or locally for each supplier-retailer relation.
6. Profile maintenance function (centrally, supplier or retailer). When the article data is ready to be sent to the retailer, it still has to be customized to the retailer's specifications. This means that from the complete IP specification, the retailers can specify which data fields they are interested in (Most retailers only want a subset of all the available information). Furthermore, the retailer can indicate on which articles he wants to receive information. Thus, a function is needed that customizes the update to the exact retailer's specifications. In the centralized scenario, this function is performed in the assortment definition step. In the decentralized scenario it is accomplished in the assortment module. The function can be performed at a central location, at the supplier's or at the retailer's.

Step 3.3 Objective
Reviewing the description of the current situation, and the project plans of the two extreme scenarios, we found the following costs that are involved:

- Administration costs covering processing updates of the MIG, making implementation agreements and developing message exchange agreements.
- Assignment costs for entering the article information in the database (at the supplier).
- Distribution costs pertaining to sending messages bilaterally or via the EAN-DAS system.
- Checking costs (manual and automated at retailer, or automated at EAN-DAS system).
- Software development, testing and maintenance costs related to implementing either the bilateral EDI article message, or the EAN-DAS system at both supplier and retailer. These costs can be split up in:
    - The cost of software checking application
    - The costs of profile maintenance application
- Message testing costs.

Step 4: Scenarios
Using the six design parameters we described, we could generate 1*2*2*2*2*3 = 49 different scenarios. For this case study, we chose to investigate only the two extreme distribution scenarios. In addition, we will describe both extreme scenarios using a representation notation from the world of physical distribution.

Description method
To represent the different alternatives, we developed a simple representation method, which uses the following three symbols:

- The *rectangle* represents the location of each of the different functions. All functions are located at no more than three locations: centrally, at the supplier or at the retailer. Functions that are performed per supplier-retailer relation are located at both the supplier and the retailer.
- The *triangle* represents the location of the profile maintenance function, which we refer to as the Information Decoupling Point (IDP). This IDP is similar to the CODP, which we described in 7.2.2. Here, the IDP is defined as the point in the supply chain from which the article information is customized to the specific users' needs.
- The *square* represents the different parties in the chain: the supplier, the retailer and the central organization.
- The *arrow* represents the flow of information.

Abstract description centralized and decentralized distribution scenario

Figure 8-5 represents both scenarios graphically.

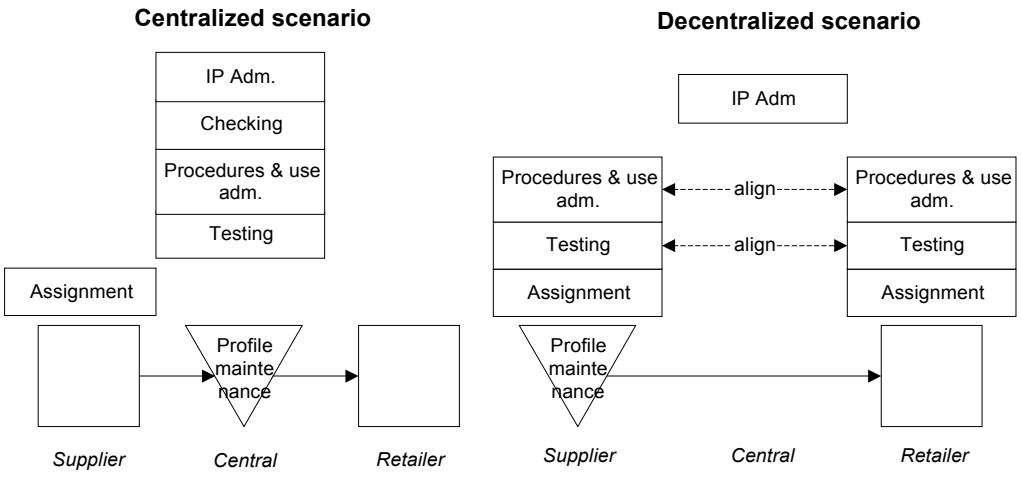


Figure 8-6: The centralized distribution scenario

The centralized distribution scenario works as follows. When an update of a new or existing article occurs, the supplier sends the article data according to the specifications of the IP to EAN-DAS (the central database system). There, the data is checked (check 2.c, 4.g and 5.h). If an error occurs, the data is returned to the supplier with the request to correct the problem. Thus, the main task of the checking function is to guarantee that all communication about article master data is according to the specifications of the IP. Next to checking the data, also the message syntax is checked (conformance to the EDI article message specification). Thus message testing is performed continuously. Finally, the data is customized to the retailer's specifications using the assortment definition of the retailer, after which the customized data is

distributed to the retailer. Since profile maintenance is performed at the central database, the IDP is located centrally in the chain. As we can see from Figure 8-6, all functions are performed centrally, except for the assignment function. This we decided because a central assignment scenario would mean that each new or altered *physical* product should be send to the central organization. There, the administrators should determine the characteristics of the products (size, measurements, packing units, trade units etc.) and enter it in the central database according to the IP specification. This would mean that over 25,000 articles annually with per article about 70-100 fields of information have to be determined at the central organization. Since all stakeholders (The CBL, EAN Nederland and SMA) in the food sector have indicated that this is not a realistic situation, we excluded this situation in this case.

The decentralized distribution scenario works as follows. Before data alignment can start, both retailer and supplier have to make individual agreements about exchange procedures and rules of conduct. Additionally, they have to exchange test messages to test the reliability of the link. After that the actual exchange can take place. When an update of a new or existing article occurs, the supplier has to send an EDI article message to all interested retailers. To do this, the supplier uses an assortment module where all assortment definitions of all retailers are stored (who want which article data fields). Thus, the IDP is located at the supplier. When the retailer receives the EDI message, he checks the article data (all checks). If errors occur, the supplier is informed. Finally, the update is stored in the retailer's database.

Step 5: Quantitative evaluation

In the quantitative evaluation we have compared the costs of each distribution scenario and determined which scenario is best. The analysis is based on the assumptions described in step 2.2.

Distribution costs
The results of estimating the distribution costs of both scenarios are presented in Table 8-5.

Step	Description	**Decentralized** Costs	**Centralized** Costs
C1	Subscription	600	3,000
D14/C15	Updates	1330	2.600
	Subtotal Supplier	**1,930**	**5,600**
C8	Subscription	600	20,000
C16	Updates	0	2.600
	Subtotal Retailer	**600**	**22,600**

Table 8-5: distribution costs of both scenarios

The table shows that the costs for the supplier in the decentralized scenario consist of a yearly EDI subscription cost and the cost of sending updates. The EDI subscription cost is 50 NLG per month per EDI mailbox, which means that the supplier pays 600 NLG annually. The one time installation cost is not included since all suppliers (and retailers) already use EDI. We estimated the cost of sending updates at 1330 NLG annually. This estimate is based on 200 updates annually per supplier sent to 25 retailers at a tariff of 0.266 NLG per message (see Vermeer 1998:2, pp. 25). The annual subscription cost for the retailer in the decentralized scenario is also 600 NLG. Since the retailer only receives messages, we assume that the cost of sending messages is zero (thus, the retailer does not send a conformation message).

The costs in the centralized scenario also consist of a subscription cost and the cost of sending messages. The subscription cost is 3000 NLG per supplier annually and 20,000 NLG per retailer annually. The cost of sending *and* retrieving messages are 50 NLG per session. In a session one or more article messages can be exchanged. If we assume that suppliers and retailers will send and retrieve messages once per week, this will cost both suppliers and retailers (50 NLG * 52 weeks =) 2600 NLG annually. All costs of the centralized scenario are based on the tariffs of the concept Business Plan for the EAN-DAS system.

From the comparison we see that the distribution cost of the centralized scenario are much higher than for the decentralized scenario. This was expected, since the centralized scenario does more than only distribute the data (for instance, it contains the tasks of central IP administration, central checking, and central profile maintenance).

Costs of investment
The costs of investment occur in the preparation phase of both scenarios (step 2-12, ex 8 in the centralized scenario and step 1-12 in the decentralized scenario). As we saw in the description of both scenarios, we distinguish between the one-time investments for adapting the IT systems for electronic communication of article data, and the investments per supplier-retailer relation that occur each time a new connection is established. For each activity in the detailed scenario descriptions we estimated the number of days required for software development[1], which we multiplied with a standard tariff of 1,500 NLG per day. This is an average of internal and external development tariffs.

An overview of the costs of one-time investments for both scenarios is shown in Table 8-6.

[1] These estimated were established through consulting the EDI expert of one of the suppliers who already had extensive experience with EAN-DAS and EDI article message exchange pilots.

Step[1]	Description	Decentralized Days	Decentralized Costs	Centralized Days	Centralized Costs
D1	Development of address maintenance module	1			
D2	Development of assortment maintenance module	40			
D3/C2	Collect all available information per EAN code	40		40	
D4/C3	Supplement all remaining information that is not available	20		20	
D5/C4	Develop connection to EDI gateway	15		15	
D7/C5	Implementation of three checks in own system	25		25	
	Subtotal Supplier	**141**	**211,500**	**100**	**150,000**
C9	Create assortment definition			5	
D6/C10	Create connection from EDI gateway applications	15		15	
D7	Implementation of three checks in own system	25			
	Subtotal Retailer	**40**	**60,000**	**20**	**30,000**

Table 8-6: One-time investments

Table 8-6 shows a difference of 61,500 NLG for the supplier in favor of the centralized scenario, which is almost completely the result of the extra cost of developing an assortment module in the decentralized scenario. For the retailer we find a difference of 30,000 NLG, which results from the extra costs of implementing the three checking activities (2.c, 4.g and 5.h) in the retailer's IT systems.

Table 8-7 gives an overview of the investments per supplier-retailer relation. As we can see from the table, we chose to include several activities from the centralized scenario to make a better comparison possible between the two scenarios. These specific activities are dependent on the number of supplier-retailer relations for the decentralized scenario, but also occur as one-time investments in the centralized scenario.

Step	Description	Decentralized Days	Decentralized Costs	Centralized Days	Centralized Costs
D9,10	Define the article message implementation agreements, exchange procedures and directions for use (3days)	75			
D11	Enter data in assortment module (2 days)	50			
D12/C7	Test messages (3 days decentralized, 10 days centralized)	75		10	
	Subtotal supplier (with 25 retailers)	**200**	**300,000**	**10**	**15,000**
D8/C6,11	Match corporate article databases (12400 articles; 5 min/art)	129		129	
D9,10	Define the article message implementation agreements, exchange procedures and directions for use (3days)	450			
D12/C12	Test messages (3 days decentralized, 10 days centralized)	450		10	
	Subtotal retailer (with 20% of 750 suppliers)	**1029**	**1,543,750**	**139**	**208,750**

Table 8-7: Investments per supplier-retailer relation

Table 8-7 shows that especially the costs of implementing the local agreements about article message implementation, exchange procedures and rules of conduct (d9/d10) and the cost of sending test messages (d12) require large investments for both suppliers and retailers in the

[1] The numbers in this column refer to the steps in the centralized and decentralized distribution scenario (see Table 8-3 and Table 8-4). D1 = Decentralized scenario step 1. C7 = Centralized scenario step 7.

decentralized scenario. This results in a difference of 285,000 NLG for the supplier and 1,3 million NLG for the retailer[1] in favor of the centralized scenario.

Finally, Table 8-8 gives a summary of all costs for both scenarios. The table shows a distinction between investment costs and yearly recurrent costs. In the yearly recurrent costs we added an extra entry for software maintenance, which we valued at 10% of development costs, since the investments in software play such an important role.

		Supplier			**Retailer**		
		Decent-ralized (a)	Centra-lized (b)	Difference (%) (1-b/a)	Decen-tralized (a)	Centra-lized (b)	Differen ce (%) (1-b/a)
Investment	One time investment	211,500	150,000		60,000	30,000	
	Investment per relation	300,000	15,000		1,543,750	208,750	
	Total investment	**511,500**	**165,000**	**68%**	**1,603,750**	**238,750**	**85%**
Yearly costs	SW maintenance cost	51,150	16,500		160,375	23,875	
	Distribution cost	1,930	5,600		600	22,600	
	Total yearly costs	**53,080**	**22,100**	**58%**	**160,975**	**46,475**	**71%**

Table 8-8: Total costs for both scenarios

Table 8-8 shows that the suppliers of our two mini chains have 68% less investment costs in the centralized scenario and 58% less yearly recurrent costs. The retailers have 85% less investment costs and 71% less yearly costs.

Furthermore, the table shows that the distribution cost is less than 1% in the decentralized scenario for both supplier and retailer, and approximately 10% less for the centralized scenario. However, the distribution cost of the centralized scenario contains much more than only the *transportation* cost of the data (e.g. also the administration, checking and profile maintenance function). If we assume that the distribution cost of the decentralized scenario is a good estimation for the transportation cost, than we may conclude that the transportation cost is only a fraction of the total costs.

Step 6: Current data quality
To measure the current data quality we compared the databases of supplier and retailer in the two mini chains. We performed three tests to measure the current data quality in the two mini chains: (1) Information field completeness, (2) Record completeness, and (3) Accuracy of the data fields (see Vermeer 1998:2, pp. 5-7). This assessment resulted in the following two conclusions:

- The current data quality in both mini chains is very low. Each article update contains at least one error. Since all four companies are highly regarded in the Dutch food sector -all having experience with automated processes and EDI communication for a long period of time- this result is considered to be an optimistic estimate of the data quality in the rest of the sector. Thus, it is expected that the data quality in the rest of the sector is even worse.

[1] In the calculations we assumed that the retailer implements EDI article messages with only 20% of the 750 suppliers.

- The pallet data of the articles have the worst data quality. This is consistent with feelings in the sector that pallet data are seldom correct.

Step 7: Determining the root causes of insufficient data quality
As described in the case methodology, we determined the root causes through identifying all UDEs from which we constructed a CRT. For detailed information, we refer to the case study report (Vermeer 1998:2, pp. 15 -21). The CRT is largely comparable to the CRT in Appendix B that was constructed for the problem analysis in Chapter 3.

The following four root causes were identified in the two mini-chains:

1. There are no central agreements about the structure of article data, data definitions and restrictions (a central data model), which are actually used in practice.
2. Agreements between suppliers and retailers about article data are often not documented.
3. Both internal procedures (between departments for collecting data) and external procedures (for data exchange) are often not explicit.
4. The data is not sufficiently checked. At the moment only the completeness of fields and some consistency checks are performed. However, the other checks (discussed in Table 8-2) are not performed.

Step 8: How effectively do both scenarios solve the data quality problem?
Using the positive and negative arguments of all participants about both extreme scenarios, we constructed several FRTs to explain how both scenarios solved the four root causes. Here, we will describe only the most important arguments. For more details we refer to Vermeer 1998:2, pp. 15-21).

Cause 1: There are no central agreements about the structure of article data, data definitions and restrictions, which are actually used in practice.

There are two arguments that we will address:
a. Both the centralized and decentralized scenarios provide a solution based on a single data model (the IP) with a description of the structure of article data, the definitions and the constraints. The major difference is that in the decentralized scenario only the IP is centrally defined, whereas other agreements are still locally defined.
b. A central data model is not static, but is evolving. Through the centralized checking function, the centralized scenario has a strong mechanism to signal new adjustments for the central model. As well, the checking mechanism forces both suppliers and retailers to adjust their private translation and checking applications whenever a new adjustment occurs. (The translation application translates the data from the local data model to the centralized data model and back). These mechanisms do not exist in the decentralized scenario.

Cause 2: Agreements between suppliers and retailers about article data are often not documented.

Documentation is a subject that is hardly addressed both in the EAN-DAS project (the centralized scenario) and in the pilot projects, which are based on communication via EDI

article messages (the decentralized scenario). However, the availability of good documentation about the central data model will be essential for both scenarios to become successful.

Cause 3: Both internal procedures (between departments for collecting data) and external procedures (for data exchange) are often not explicit.

The active checking mechanism in the centralized scenario forces suppliers to improve their internal procedures, because otherwise they may receive a fine. In the decentralized scenario the pressure to improve the internal procedures is much less, because the effectiveness of decentralized checking depends on the agreements that are made between suppliers and retailers and to the degree that retailers provide feedback about errors to their suppliers. With regard to the external procedures, both scenarios improve them because both scenarios depend on *electronic* data exchange, which in it self requires detailed and explicit exchange procedures. However, an extra impulse in the centralized scenario is provided by the contractual agreement, which explicitly prescribes how the data should be exchanged.

Cause 4: The data is not sufficiently checked.

In the centralized scenario the interpretability check (5.h) actively examines the conformance of all article data in the sector to the central data model. This check is essential because it guarantees uniform data and thus enabling undisturbed translation to the internal format. In the decentralized scenario nothing specific is arranged for the checking function. Hence, it depends on the discipline of the retailers that this check is performed and is performed in the right way. Since dependence on discipline is a weak enforcement instrument, the final quality of the data in the decentralized scenario will be less than in the centralized scenario.

Based on these arguments, we conclude that the centralized scenario is more effective in solving the causes for insufficient data quality than the decentralized scenario.

8.5.6 Case conclusions

The practical case objective was to analyze which of the extreme case scenarios was best, compared to each other and to the paper based situation. From the quantitative evaluation we draw the following conclusions (see also Vermeer 1998:2):

1. The benefits of switching from the current paper based situation to electronic distribution of article data amount up to 0.2 FTE for the supplier and 5.2 FTE for the retailer. These are the direct benefits from savings in administrative costs. The indirect benefits as a result of less prevention and fewer errors in the primary processes are not included. (This conclusion was important for the practical case, but was not covered in this account of the case study. For details we refer to Vermeer 1998:2, pp. 30.
2. The centralized scenario is for both the suppliers (with 68% less investment costs and 58% less yearly recurrent costs) and the retailers (with 85% less investment costs and 71% less yearly recurrent costs) cheaper than the decentralized scenario.
3. Compared to the software development and maintenance costs, the real message transportation cost is only a fraction (<1%) of the total costs for each scenario.

From the qualitative evaluation we draw the following conclusions:
1. The causes of insufficient data quality are: (1) no central agreements (2) No good documentation (3) Procedures are not explicit and (4) Data is not sufficiently checked.
2. The centralized scenario appears to be solving these problems better than the decentralized scenario.

8.5.7 Reflection

The main research objective of this case study was to test the parameters in the abstract problem description we extracted from the CBL case. Additionally, we wanted to improve the method through exploring how to measure different cost types and through adding a graphical model to describe scenarios. However, during the case we learned that apart from the quantitative problem we were solving, there also existed a more qualitative problem about what exactly the problem was, and why each of the extreme scenarios solved this problem. Therefore, we decided to add a fourth objective, which was a qualitative evaluation of the problem. With respect to these four objectives we draw the following conclusions:

Testing the abstract problem description
Figure 8-7 gives an overview of all parameters that we found in this case plus the parameters from the CBL case.

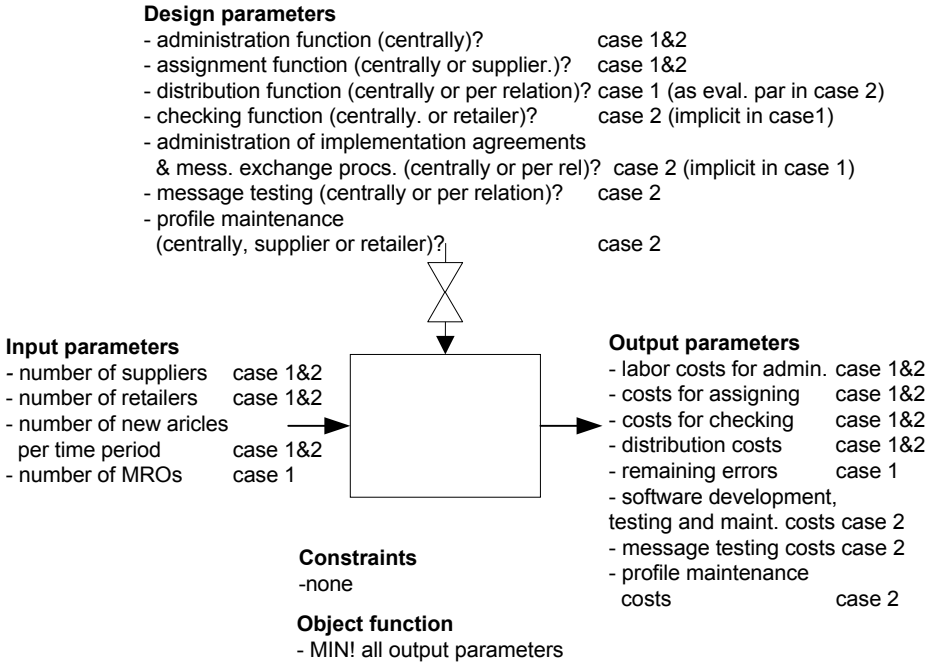


Figure 8-7: Comparison abstract problem description both cases

The comparison shows:

- With respect to the *design parameters,* that 5 of the 7 parameters are similar to the parameters from the CBL case. The administration and assignment parameters of case 1 (CBL) were also found in case 2 (EAN-DAS). The distribution parameter from case 1 was implicitly present in case 2, but then as an evaluation parameter. The reason for this was that case 2 showed that the location of the distribution function was primarily determined by the location of other functions, such as the assignment, administration or profile maintenance functions. Furthermore, the new checking function in case 2 was implicitly present in the assignment function of case 1. Finally, since the administration of implementation agreements and message exchange procedures is a further detailed parameter of the administration parameter, we can combine this into one parameter. This means that two new design parameters were found: the message testing and profile maintenance functions.
- With respect to the *input parameters,* that three of the four parameters in both cases were similar, except for the number of MROs parameter. However, this parameter should be excluded, since it depends on the specific characteristics of the CBL case.
- With respect to the *output parameters,* that 4 of the 8 parameters are similar to the parameters from the CBL case. The remaining errors parameter was not measured in the EAN-DAS case, since at this time we did not know what the effect of central checking on the quality of the data would be. The other 3 parameters (message testing cost, software development and maintenance cost, and profile maintenance cost) are new parameters.
- With respect to the *object function,* that the object function in this case differs from the object function in the previous case in that it measures only costs, not quality. The independent quality parameter was transferred to a cost function, by means of adding the extra cost of developing a quality checking application.
- With respect to *constraints,* that in both cases no constraints were identified.

Based on this comparison, we conclude that almost all parameters in the abstract problem description of the CBL case were found in the EAN-DAS case. However, defining the abstract problem description of the distribution system is not finished, because we found several new parameters in the EAN-DAS case that might improve the abstract problem description.

Measuring different cost types
An important question from the CBL case was that it was not clear how several cost types should be measured. In this case we defined explicitly how we measured several cost types such as distribution cost, software development and maintenance costs and message testing cost. Especially the 'estimating project time and costs' approach for estimating software development, maintenance and message testing costs proved to be useful.

Graphical representation method
In this case, we developed a graphical representation method through using a notation form from the physical distribution world. Although this notation appeared to be useful, there were several problems with the notation:

- In representing the structure of the distribution situation (at least) 6 different functions exist. The distribution of these functions over the network appears to be critical to represent each scenario. In the physical distribution world only the location of a Distribution Center (or Warehouse) is important.
- The flow of goods plays an important role in describing a physical distribution structure. However, in our case, the flow of information is much less important than the location of the 6 different functions. One important reason might be that this case shows that the distribution cost is much less important than other costs.

Qualitative evaluation
There are two conclusions we draw regarding the qualitative evaluation:

- The qualitative evaluation proved to be very important. As described in the setting of the case, an important reason for studying two extreme scenarios was that within the steering committee there were different opinions about which scenario was best. Although developing the project plans in the quantitative evaluation cleared up many questions considering the exact definitions of each scenario, the DAL method could not handle all the different arguments that were used in the steering committee to defend each of the scenarios. Because the qualitative evaluation showed how each of the scenarios solved the root problems of insufficient data quality, and how advantages and disadvantages of each scenario were related to these root problems, the steering committee got a clearer picture of the problem. This also resulted in more acceptance of the outcomes of the quantitative evaluation.
- Goldratt's Chain Analysis methods proved to be very helpful in analyzing different arguments in a structured way. They appeared to be very adequate in connecting many seemingly unrelated problems to each other and eliminating many lines of argument, hence resulting in a clear and much simpler picture of the problem situation.

8.6 Conclusions

The objective of the cases was to test the DAL method. We did this using an evolving case study strategy, which both tested and further explored the DAL method.

In the CBL case, we tested the first step of the method. In addition, we explored the content of the method. The case showed that with respect to the *translation* part of the method, the definition of the objective was important because there proved to be two different, and partly conflicting, objectives. With respect to the distribution part of the method, we found three parameters that defined the static structure of the distribution problem.

In the EAN-DAS case, we focused on testing the parameters that determine the structure of the distribution network. The case showed that almost all parameters, which determined the structure of the distribution in the CBL case, were also relevant in the EAN-DAS case.

Based on these results, and a comparison with the requirements from Chapter 7, we draw two conclusions:

1. The current version of the DAL method solves both the translation part of the data alignment problem (as was confirmed in the CBL case), and the static part of defining a fact update distribution structure (as was confirmed in the EAN-DAS case).
2. The current version of the DAL method does not only fulfill the first criterion of a good design method, namely that it works. It also fulfils Simon's criteria for design methods, which means that it has a problem representation, it states objective and constraints, it uses different scenarios, it has an optimization function, and it states how results are evaluated.

9. Towards a Formal Model for Data Alignment

9.1 Introduction

In this chapter, we will use the final version of the Data Alignment through Logistics (DAL) method to develop a more formal description for data alignment. In Section 9.2 we will use the parameters from the comparison of the abstract problem descriptions of both cases (Figure 8-7) to describe a formal model for data alignment. Specifically, we will precisely define the input, decision, and output parameters of the model. Next, in Section 9.3 we will construct cost functions for each of the evaluation parameters, and combine these cost functions in an object function. This results in a description of the DAL method, in terms of a Binary Integer Programming Model. Using this model, an information network designer is able to determine an optimal solution for a specific interorganizational business network.

Since there are only 72 possible scenarios, we will further analyze the scenarios to understand under which conditions different scenarios will prevail. This analysis is presented in Section 9.4. This results in a decision tree, that an information network designer may use to understand why a specific solution is best under different circumstances.

9.2 Definition of a formal model for data alignment

Based on the situations of the CBL and EAN-DAS case, we can now combine the parameters of both cases to develop a formal model for data alignment. We will do this through using the parameters from Figure 8-7, defining each parameter more clearly and combine several of them.

Input parameters
With respect to the input parameters, they are the same for both cases, except for the number of MROs. This parameter will not be included, since it depends on the specific characteristics of the CBL case.

We define the following input parameters:

- S = Number of suppliers
- R = Number of receivers
- F_{mu} = Number of updates (or changes) per time period (the mutation frequency)

Design parameters
When we look at the design parameters, two simplifications are possible. Firstly, when administration of agreements is centralized, decentralized implementation procedures and message exchange procedures will always be much more expensive, which means that these procedures should also be centralized. On the other hand, when administration of agreements is decentralized, centralized administration of other agreements and procedures is not feasible. Therefore, we will combine these two design parameters into one parameter for administration.

Secondly, when we look at the distribution parameter from the CBL case, another simplification is possible. In the CBL case, we defined the distribution function as concerned with the transport of messages via a central point, or bilaterally between each supplier and receiver. However, the EAN-DAS case showed that the distribution cost in terms of message *transportation* is just a fraction (<1%) of the total cost of the network, which implied that the location of the distribution function was primarily determined by the location of other functions. Since we assume that the outcome of the EAN-DAS case with respect to the cost of message transportation is representative in general, this means that we will eliminate the distribution parameter from the information logistics model.

We make one remark. In general, distribution means more than only transportation of messages. Specifically, one other important characteristic of information distribution, besides transportation, is the mechanism of supply and demand. However, in our model, the profile maintenance function already performs this function. In fact, because distribution means more than only message transportation, it is exactly the reason why we conducted two cases to research information distribution.

We define the following design parameters:

- g = Where administration of agreements (central, bilateral)
- h = Where assignment (central, local at supplier)
- i = Where checking (central, local at receiver, none)
- j = Where testing (central, bilateral)
- k = Where profile maintenance (central, supplier, retailer)

Each parameter (g, h, i, j, k) is defined as follows:

- Administration = The task of discussing and implementing changes in agreements, with respect to data structure agreements (which are contained in the Information Product), message exchange procedures, and message implementation and usage agreements. In our model, administration is performed centrally on the community level, or bilaterally, between each supplier – receiver relationship.
- Assignment = The task of entering article into the article database (of new and existing products) according to the structure, described in the IP. In our model, this function can be performed at a central organization and then distributed to all participants, or locally at the supplier. The assignment task is performed centrally, after which the data is send back to all participants, or locally, at the supplier.
- Checking = The task of performing manual sample checks and defining business rules for a checking application, to check for mapping errors, and assignment errors. In the model we will not check for synchronization errors because we assume that product updates are not saved up to save on transportation cost. Since the transportation cost of updates is close to zero, there is no point in saving up product updates to send them in batch. In our model, central checking means that a central organization performs the checking function; local checking means that all receivers perform the checking task individually; and, no checking means that no updates are checked for errors. We are assuming that no checking will normally not occur, unless assignment takes place centrally. In that case, the update information is in conformity with the IP by definition. Therefore, we will assume that no data entry errors occur with centralized assignment. With decentralized assignment, no checking will only occur, when the IP is so simple that the error level at the end of the distribution process is still within acceptable levels. Since it is very unlikely that no checking is feasible, we will eliminate this situation from the model.
- Testing = The task of testing messages. In our model, we assume that testing is performed centrally, which means that each sender and receiver sends test messages to a central testing certification authority, or bilaterally, which means that with each new link between a sender and receiver test messages are exchanged.
- Profile maintenance = The task of administering which information receivers would like to receive (specifically, about which articles, and which specific fields per article). In our model, we assume that profile maintenance takes plays at the sender, centrally or at the receiver. *Profile maintenance at the sender* means that each sender administers the profile maintenance information for all receivers that he supplies articles to. This is normally accomplished by each sender receiving a Standard Article Form from the receiver, specifying which articles and which fields are of interest to the receiver. The sender enters this information into a profile maintenance software module and administers the changes. Central *profile maintenance* means that a central organization administers a profile maintenance software module, where receivers themselves can enter their profile information. Finally, *profile maintenance at the receiver* means that each receiver administers his own profile information in his own profile maintenance software module.

Output parameters

Finally, when we look at the output parameters, three changes are necessary.

Firstly, the cost of software development is unclear, which means that we have to specify it further. Software development is necessary for three purposes: (1) developing or buying a

translation program that translates between the local data format and the central data format, or between two local formats. (2) Developing or buying a checking application to check for errors. (3) The cost of creating or buying a profile maintenance software package.

Secondly, as we saw in the EAN-DAS case, it is important to make a distinction between one time project costs and operational costs. This has consequences for the translation, checking and profile maintenance costs.

The one time project cost of *translation* is the cost of building or buying a translation program that maps between the local and central data format, or the local formats between sender and receiver respectively. The operational cost is the cost of implementing new mappings in the translation program, as a result of local data model changes, or central (or bilateral) IP model changes.

The one-time project cost of *checking* is the cost of developing, or buying a checking application. The on-going checking cost is the labor cost of checking staff, who define business rules to further develop the checking application, and who evaluate errors that are reported by the checking application.

The one time cost of *profile maintenance* is the cost of creating or buying a profile maintenance package. The ongoing cost of profile maintenance is the administration cost of first time entering, and maintaining changes in the retailer's assortment specification.

Thirdly, the question is what the impact of poor data quality will be. Reviewing the two DAL cases, we identified three sources of quality errors:

3. Assignment errors. These errors are the result of errors in assigning values to attribute fields when entering the data in the article database (i.e. data entry errors).
4. Translation or mapping errors. These errors are the result of wrong mappings between local data models and the IP (which is either a central or a bilateral data models). These mapping errors result from either human errors in making mappings, or changes in local or central data models, which did not result in required changes in mappings.
5. Synchronization errors. Synchronization errors are the result of the sender who sends updates too late. This is normally a human error, because sending updates normally requires human intervention. Two examples are: (1) a sales or account manager that gives permission to process new updates too late (this permission is still required because sending updates is part of the process of marketing communication); (2) if sender and receiver made an error in their exchange procedure agreements, a synchronization error will occur.

Another question is how we will handle errors in our model of data alignment. In this formal definition of data alignment we will follow the suggestion of Mourits & Evers (1995) who suggested to reduce the complexity of a logistics model through separating structural aspects from dynamic aspects. Therefore, we will *assume* that the testing and checking functions for all possible scenarios are sufficient in providing the required level of quality to all information receivers. The dynamic aspects of balancing testing and checking activities to control the quality levels for each retailer under different rates of local and central data model changes through time, is a separate problem that we should address after the structure of the network is

accomplished. This means that in our model for data alignment we will focus on the static, structural aspects of an information distribution network, which means that the dynamic aspects of controlling data quality levels are not incorporated in our formal model for data alignment.

We define the following output parameters:

One time project costs

Building/buying translation program cost	: $C_{transappl}$
Building checking application cost	: $C_{checkappl}$
Testing cost	: C_{test}
Building profile maintenance program cost	: $C_{profappl}$

Operational costs

Administration cost	: C_{adm}
Assignment cost	: C_{ass}
Online checking cost	: C_{check}
Profile maintenance cost	: C_{prof}

Throughout this chapter, we will use C (upper case) as the term for total costs, while c (lower case) is used for costs per unit.

These parameters are combined in an object function that will determine the optimal fit of the design parameters. This results in the following model for data alignment (see Figure 9-1).

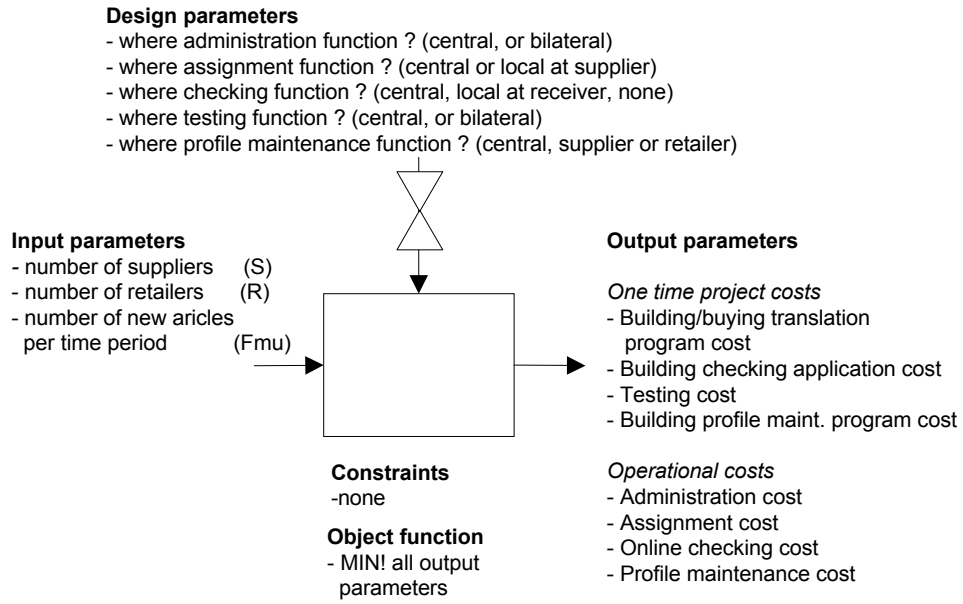


Figure 9-1: Parameters formal model for data alignment

9.3 The object function

The object function describes how each of the decision parameters affects the evaluation parameters. Therefore, we will describe which decision parameters relate to which evaluation parameters and develop a cost function per evaluation parameter. To do this, we will use a specific notation to describe the relation between different design parameters.

We define (g, h, i, j, k, m) as a scenario, consisting of six design parameters, where:

g	=	Where administration of agreements (central, bilateral)	(g=1,0)
h	=	Where assignment (central, local at supplier)	(h=1,0)
i	=	Where checking (central, local at receiver, none)	(i=2,1,0)
j	=	Where testing (central, bilateral)	(j=1,0)
k	=	Where profile maintenance (central, supplier, retailer)	(k=2,1,0)

We will use (g, h, i, j, k, m) as the notation that describes possible scenarios for each cost function.

9.3.1 Developing cost functions per evaluation parameter

Translation program cost
The translation program cost ($C_{transappl}$) depends on only one parameter: where administration takes place. Central administration means that only one IP will be used, while bilateral administration means that each sender and receiver who are connected need his or her own translation program. Thus, in case of centralized administration, each sender and each receiver needs one translation program, which means that we need (S+R) translation programs. In case of bilateral administration, the number of translation programs per sender is equal to twice the number of bilateral links between senders and receivers. If we define anr_s as the average number of receivers per supplier, then the required number of translators is $2 \cdot S \cdot anr_s$ for bilateral administration. And, if we define $c_{transappl}$ as the cost of developing one translation program, then this leads to the following cost function:

$C_{transappl}$ =

g (=adm)	Cost term
1	$(R+S) \cdot c_{transappl}$
0	$2 \cdot S \cdot anr_s \cdot c_{transappl}$

The table shows that $C_{trans} = (R+S) \cdot c_{transappl}$ if $g=1$ (central administration), and that $C_{trans} = 2 \cdot S \cdot anr_s \cdot c_{transappl}$ if $g=o$.

Specifically, we make the following assumptions:
a. Each sender and receiver requires a translation program, also in the case of central assignment with central administration. In this specific case, one could argue that a translation program at the sender might not be necessary, because assignment does not depend on the sender. However, we assume that the sender will always perform the

assignment function and only then sends the physical product to the central organization. Since the sender will receive the assignment result back to the sender, this means that the sender will compare the result with his or her own assignments. Therefore, both sender and receivers will use a translation program.

b. Also, from the situation in assumption a, it follows that each sender and receiver has a link with central, which needs to be tested, because by definition, testing is necessary whenever a translation program and exchange procedures exist for a link.

Cost of building checking application

We make the following assumption:

c. Both senders and receivers already have a checking application, because normally checking functionality is required also for message testing and for checking transaction messages.

Thus, we are only interested in the extra cost of building a centralized checking application.

The cost of building a checking application depends on one parameter: where checking (i). In case of centralized checking, we need one checking application. In case of decentralized checking, we do not need a checking application (because the receivers already have one). Finally, in case of no checking, the costs are infinite, because we make the following assumptions:

d. Checking is always necessary, because for one we separated static aspects from dynamic aspects. Although we will not address dynamic aspects to define the desired quality level, because of this separation we must assume in our model that a form of quality control is available. And then, checking is necessary since this is normal practice.

e. Checking is not necessary in case of central administration with central assignment, because the data is uniformly assigned, which results in a quality of 100%.

This leads to the following cost function:

$C_{checkappl}$ =

g (= adm)	h (=ass)	I (=check)	Cost term
1	1	0	0
0∨1	0∨1	2	$1 \cdot c_{checkappl}$
0∨1	0∨1	1	0
0∨1	0∨1	0	$\propto$ (not allowed)

In this table, we defined $c_{checkappl}$ as the cost of building or buying one checking application. For developing the cost function, we assume two more things:

f. Adapting an existing checking application costs as much as building a new one.

g. The cost of building or buying a checking application does not vary with the number of checks that need to be performed.

Cost of message testing

With testing, sender and receiver check if their message exchange succeeds, which means that they use their IP and procedural agreements to check their message exchange. Thus, testing is basically a first time check. Since testing depends on the IP, two parameters determine the cost of message testing: where administration (g), and where testing (j). In case of central

administration, with central testing (g=1, j=1), we need to test with both suppliers and receivers. If we define $c_{testing}$ as the cost of testing one connection, then the cost of testing $=(S+R) \cdot c_{testing}$. In case of central administration with decentralized testing (g=1, j=0), the same test is performed for each link. This means that the total costs are $(S^* anr_s * c_{testing})$. In case of decentralized administration, there is no difference between centralized or decentralized testing (g=0, j=0∨1), because each link has a different IP. Thus, each bilateral link must be tested, whether it is done centrally, or decentralized. Thus, the costs for both $= (S^* anr_s * c_{testing})$

$C_{testing}$ =

g=(adm)	j (=test)	Cost term
1	1	$(S+R) \cdot c_{testing}$
1	0	$S.anr_s \cdot c_{testing}$
0	1	$S.anr_s \cdot c_{testing}$
0	0	$S.anr_s \cdot c_{testing}$

There is no influence of the assignment parameter, (as is the case in the checking cost) because the concept of testing is checking the translation program and the procedures (e.g. checking the quality of the links). In case of central assignment with central administration, it is still necessary that both links with sender and receiver are checked, because both sender and receiver will use translation programs (see also assumption a and b).

Cost of building profile maintenance program

The cost of building or buying a profile maintenance program depends on the profile maintenance parameter (k). Central profile maintenance means that one profile maintenance program is needed at the central organization. Profile maintenance at the sender means that S profile maintenance programs are necessary. Finally, profile maintenance at the receiver means that R profile maintenance programs are necessary. This leads to the following cost function.

$C_{profappl}$ =

k (=prof)	Cost term
2	$1 \cdot c_{profappl}$
1	$S \cdot c_{profappl}$
0	$R \cdot c_{profappl}$

In this table, we define $c_{profappl}$ as the cost of building or buying one profile maintenance application.

Cost of administration

The cost of administration depends on one parameter: the administration parameter (g). In case of centralized administration, we need to administer only one IP. When administration is bilateral, we need to administer as many IPs as there are links between suppliers and retailers. Thus, the costs are equal to $S \cdot anr_s \cdot at \cdot c_{adm}$, where c_{adm} is defined as the cost of administering IP per time unit (e.g. a year). This leads to the following cost function for this evaluation parameter.

C_{adm} =

g(=adm)	Cost term
1	$1 \cdot c_{adm}$
0	$S.anr_s \cdot c_{adm}$

In this cost function we make an important assumption, namely:

h. The complexity of a centrally defined IP is equal to the complexity of a bilaterally defined IP.

This assumption is realistic if there is a large degree of commonality between bilaterally defined IP's.

Cost of assigning

The cost of assigning depends on the administration parameter (g), and the assignment parameter (h). We make the following assumption:

i. The number of assignments depends primarily on the number of updates annually (and not per IP).

If we define F_{mu} as the number of mutations per supplier per time unit (e.g. a year), and c_{ass} as the cost of labor for assigning one update, then the cost of assigning all updates once annually is equal to $S \cdot F_{mu} \cdot c_{ass}$.

Furthermore, we make the following assumption:

j. Based on assumption a., we assume that in case of centralized assignment, the amount of assignment labor is twice as high as in the case of decentralized assignment.

This assumption is reasonable, because senders will always enter a product in their own database according to their own rules, and only then send the product to the central organization. Also, central assignment means that each new physical product is sent to the central organization. If we define the cost of physically sending a product as c_{trans}, the extra cost of central assignment is $S \cdot F_{mu} \cdot c_{trans}$. The question is, what is the advantage of central assignment? The advantage of this strategy is that checking is no longer necessary. The disadvantage is that assignment is performed twice, and physical article transport to the assignment organization is necessary. We will define the transportation cost as $S \cdot F_{mu} \cdot c_{trans}$. Finally, the location of administration also has an impact on the assignment cost. We make the following assumption:

k. In case of central assignment with decentralized administration, we assume that the cost of assigning one update is f times higher than assigning one update with centralized administration. The higher cost of assigning is the result of extra labor time that will be necessary to find the right IP, interpret it, and only then assign, according to the specific IP for the specific bilateral connection, values to the database.

This results in the following cost function for this evaluation parameter.

C_{ass} =

g=(adm)	h(=ass)	Cost term
1	1	$2 \cdot (S \cdot F_{mu} \cdot c_{ass}) + S \cdot F_{mu} \cdot c_{trans}$
1	0	$S \cdot F_{mu} \cdot c_{ass}$
0	1	$2 \cdot (S \cdot F_{mu} \cdot f.c_{ass}) + S \cdot F_{mu} \cdot c_{trans}$
0	0	$S \cdot F_{mu} \cdot c_{ass}$

Where $f > 1$.

Cost of online checking

The cost of online checking depends on three parameters: administration, assignment and checking. Before we will define the specific costs per situation, we will make 3 more assumptions, which together define how the online checking cost is affected in different situations. These assumptions are:

l. Based on assumption e., it follows that in case of central assignment with central administration no checking is allowed, which means that the online checking cost is zero.

m. Central administration means that only one IP is developed with central procedural agreements. Decentralized administration means that each bilateral connection has its own IP and procedural agreements.

n. Central assignment with decentralized administration means that all assignments take place using the bilateral IP's. This means that we loose the advantage of uniform assignment. The result is that the assignment quality with decentralized checking will be no better than with decentralized assignment. Therefore, the checking cost in the centralized situation is equal to the cost in the decentralized checking situation where the checking amount of work depends on the number of IP's.

This results in the following cost function:

C_{check} =

g(=adm)	h (=ass)	l (=check)	Cost term online checking
1	1	2	$1 \cdot c_{check}$
1	1	1	$R \cdot c_{check}$
1	1	0	0
1	0	2	$1 \cdot c_{check}$
1	0	1	$R \cdot c_{check}$
1	0	0	$\propto$ (not allowed)
0	1	2	$S.anr_s.c_{check}$
0	1	1	$S.anr_s.c_{check}$
0	1	0	$\propto$ (not allowed)
0	0	2	$S.anr_s.c_{check}$
0	0	1	$S.anr_s.c_{check}$
0	0	0	$\propto$ (not allowed)

In this table, c_{check} is the cost of online checking for one IP per time unit (e.g. annually). As we can see, central administration with central assignment results in no cost of checking. Central administration with decentralized assignment leads to the cost of checking one IP in case of centralized checking; to the cost of R · one IP in case of decentralized checking; and, to infinite cost in case of no checking. Finally, following rule 3, we see that the cost of decentralized administration results in the number of bilateral links · cost of checking one IP, independently of where checking takes place.

Cost of ongoing profile maintenance
The cost of ongoing profile maintenance is the administration cost of first time entering and maintaining changes in the retailer's assortment specification. We will make the following assumption:
o. The online profile maintenance cost is equal for all 3 situations.

This is explained as follows. In case of sender based profile maintenance, we assume that the receiver enters and updates his profile via the Internet in the sender's profile maintenance application. In case of central profile maintenance, we assume the same happens, with the only difference that only one profile maintenance program is required. Finally, in case of receiver's based profile maintenance, the sender sends a standard update to the receiver, after which the receiver's profile maintenance program automatically selects the right information. Thus, in this case the receiver enters the information directly in his own profile maintenance application. Evaluating all 3 situations, we see that the cost of ongoing profile maintenance depends on the amount of work and the cost of transportation. Since the workload for profile maintenance is depending only on the receiver in all 3 situations, we see that the average ongoing labor cost of profile maintenance per retailer per time unit (c_{prof}) is equal in all 3 situations. Furthermore, since we assume that the transportation cost is not relevant, the total cost of profile maintenance $C_{prof} = R \cdot c_{prof}$ for all situations.

This leads to the following cost function:

C_{prof} =

k(=prof)	Cost term
2	$R \cdot c_{prof}$
1	$R \cdot c_{prof}$
0	$R \cdot c_{prof}$

9.3.2 Restricting the number of scenarios

We will restrict the total number of scenarios as much as possible to simplify the description of the object function. Based on the six design parameters we identified, the total number of possible scenarios is:

Where administration of agreements (central, local per pair) = 2

Where assignment (central, local at supplier)	= 2
Where checking (central, local at retailer, none)	= 3
Where testing (central, bilateral)	= 2
Where profile maintenance (central, supplier, retailer)	= 3 +
Total number of scenarios	72

However, there is one set of scenarios, which is not allowed because of the following restriction:

- Only in the case of central assignment with central administration, no checking is allowed.

This means that 18 scenarios are excluded, which results in 54 scenarios allowed. We have solved this by assuming infinite costs (∞) in these situations.

9.3.3 The object function in terms of a Binary Integer Programming model

We define the following object function:

Min! $C_{transappl} + C_{checkappl} + C_{test} + C_{profappl} + C_{adm} + C_{ass} + C_{check} + C_{prof} =$

Translation: $(1-g) \cdot 2 \cdot S \cdot anr_s \cdot c_{transappl} + g \cdot (R+S) \cdot c_{transappl} +$

Checkapplication: $(1-g-h-i) \cdot INF + (2g-4h+i) \cdot 0 + (i-g-h-1) \cdot 1 \cdot c_{checkappl} + (h-g-i) \cdot INF + (h-3g+2i-2) \cdot 0 + (h-g+i-2) \cdot 1 \cdot c_{checkappl} + (g-h-i) \cdot INF + (g-3h+2i-2) \cdot 0 + (g-h+i-2) \cdot 1 \cdot c_{checkappl} + (g+h-i-1) \cdot 0 + (g+h+3i-4) \cdot 0 + (g+h+i-3) \cdot 1 \cdot c_{checkappl} +$

Testing: $(1-g-j) \cdot S.anr_s \cdot c_{testing} + (g-j) \cdot S^* anr_s * c_{testing} + (j-g) \cdot S^* anr_s * c_{testing} + (g+j-1) \cdot (S+R) \cdot c_{testing} +$

Profappplication: $(1-k) \cdot R \cdot c_{profappl} + k \cdot S \cdot c_{profappl} + (k-1) \cdot 1 \cdot c_{profappl} +$

Administration: $(1-g) \cdot S.anr_s \cdot c_{adm} + g \cdot 1 \cdot c_{adm} +$

Assigning: $(1-g-h) \cdot S \cdot F_{mu} \cdot c_{ass} + (g-h) \cdot S \cdot F_{mu} \cdot c_{ass} + (h-g) \cdot \{2 \cdot (S \cdot F_{mu} \cdot f.c_{ass}) + S \cdot F_{mu} \cdot c_{trans}\} + (g+h-1) \cdot \{2 \cdot (S \cdot F_{mu} \cdot c_{ass}) + S \cdot F_{mu} \cdot c_{trans}\} +$

Online checking: $(1-g-h-i) \cdot INF + (2g-4h+i) \cdot S.anr_s.c_{check} + (i-g-h-1) \cdot S.anr_s.c_{check} + (h-g-i) \cdot INF + (h-3g+2i-2) \cdot S.anr_s.c_{check} + (h-g+i-2) \cdot S.anr_s.c_{check} + (g-h-i) \cdot INF + (g-3h+2i-2) \cdot R \cdot c_{check} + (g-h+i-2) \cdot 1 \cdot c_{check} + (g+h-i-1) \cdot 0 + (g+h+3i-4) \cdot R \cdot c_{check} + (g+h+i-3) \cdot 1 \cdot c_{check} +$

Ongoing profile maint.: $(1-k) \cdot R \cdot c_{prof} + k \cdot R \cdot c_{prof} + (k-1) \cdot R \cdot c_{prof}$

Under the following restrictions:

- $g, h, j \in \{0,1\}$ (decision parameters g, h and j are either 0 or 1)
- $i, k \in \{0,1,2\}$ (decision parameters i, k are either 0, 1 or 2)
- $INF = 1{,}0 * 10^{10}$ (INF must be a very large number)
- $R \geq 1, S \geq 1$ (a network always consists of at least one sender and one receiver)
- $f>1$ (parameter in assignment function)
- $0 \leq$ all terms in object function ≤ 1, which means:

Terms translation
- $0 \leq 1-g \leq 1$
- $0 \leq g \leq 1$

Terms check application
- $0 \leq (1-g-h-i) \leq 1$
- $0 \leq (2g-4h+i) \leq 1$
- …….
- $0 \leq (g+h+i-3) \leq 1$

Terms Testing
- ……

Terms Ongoing Profile maintenance
- $0 \leq (1-k) \leq 1$
- $0 \leq k \leq 1$
- $0 \leq (k-1) \leq 1$

9.4 Analysis

The object function in the previous section is very easy to solve if the required cost variables are known. Since there are only 72 possible solutions, a linear programming software program will require only a fraction of a second to calculate the optimum. The optimum could even be easily calculated by evaluating all scenarios manually.

However, to get a picture about which scenario is best under what circumstance, we will further analyze the object function. The objective of this analysis is to find the conditions under which a certain scenario will prevail. We can use this information to construct a decision tree, helping update distribution network designers to make choices for the construction of the network. However, whenever detailed information about the costs are available, we strongly suggest to calculate the solution using the object function of the Binary Integer Programming model.

9.4.1 Qualitative picture

The objective of the analysis is to eliminate as much of the 72 scenarios as possible. To do this, we will first analyze how the decision parameters affect the cost functions. This picture is shown in Figure 9-2.

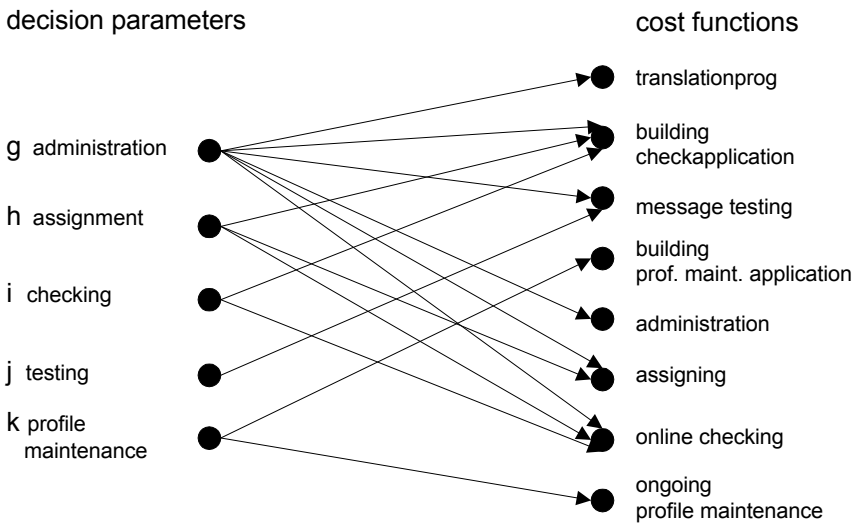


Figure 9-2: Relations between the design and evaluation parameters

Looking at this picture, we see two things:

1. The profile maintenance parameter affects the building profile maintenance program cost + the ongoing profile maintenance cost, which are only affected by this parameter. If we evaluate the profile maintenance parameter first, this leads to a choice between k=0, k=1 or k=2. This means that the total number of scenarios is reduced with a factor 3.
2. The other decision parameters ultimately depend on each other to select the optimal scenario.

Based on these observations, we will first analyze the profile maintenance parameter. Next, we will analyze the remaining 24 scenarios collectively.

9.4.2 The profile maintenance parameter

Table 9-1 gives an overview of the possible values of the profile maintenance parameter (k) and the total profile maintenance costs ($C_{profappl} + C_{prof}$).

k(=prof)	$C_{profappl} + C_{prof}$	
0	$R \cdot c_{profappl} +$	$R \cdot c_{prof}$
1	$S \cdot c_{profappl} +$	$R \cdot c_{prof}$
2	$1 \cdot c_{profappl} +$	$R \cdot c_{prof}$

Table 9-1 : The total cost of profile maintenance

If we compare these cost functions, we see that k=2 (that is centralized profile maintenance) is always cheapest, unless we have a network of only 1 sender and 1 receiver, which is the minimum number of participants to form a network. Only in that case profile maintenance may be either centralized (k=2), at the supplier (k=1), or at the receiver (k=0). Since in that case we may choose a solution, we select centralized profile maintenance. This means that in all situations, centralized profile maintenance is preferable ((k=2). This leads to rule 1:

Rule 1: Centralized profile maintenance is always preferable. (k=2).

9.4.3 Analyzing the remaining cost functions

We will describe the 24 remaining scenarios in a matrix for all values of the decision parameters g,h,i,j. To do this, we will first present the remaining 6 cost parameters with all their possible values (see Table 9-2).

Cost term	Values				
$C_{transappl}$	$(R+S) \cdot c_{transappl}$	$2 \cdot S \cdot anr_s \cdot c_{transappl}$			
$C_{checkappl}$	$1 \cdot c_{checkappl}$	0	$\propto$		
C_{test}	$(R+S) \cdot c_{testing}$	$S \cdot anr_s \cdot c_{testing}$			
C_{adm}	$1 \cdot c_{adm}$	$S \cdot anr_s \cdot c_{adm}$			
C_{ass}	$2 \cdot (S \cdot F_{mu} \cdot c_{ass})$	$S \cdot F_{mu} \cdot c_{ass}$	$2 \cdot (S \cdot F_{mu} \cdot f.c_{ass})$		
	$+ S \cdot F_{mu} \cdot c_{trans}$		$+ S \cdot F_{mu} \cdot c_{trans}$		
C_{check}	$1 \cdot c_{check}$	$R \cdot c_{check}$	$S \cdot anr_s \cdot c_{check}$	0	$\propto$

Table 9-2: Overview of cost parameters and all possible values

Based on this overview, we created a matrix of the remaining 24 scenarios, which is presented in Table 9-3 on page 171. Looking at the cost functions in this table, we can immediately eliminate a number of scenarios.

Firstly, scenarios 11, 12, 17, 18, 23, 24 can be eliminated because the cost functions are infinite. This leads to the following rule:

Rule 2: scenarios 11, 12, 17, 18, 23, 24 can be eliminated because the cost functions are infinite.

We will now analyze the last 12 scenarios, and then the first 12 scenarios.

The last 12 scenarios

Looking at the last 12 scenarios (13 to 24), we see that:

scenario 13 = scenario 15 + $1 \cdot c_{checkappl}$, and
scenario 14 = scenario 16 + $1 \cdot c_{checkappl}$

This leads to the following rule:

Rule 3: Scenarios 13 and 14 can be eliminated because scenarios 15 and 16 are always cheaper.

scen. nr.	g	h	i	J	$C_{transappl}$	$C_{checkappl}$	C_{test}	C_{adm}	C_{ass}		C_{check}
					$c_{transappl}$	$c_{checkappl}$	c_{test}	c_{adm}	$S \cdot F_{mu} \cdot c_{ass}$	$S \cdot F_{mu} \cdot c_{trans}$	c_{check}
1	1	1	2	1	R+S	1	R+S	1	2	1	1
2	1	1	2	0	R+S	1	$S \cdot anr_s$	1	2	1	1
3	1	1	1	1	R+S	0	R+S	1	2	1	R
4	1	1	1	0	R+S	0	$S \cdot anr_s$	1	2	1	R
5	1	1	0	1	R+S	0	R+S	1	2	1	0
6	1	1	0	0	R+S	0	$S \cdot anr_s$	1	2	1	0
7	1	0	2	1	R+S	1	R+S	1	1	0	1
8	1	0	2	0	R+S	1	$S \cdot anr_s$	1	1	0	1
9	1	0	1	1	R+S	0	R+S	1	1	0	R
10	1	0	1	0	R+S	0	$S \cdot anr_s$	1	1	0	R
11	1	0	0	1	R+S	∞	R+S	1	1	0	∞
12	1	0	0	0	R+S	∞	$S \cdot anr_s$	1	1	0	∞
13	0	1	2	1	$2 \cdot S \cdot anr_s$	1	$S \cdot anr_s$	$S \cdot anr_s$	$2 \cdot f$	1	$S \cdot anr_s$
14	0	1	2	0	$2 \cdot S \cdot anr_s$	1	$S \cdot anr_s$	$S \cdot anr_s$	$2 \cdot f$	1	$S \cdot anr_s$
15	0	1	1	1	$2 \cdot S \cdot anr_s$	0	$S \cdot anr_s$	$S \cdot anr_s$	$2 \cdot f$	1	$S \cdot anr_s$
16	0	1	1	0	$2 \cdot S \cdot anr_s$	0	$S \cdot anr_s$	$S \cdot anr_s$	$2 \cdot f$	1	$S \cdot anr_s$
17	0	1	0	1	$2 \cdot S \cdot anr_s$	∞	$S \cdot anr_s$	$S \cdot anr_s$	$2 \cdot f$	1	∞
18	0	1	0	0	$2 \cdot S \cdot anr_s$	∞	$S \cdot anr_s$	$S \cdot anr_s$	$2 \cdot f$	1	∞
19	0	0	2	1	$2 \cdot S \cdot anr_s$	1	$S \cdot anr_s$	$S \cdot anr_s$	1	0	$S \cdot anr_s$
20	0	0	2	0	$2 \cdot S \cdot anr_s$	1	$S \cdot anr_s$	$S \cdot anr_s$	1	0	$S \cdot anr_s$
21	0	0	1	1	$2 \cdot S \cdot anr_s$	0	$S \cdot anr_s$	$S \cdot anr_s$	1	0	$S \cdot anr_s$
22	0	0	1	0	$2 \cdot S \cdot anr_s$	0	$S \cdot anr_s$	$S \cdot anr_s$	1	0	$S \cdot anr_s$
23	0	0	0	1	$2 \cdot S \cdot anr_s$	∞	$S \cdot anr_s$	$S \cdot anr_s$	1	0	∞
24	0	0	0	0	$2 \cdot S \cdot anr_s$	∞	$S \cdot anr_s$	$S \cdot anr_s$	1	0	∞

Table 9-3: The costs of the remaining 24 scenarios

Secondly, we see that:

scenario 19 = scenario 21 + $1 \cdot c_{checkappl,}$, and
scenario 20 = scenario 22 + $1 \cdot c_{checkappl}$

This leads to the following rule:

Rule 4: Scenarios 19 and 20 can be eliminated because scenarios 21 and 22 are always cheaper.

Thirdly, if we compare scenarios 15 en 16 with scenarios 21 and 22, we see that

scenario 15 = scenario 21 + $(2 \cdot f - 1) \cdot S \cdot F_{mu} \cdot c_{ass} + S \cdot F_{mu} \cdot c_{trans}$, and

scenario 16 = scenario 22 + $(2 \cdot f - 1) \cdot S \cdot F_{mu} \cdot c_{ass} + S \cdot F_{mu} \cdot c_{trans}$, and

Since $f > 1$, and both $S \cdot F_{mu} \cdot c_{ass}$ and $S \cdot F_{mu} \cdot c_{trans} > 0$, this leads to the following rule:

Rule 5: Scenarios 15 and 16 can be eliminated because scenarios 21 and 22 are always cheaper.

The first 12 scenarios

Looking at the first 12 scenarios (1 to 12), we see that:

scenario 1 = scenario 5 + $1 \cdot c_{checkappl,} + 1 \cdot c_{check}$, and

scenario 2 = scenario 6 + $1 \cdot c_{checkappl} + 1 \cdot c_{check}$

This leads to the next rule:

Rule 6: Scenarios 1 and 2 can be eliminated because scenarios 5 and 6 are always cheaper.

Secondly, if we compare scenarios 3 and 4 with scenarios 5 and 6, we see that:

scenario 3 = scenario 5 + $R \cdot c_{check}$, and

scenario 4 = scenario 6 + $R \cdot c_{check}$

This leads to the next rule:

Rule 7: Scenarios 1 and 2 can be eliminated because scenarios 5 and 6 are always cheaper.

Analyzing the remaining eight scenarios

If we apply rule 2 – 7 on Table 9-3, we have the following remaining eight scenarios:

scen.	g	h	i	j	$C_{transappl}$	$C_{checkappl}$	C_{test}	C_{adm}	C_{ass}		C_{check}
nr.					$c_{transappl}$	$c_{checkappl}$	c_{test}	c_{adm}	$S \cdot F_{mu} \cdot c_{ass}$	$S \cdot F_{mu} \cdot c_{trans}$	c_{check}
5	1	1	0	1	R+S	0	R+S	1	2	1	0
6	1	1	0	0	R+S	0	$S \cdot anr_s$	1	2	1	0
7	1	0	2	1	R+S	1	R+S	1	1	0	1
8	1	0	2	0	R+S	1	$S \cdot anr_s$	1	1	0	1
9	1	0	1	1	R+S	0	R+S	1	1	0	R
10	1	0	1	0	R+S	0	$S \cdot anr_s$	1	1	0	R
21	0	0	1	1	$2 \cdot S \cdot anr_s$	0	$S \cdot anr_s$	$S \cdot anr_s$	1	0	$S \cdot anr_s$
22	0	0	1	0	$2 \cdot S \cdot anr_s$	0	$S \cdot anr_s$	$S \cdot anr_s$	1	0	$S \cdot anr_s$

Table 9-4: Remaining eight scenarios

We can further reduce these scenarios as follows.

First, we will compare scenarios 7, 8 with scenarios 9, 10. We see that:

scenario 7 = scenario 9 + $1 \cdot c_{checkappl} + (1 - R) \cdot c_{check}$, and

scenario 8 = scenario 10 + $1 \cdot c_{checkappl} + (1 - R) \cdot c_{check}$.

This leads to the following rule:

Rule 8: If $1 \cdot c_{checkappl} + 1 \cdot c_{check} > R \cdot c_{check}$, then scenarios 9,10 are cheaper, otherwise scenarios 7, 8.

Next, we will compare scenarios 5, 6 with scenarios 7,8 and scenarios 9,10 respectively. We see that:

$$Sc.\ 5 - Sc.\ 7 = -1 \cdot c_{checkappl} + S \cdot F_{mu} \cdot c_{ass} + S \cdot F_{mu} \cdot c_{trans} - 1 \cdot c_{check}) \rightarrow$$
$$Sc.\ 5 = Sc.\ 7 + S \cdot F_{mu} \cdot (c_{ass} + c_{trans}) - 1 \cdot c_{checkappl} - 1 \cdot c_{check}$$

The same holds for scenario 6 and 8.

$$Sc.\ 5 - Sc.\ 9 = S \cdot F_{mu} \cdot c_{ass} + S \cdot F_{mu} \cdot c_{trans} - R \cdot c_{check} \rightarrow$$
$$Sc.\ 5 = Sc.\ 9 + S \cdot F_{mu} \cdot (c_{ass} + c_{trans}) - R \cdot c_{check})$$

The same holds for scenario 6 and 10.

This leads to the following rules:

Rule 9a: if $S \cdot F_{mu} \cdot (c_{ass} + c_{trans}) > (1 \cdot c_{checkappl} + 1 \cdot c_{check})$, then scenarios 7,8 are cheaper, otherwise scenarios 5,6

Rule 9b: if $S \cdot F_{mu} \cdot (c_{ass} + c_{trans}) > R \cdot c_{check}$, then scenario 9,10 are cheaper, otherwise scenarios 5,6

Finally, we still have to compare scenarios 5,6 scenarios 7, 8 and scenarios 9,10 with scenarios 21, 22. Since this is not very easy, we will make a reasonable assumption, namely that the number of links between suppliers, retailers and a central point is smaller than the number of bilateral links between suppliers and retailers. Thus:

Assumption 1: $(R+S) < S.anr_s$

In plain English, this assumption means that the number of central links < number of bilateral links. Applying this assumption to the scenarios in Table 9-4, we see that in that case centralized testing is always preferred (eliminating scenarios 6, 8, 10 and 22). This leads to the following result table (see Table 9-5).

scen. nr.	G	h	l	j	$C_{transappl}$	$C_{checkappl}$	C_{test}	C_{adm}	C_{ass}		C_{check}
					$c_{transappl}$	$c_{checkappl}$	c_{test}	c_{adm}	$S \cdot F_{mu} \cdot c_{ass}$	$S \cdot F_{mu} \cdot c_{trans}$	c_{check}
5	1	1	0	1	R+S	0	R+S	1	2	1	0
7	1	0	2	1	R+S	1	R+S	1	1	0	1
9	1	0	1	1	R+S	0	R+S	1	1	0	R
21	0	0	1	1	$2 \cdot S \cdot anr_s$	0	$S \cdot anr_s$	$S \cdot anr_s$	1	0	$S \cdot anr_s$

Table 9-5: Resulting four scenarios under assumption 1

We will compare scenarios 5, 7, 9 with scenario 21 respectively. First, if we compare scenario 9 with 21, we see that:

(1) $Sc\ 21 - Sc\ 9 = \{2 \cdot S \cdot anr_s - (R+S)\} \cdot c_{transappl} + \{S \cdot anr_s - (R+S)\} \cdot c_{test} + \{S \cdot anr_s - 1\} \cdot c_{adm} + \{S \cdot anr_s - R\} \cdot c_{check}$

Since $S{\cdot}anr_s > R+S$, and $R+S > 2$ (because we have at least one supplier and one retailer in a network), and $\{S{\cdot}anr_s - R\}$ is at least $R+S-R$, we see that all terms at the right side of the equation are always larger than 0. Thus, Scenario 21 is always more expensive than scenario 9. in plain English, this leads to the next rule.

Rule 10: Under the assumption that # central links < # bilateral links$_s$ scenario 9 is always cheaper than scenario 21.

Second, if we compare scenario 7 with scenario 21, we see that:

(3) $Sc.\ 21 - Sc\ 7 = \{2{\cdot}S{\cdot}anr_s - (R+S)\}{\cdot}c_{transappl} - 1{\cdot}c_{checkappl} + \{S{\cdot}anr_s - (R+S)\} {\cdot}c_{test} + \{S{\cdot}anr_s - 1\}{\cdot}c_{adm} + \{S{\cdot}anr_s - 1\}{\cdot}c_{check}$

If we define $\{2{\cdot}S{\cdot}anr_s - (R+S)\}{\cdot}c_{transappl} + \{S{\cdot}anr_s - (R+S)\} {\cdot}c_{test} + \{S{\cdot}anr_s - 1\}{\cdot}c_{adm} + \{S{\cdot}anr_s - 1\}{\cdot}c_{check}$ as V_1 we get the following equation:

(4) $Sc.\ 21 = Sc\ 7 + V_1 - 1{\cdot}c_{checkappl}$, where V_1 is a positive constant.

Equation (4) shows that Scenario 7 is always cheaper than scenario 21, unless $1{\cdot}c_{checkappl} > V_1$, where $V_1 = \{2{\cdot}S{\cdot}anr_s - (R+S)\}{\cdot}c_{transappl} + \{S{\cdot}anr_s - (R+S)\} {\cdot}c_{test} + \{S{\cdot}anr_s - 1\}{\cdot}c_{adm} + \{S{\cdot}anr_s - 1\}{\cdot}c_{check}$.

Third, if we compare scenario 5 with scenario 21, we see that:

(5) $Sc.\ 21 - Sc\ 5 = \{2{\cdot}S{\cdot}anr_s - (R+S)\}{\cdot}c_{transappl} + \{S{\cdot}anr_s - (R+S)\} {\cdot}c_{test} + \{S{\cdot}anr_s - 1\}{\cdot}c_{adm} - 1 {\cdot}S{\cdot}F_{mu} {\cdot}c_{ass} - 1 {\cdot}S{\cdot}F_{mu} {\cdot}c_{trans} + \{S{\cdot}anr_s\}{\cdot}c_{check}$

If we define $\{2{\cdot}S{\cdot}anr_s - (R+S)\}{\cdot}c_{transappl} + \{S{\cdot}anr_s - (R+S)\} {\cdot}c_{test} + \{S{\cdot}anr_s - 1\}{\cdot}c_{adm} + \{S{\cdot}anr_s\}{\cdot}c_{check}$ as V_2 we get the following equation:

(6) $Sc.\ 21 = Sc.\ 5 + V_2 - 1 {\cdot}S{\cdot}F_{mu} {\cdot}(c_{ass} + c_{trans})$

Equation (6) shows that Scenario 5 is always cheaper than scenario 21, unless $1 {\cdot}S{\cdot}F_{mu} {\cdot}(c_{ass} + c_{trans}) > V_2$ where $V_2 = \{2{\cdot}S{\cdot}anr_s - (R+S)\}{\cdot}c_{transappl} + \{S{\cdot}anr_s - (R+S)\} {\cdot}c_{test} + \{S{\cdot}anr_s - 1\}{\cdot}c_{adm} + \{S{\cdot}anr_s\}{\cdot}c_{check}$

To make equation (4) and (6) understandable in plain English, we will make a more restricting assumption than assumption 1, which is as follows:

Assumption 2: $\mathbf{(R+S) << S.anr_s}$ (the number of links to a central point is much smaller than the number of bilateral links), so that $S.anr_s - (R+S) \approx S.anr_s$.

In a normal business network, this is still a reasonable assumption. This leads to the following 2 equations for equation (4) and (6):

(7) $Sc.\ 21 = Sc\ 7 + V_1 - 1{\cdot}c_{checkappl}$, where $V_1 = S{\cdot}anr_{s^*} (2{\cdot}c_{transappl} + c_{test} + c_{adm} + c_{check})$

(8) $Sc.\ 21 = Sc.\ 5 + V_2 - 1 {\cdot}S{\cdot}F_{mu} {\cdot}(c_{ass} + c_{trans})$, where $V_2 = V_1 = S{\cdot}anr_{s^*} (2{\cdot}c_{transappl} + c_{test} + c_{adm} + c_{check})$

Using this assumption, we can now construct the last 2 rules in plain English:

Rule 11 (based on equation 7):
Under the assumption that there are much more bilateral than central links, scenario 7 is always cheaper than scenario 21, unless the cost of building one central checking application is higher than the total costs of $\{2 \cdot c_{transappl} + c_{test} + c_{adm} + c_{check}\}$ per link combined.

Rule 12 (based on equation 8):
Under the assumption that there are much more bilateral than central links, scenario 5 is always cheaper than scenario 21, unless the cost of physical transportation and assigning all product mutations is higher than the total costs of $\{2 \cdot c_{transappl} + c_{test} + c_{adm} + c_{check}\}$ per link for all links combined.

9.4.4 Constructing a decision tree

Applying rules 8 to 12, we can construct a decision tree to choose between the remaining 8 scenarios. Figure 9-3 on page 177 shows this decision tree. In the tree we substituted the mathematical equations of the rules in plain English. The tree shows the decision path of the most probable decisions. These decisions are respectively:

(1) In a normal business network, the number of central links is much smaller than the number of bilateral links.
- 4 scenarios left: scenarios 5, 7, 9, and 21

(2) The cost of having 1 central online checking function + the cost of building a central checking application, are normally lower than the total costs of all retailers having a separate online checking function per IP.
- Scenario 9 eliminated.

(3) The cost of having 1 central online checking function + the cost of building 1 checking application are normally much lower than the total costs of physical transportation and assigning all product mutations in the sector centrally.
- Scenario 5 eliminated.

(4) Normally, the cost of building or buying a central checking application is much lower than the costs of {2* building a translation application + testing, administration and checking per link} for all links combined.
- Scenario 21 eliminated.

This leads to the conclusion that under normal circumstances (as described in (1) to (4)) central administration, with decentralized assignment and with central checking, testing and profile maintenance is preferred, which is consistent with our case study findings.

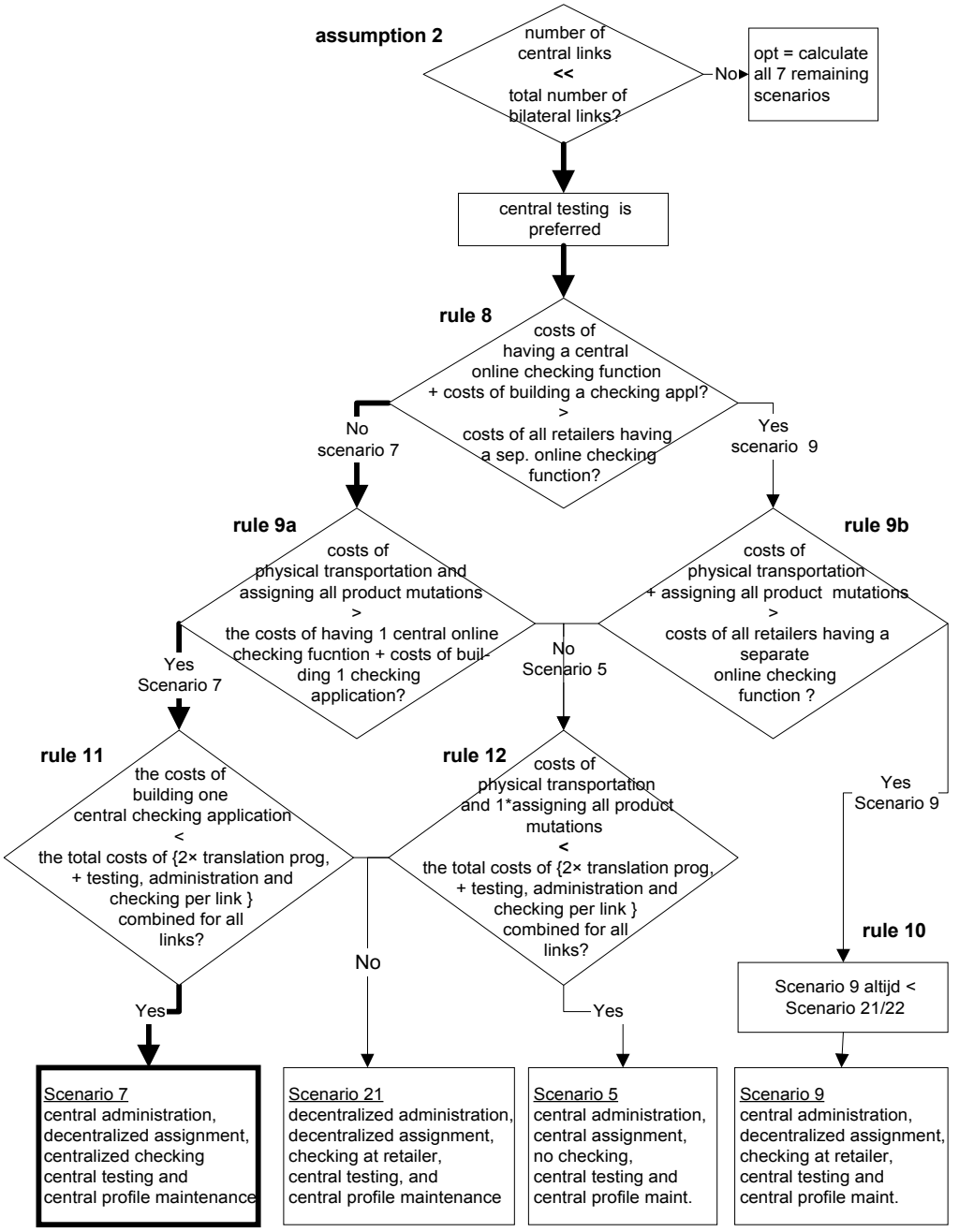


Figure 9-3: Decision tree

10. Conclusions and Further Research

10.1 Introduction

The role of data quality in an EDI-enabled interorganizational business process as a result of insufficient context quality, and how to solve this problem through data alignment, is the central issue in this thesis.

In the first part of this thesis we described the problem of insufficient product data quality and how it results from insufficient data alignment in the contexts of senders and receivers. This ended in a proposition about how context quality affects an EDI-enabled business process we tested in the Schwartz case. In the second part of this thesis we developed the Data Alignment through Logistics (DAL) method to solve the data alignment problem. This method has been further developed and tested in the CBL and EAN-DAS case. Additionally, based on the results of these two cases, we constructed a Binary Integer Programming model that helps network or community builders to design the data alignment structure of a large (>50 participants) interorganizational B2B network.

In this final chapter we will present the conclusions of this research. We will start with conclusions about the research methodology in Section 10.2. Next, we will present the conclusions with respect to the research results in Section 10.3. In Section 10.4 we will reflect on the implications of our conclusions with respect to theory and practice. Finally, in Section 10.5 we will present recommendations for further research.

10.2 Conclusions research methodology

Normally, a researcher describes the research methodology, so that other researchers can examine how the research was conducted, if it was conducted using proper scientific methods, and possibly to reconstruct the results of the scientific study.

In Chapter 2, we argued that because of the variety of philosophical backgrounds in IS research, unbiased evaluation of the research results is only possible when the philosophical perspective of the research is made clear to other researchers. Otherwise, the interpretation of the research results is troubled with the own philosophical assumptions of the evaluator. Once the character of the research is clear, other researchers must understand what research strategy the researcher used, and which methods were used in this strategy. In this way, other researchers are able to check if the research strategy and method(s) that the researcher has chosen are consistent with the character of the research.

Therefore, we summarized the research perspective, strategy and methods of this research in Figure 10-1.

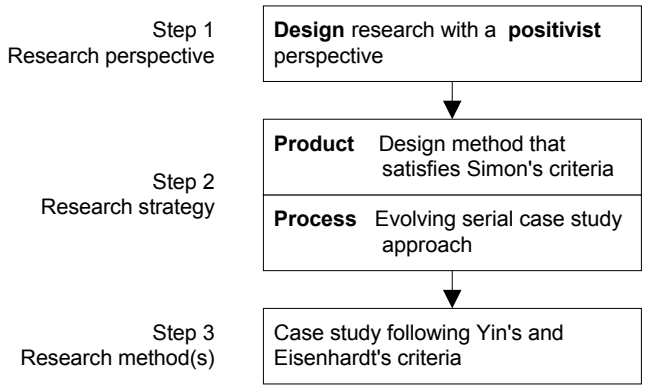


Figure 10-1: Research perspective, strategy and method of this research

Wit respect to the research methodology of this research we draw the following conclusion:

- This research can be classified as *design* research with a *positivist* perspective. This research was design oriented because the objective of our research was to develop a method that solves a specific problem (namely the problem of insufficient data quality of product information on sector level). Furthermore, this research was based on a positive epistemology because the objective of the method was to provide a *general* method for data alignment to be used by data integration professionals.

10.3 Conclusions research results

The conclusions with respect to the results of this research will be discussed by answering the research questions we presented in Secion 1.4.2. Finally, we will discuss to what extent we have achieved our general research objective we presented in Section 1.4.1.

10.3.1 Relation between context and data quality

Research question 1
What is the relation between context and data quality?

Answer question 1
There are three views on data quality, the reliability view, the relevance view, and the context view. The *reliability* view defines data quality as conformance of data to reality. The *relevance* view defines data quality as conformance of the data to the user's expectations (=fitness for use). The *context* view defines data quality as conformance to the shared context of sender and receiver in a communication process. This multiple view model of data quality is shown in Figure 10-2.

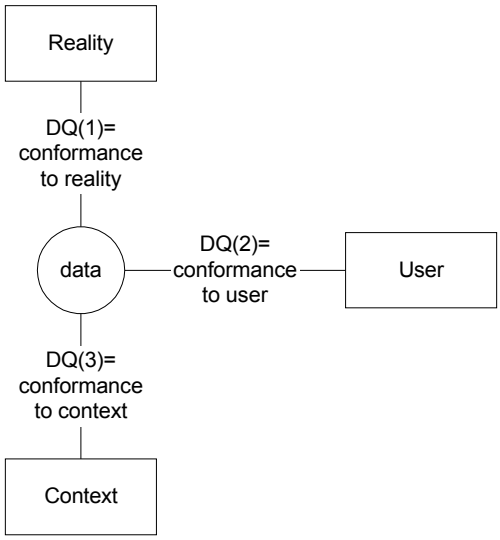


Figure 10-2: Multiview model of data quality

Using this model we can now answer our first research question. With respect to electronic message exchange in a communication process, the data quality of the message depends on the shared context. Since the shared context also has a certain quality (defined as the context quality), the relation between data quality and context is that the data quality of a message depends on the context quality.

10.3.2 The role of data quality in EDI

Research question 2
What is the role of data quality in an EDI-enabled business process?

Answer question 2
An explanation of the role of data quality in an EDI-enabled business process is shown in Figure 10-3.

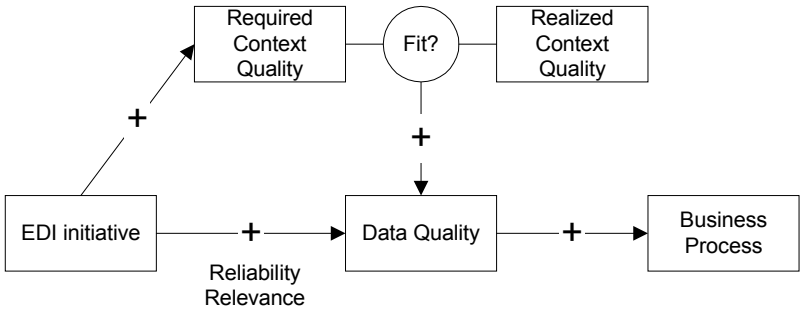


Figure 10-3: Explanation of the role of data quality in an EDI enabled business process

Figure 10-3 shows a direct and an indirect relation between EDI and data quality.

The direct relation between EDI and data quality is explained as follows. EDI positively influences the data quality of the information in an EDI message, because EDI increases both the reliability and the relevance of that information. The reliability of the information is improved because EDI leads to fewer data entry errors. Furthermore, the relevance of an EDI message is improved because EDI leads to a *complete* specification of the EDI information, and to faster communication of the information. Since relevance is defined in terms of completeness and timeliness of information, EDI use increases the relevance of the information in the EDI message. Thus, EDI positively improves the reliability and the relevance of the information in an EDI message, and hence positively influences data quality.

The indirect relation between EDI and data quality is explained as follows. An EDI initiative results in higher requirements for the quality of the context between sender and receiver, because EDI implementation eliminates human intervention in the transaction communication between business processes. As we learned from the first research question, the quality of the context positively influences the quality of the data in the EDI message.
Without EDI in the business process, many errors in the transaction communication that are the result of insufficient context data quality are detected and corrected through human interpretation. When EDI is implemented, this human correction mechanism is not in place. Hence, EDI results in a higher required context quality.

If the realized context quality matches the required context quality, the high quality of the context positively influences the data quality of the EDI message, and hence has a positive effect on the business process. However, when the context quality is not improved when EDI is implemented, then the required context quality exceeds the realized context quality. Since context quality positively influences the data quality of the EDI message, the insufficient quality of the context will result in a decreased quality of the data in the EDI message, and thus will negatively influence the business process.

10.3.3 How to specify agreements on the semantics of product data

Research question 3
Which design steps or guidelines should be applied on *sector* level to help organizations in a supplier-retailer network to specify agreements on the semantics of product data?

Answer question 3
From our extensive literature review on data integration and data distribution methods, we learned that in large interorganizational situations the best way to specify agreements on the semantics of product data is the definition of a central schema per specific user group. Translation between local and central schemas is achieved by defining local mappings between the local and central schema (conclusion 1 from Section 6.6).

In the Data Alignment through Logistics (DAL) method we propose, we incorporated this solution through the definition of Information Products (IPs). We define an IP as a semantic schema, which is defined and maintained by a user community (i.e. the interorganizational network users) and which will be used for a specific purpose. By viewing the central schema as a product that is used for specific purpose, we provide focus in the definitions, so that a minimal set of mutual agreements is established.

10.3.4 How to design a data distribution structure for data alignment

Research question 4
Which design steps or guidelines should be applied to help organizations in a supplier-retailer network to develop a data distribution structure for aligning the product data across the network?

Answer question 4
To answer this question we compared fact update distribution with the field of logistics, looking for principles from logistics that we could use for fact update distribution (see Chapter 7). From this review, we learned two things:

- Reduce the complexity of the problem through making a distinction between the static structure of a distribution network and the control structure that deals with the dynamics in the network.
- Find out which parameters determine both the static and dynamic structures of the distribution network.

In the DAL method that we propose, we developed a design model for the static structure of a fact update distribution network, which consists of the following five design parameters:

- Where should the administration function be located? (central, or local per pair);
- Where the assignment function? (central or local at supplier);
- Where the checking function? (central, local at receiver, none);
- Where the testing function? (central, or local per pair);
- Where the profile maintenance function? (central, supplier or retailer).

The complete model that describes the static structure of a fact update distribution network is shown in Figure 10-4.

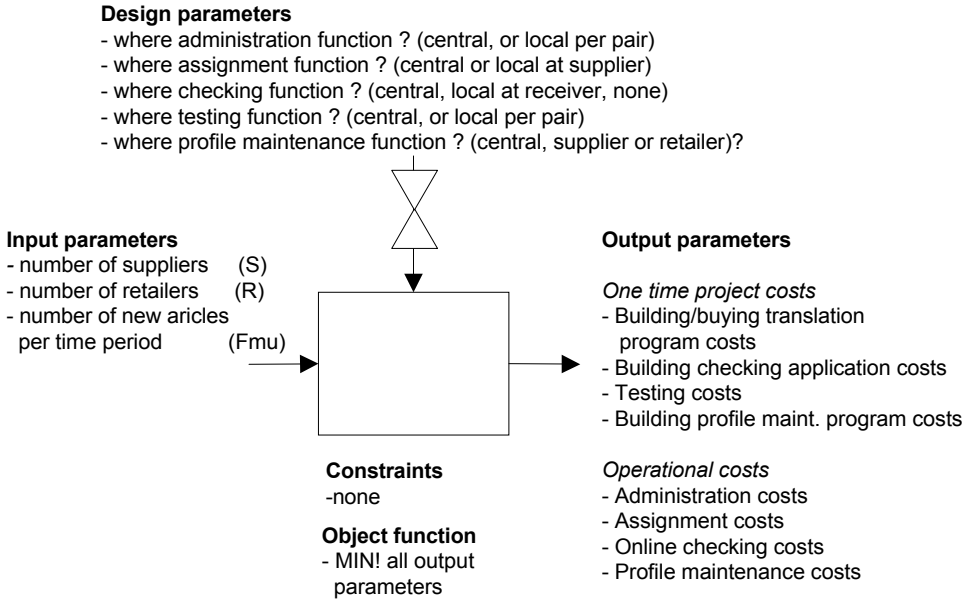


Figure 10-4: Model of the static structure of a data distribution network

Using this model, organizations in a supplier-retailer network are able to determine where to locate the five functions for fact update distribution, thus enabling these organizations to design a fact update distribution network.

10.3.5 Achievement of research objective

Finally, we will determine to what extent we have achieved the research objective we defined at the beginning of this thesis.

Research objective
Develop a method that improves the degree of alignment between *product* databases in the *food* sector, which will increase the quality of EDI order data and therefore the applicability of EDI in interorganizational business processes for *the food sector*.

Objective achieved?
From a theoretical perspective the objective has been partly achieved. Although we have developed and tested a data distribution method for fact updates, we only developed the static structure for a data distribution network. We have not addressed the control structure to control the dynamics of a fact update distribution network. This remains a subject for further research.

From a practical perspective, the objective was achieved, since EAN Nederland (the Dutch EDI standardization organization) is now operating a data alignment service (EAN-DAS) with the objective of aligning the product data of organizations in the Dutch food sector. Several basic functions of this service were developed in strong relation to this research.

10.4 Implications for research and practice

Based on the conclusions we provided in the previous paragraph, we will reflect on the implications of these conclusions for research and practice. We identified the following four implications of our research.

Transaction communication and data alignment are inseparable
Our model about the role of data quality in an EDI-enabled business process shows that an EDI- enabled business process is influenced by the degree of data alignment in its context. Especially when the required context quality is higher than the realized context quality, the direct positive effects of EDI use are decreased through the indirect negative influence of insufficient quality in the context. This finding suggests that an EDI-enabled business process, where coordination is achieved through transaction communication, cannot be separated from the process of data alignment in its context. Hence, transaction communication and data alignment are inseparable. The relation between transaction communication and data alignment is shown in Figure 10-5.

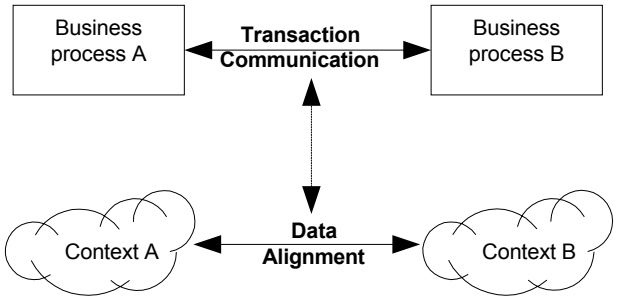


Figure 10-5: Transaction communication and data alignment are inseparable

This finding has especially consequences in the Business Process Redesign (BPR) literature, since this means that redesign of business processes using Information and Communication Technology (ICT), should include redesign of the context of these processes. For practitioners this means that they should be aware of the context in redesigning a business process. Implicitly this already happens, because in EDI-based redesign projects one of the important success factors is making detailed agreements about definitions and procedures.

Shift in multi-database research
Our finding that agreements in interorganizational networks should be based on a community defined central schema, has the following implications:

- Both database researchers and practitioners attempt data integration in interorganizational networks by constructing one overall schema for all participants in the network, which is constructed from a number of local schemas. Our research shows that much less effort is needed when multiple, community defined schemas are constructed and maintained, where the focus is not on construction but on updating through interaction with the user community. This should result in a shift from efficient schema construction methods to data quality checking methods, since in community defined schemas data quality becomes the main issue.
- When we accept the existence of multiple contexts (which is in contrast with the distributed database approach), the issue of data distribution should be replaced on the research agenda for multi-database research. Today, the problem of data distribution between databases is non-existent because the view update method is used for federated databases. However, in case of many network participants, for whom the view integration method does not work, this means that the problem of exchanging fact updates becomes relevant.

Data alignment essential for E-business
In E-business we focus on using Internet technologies in B2B networks. At the beginning of Chapter 1, we concluded that EDI fits in the upper right quadrant of Kraljic's E-business matrix, where the focus is on interorganizational partnerships based on information integration. Our research shows that data alignment plays an essential role in business processes that are redesigned using EDI. Therefore, an important implication of our research is that E-business should incorporate data alignment as an essential part, at least for information integration.

10.5 Recommendations for further research

In this section we will discuss four recommendations for further research.

10.5.1 Recommendations data quality in an EDI enabled business process

With respect to our theory about the role of data quality in an EDI enabled business process, we recommend the following two subjects for further research.

1. Further testing to construct and test a theory about the role of data quality in E-business. In our research we tested in only one case study the proposition about the role of data quality in an EDI enabled business process. This was necessary, because both the data quality and

process impact assessment methods still had to be developed. However, now we know how to measure data quality and process impact, a more substantial test using, for instance, the survey method should be possible. Also, because of the broader perspective E-business provides in interorganizational relationships, we recommend to use our proposition to construct a data quality theory in E-business that is tested using the survey method.

2. Developing a functional relation between data quality and impact. The usability of a theory about data quality in E-business would be greatly enhanced if a functional relation between data quality and impact would exist. This would lead to statements such as: "With a context quality of 85%, processing time spent on prevention will be x% of the complete processing time and the remaining number of errors will be y%". It would mean that the labor-intensive part of the method, which is measuring the actual impact, could be skipped.

10.5.2 Recommendations for the DAL method

For the DAL method, we recommend the following.

1. Further testing of the DAL method. Since the final version of the DAL method has not been tested on a large scale, this should be done first. Only then the method has both theoretical and practical value. Since many of the problems we discussed were also found in the pharmaceutical and the Electrotechnical sector, we recommend testing the DAL method for these two sectors.
2. Inclusion of dynamics in the DAL method. In this version of the DAL method, only the static structure of the fact update distribution network is addressed. Further research should focus on including the control mechanisms into the DAL method, so that data quality levels can be controlled through time.
3. Network design in case of multiple sub-usergroups. The current version of the DAL method assumes that only one central schema is developed per user group. The question is how a fact update distribution network should be constructed when many sub-usergroups exist, each with their own shared schema. Furthermore, it would be interesting to know where the balance is between keeping multiple bilateral schemas versus creating one shared schema, depending on the number of suppliers, retailers, fact update mutations, administration mutations, and possibly the complexity of the IP.
4. Combine the DAL method with research in network coordination. Data alignment basically provides a different mechanism for coordination in computer networks. Therefore, the DAL method should be combined with other research concerning coordination mechanisms in computer networks. Specifically, we recommend researching the relationship between coordination through transaction communication, and coordination through data alignment. Figure 10-6 depicts the relation between transaction communication and data alignment.

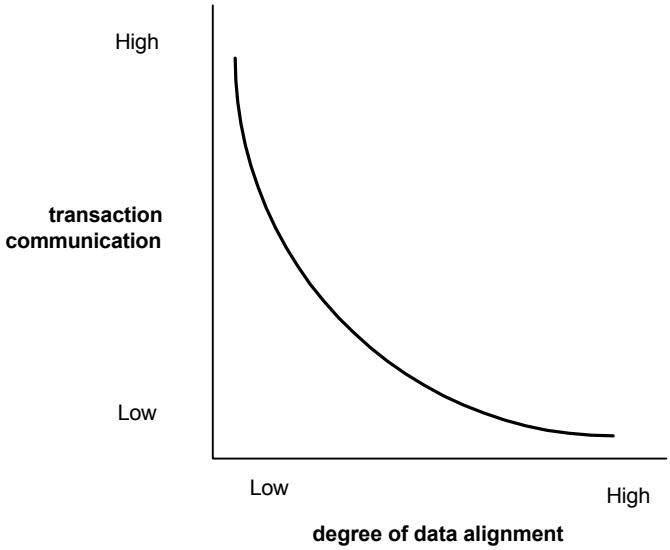


Figure 10-6: Relation between transaction communication and data alignment

According to this relation, a lower degree of data alignment results in a higher cost for transaction communication. This is the result of the fact that before a transaction is established, companies have to inform each other extensively about the required products, the type of price agreements, delivery agreements, etc. Thus, a lot of data alignment is actually established during the transaction communication process. However, when companies frequently interact, investments in data alignment will significantly lower the transaction cost.

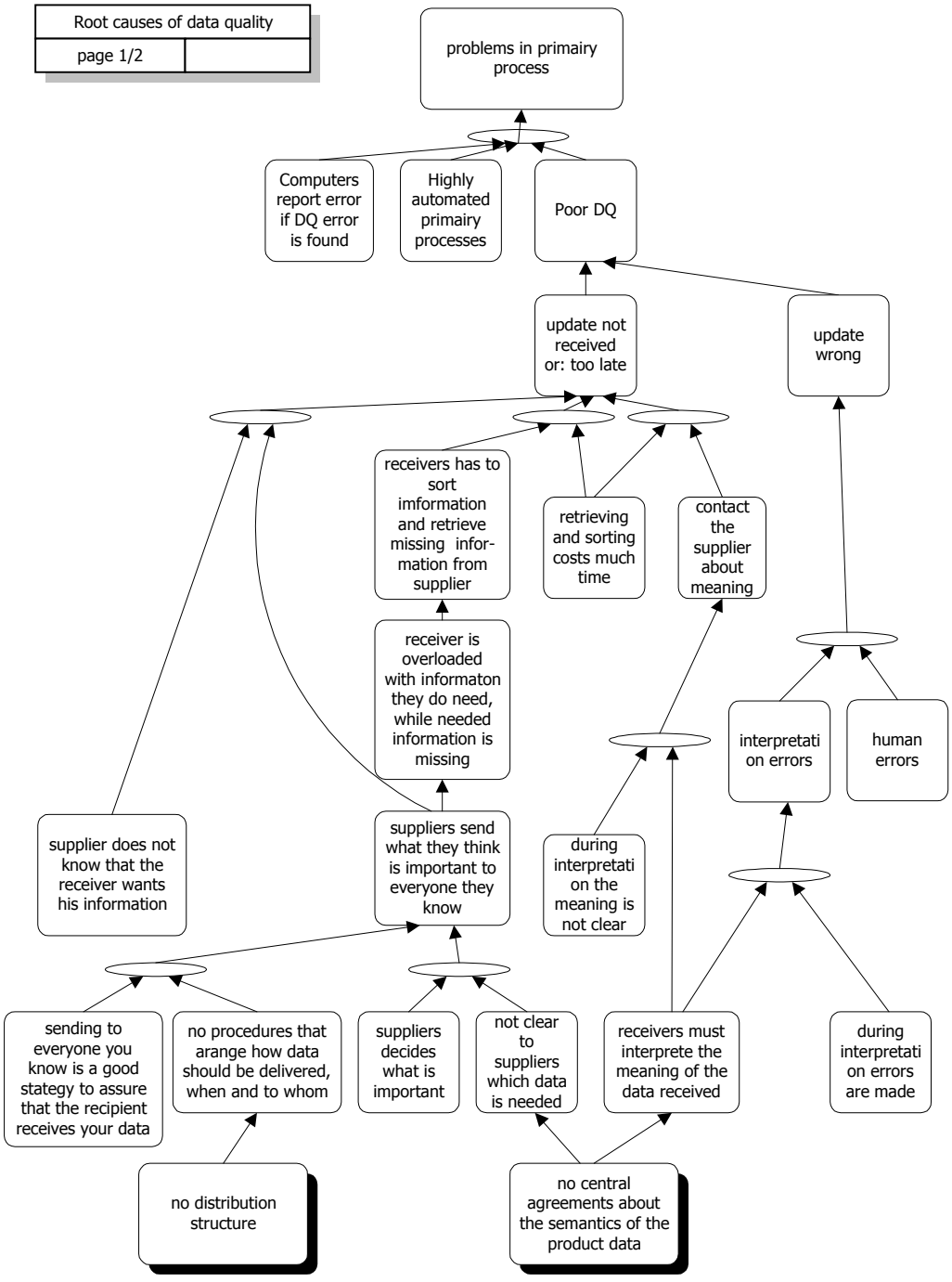
The current DAL method calculates the most economical solution for data alignment. Further research should focus on relating these investments in data alignment to the possible reduction in the cost of transaction communication. In this way, an optimal balance between transaction communication and data alignment can be established.

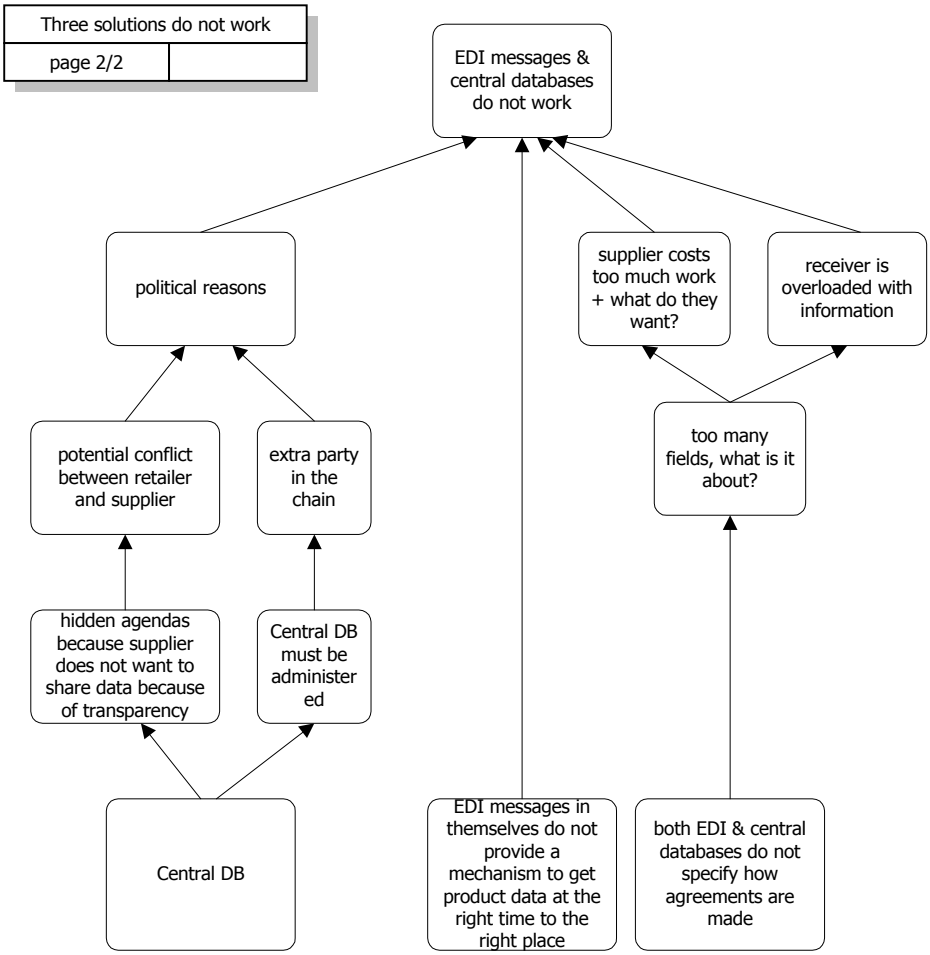
Appendix A

This appendix contains the two Current Reality Trees (CRTs) we used to analyze the three problems discovered in the cases.

- The first CRT describes the root causes of insufficient data quality.
- The second CRT describes why EDI article messages and central databases do not work.

Appendix A


Root causes of data quality
page 1/2
problems in primairy process
Computers report error if DQ error is found
Highly automated primairy processes
Poor DQ
update not received or: too late
update wrong
receivers has to sort imformation and retrieve missing infor-mation from supplier
retrieving and sorting costs much time
contact the supplier about meaning
receiver is overloaded with informaton they do need, while needed information is missing
interpretati on errors
human errors
supplier does not know that the receiver wants his information
suppliers send what they think is important to everyone they know
during interpretati on the meaning is not clear
sending to everyone you know is a good stategy to assure that the recipient receives your data
no procedures that arange how data should be delivered, when and to whom
suppliers decides what is important
not clear to suppliers which data is needed
receivers must interprete the meaning of the data received
during interpretati on errors are made
no distribution structure
no central agreements about the semantics of the product data


Three solutions do not work
page 2/2
EDI messages & central databases do not work
political reasons
supplier costs too much work + what do they want?
receiver is overloaded with information
potential conflict between retailer and supplier
extra party in the chain
too many fields, what is it about?
hidden agendas because supplier does not want to share data because of transparency
Central DB must be administer ed
Central DB
EDI messages in themselves do not provide a mechanism to get product data at the right time to the right place
both EDI & central databases do not specify how agreements are made

Appendix A

Appendix B

This appendix contains the results of the invoice analysis of all order lines in 1996 for errors resulting from insufficient context quality.

Appendix B

Process	Who?	Cause	Total	DQ error	Comment
Ordering	p/adm	other ordercosts not entered	15	x	order related, not product
	p	processed concept order too late	30	x	no DQ error
	p	order changes not processed	18	x	order related
	p	way of invoicing/payment not reported on concept order	2	x	order related
Project pur-chaser	p	unclear prices on concept order	9	9	
	p	specific price agreement(s) not reported on concept order	11	x	order related
	p	right prices/units not reported on concept order	20	20	
	p/s	incorrect ordernumber	35	x	order related
	p	price extra work not known	13	13	
Recep-tion of goods	p	Goods Reception Form (GRF) sent too late	14	x	Goods reception is after the ordering process
	p	deviations not mentioned on GRF	2	x	
	p	part or subsequent delivery not seperately reported	27	x	
	p	no return/missing registration to administration	8	x	
Order entry	cp/s	prices in ordering system not up to date	4	4	
	adm	data entry errors	7	x	human error
	p/adm	packing unit incorrect	2	2	
	adm	no separate invoice line for special project agreements	2	x	not product rel.
	adm	invoicing agreements incorrectly processed in order	3	x	order related
	adm	wrong creditor number entered	9	x	order related
	adm	order(line) processed too late	5	x	
Supplier	s	sent invoice without previous approval	21	x	After ordering process
	s	incorrect price/ number of products	9	x	supplier errors, not DQ errors of database
	s	not according to pricing conditions	16	x	
	p/s	invoice without order number	15	x	
Invoice entry	fd	data entry errors	18	x	invoice errors
	fd	incorrect creditor number	2	x	invoice errors
Other	s	more ordernumbers on one invoice	1	x	supplier related
	fz	accepted without ordernumber	5	x	order related
	s	dubbel shipping error	2	x	supplier error
	s/fz	name change creditor	4	x	not product rel.
	s	wrong ordernumber	6	x	not product related
Total			**335**	**48**	

% DQ errors	14%

Legenda:

pr	=project	cp	= central purchasing
adm	= administration	s	= supplier
fd	= financial department		

Appendix C

In this appendix we will clarify the terms in the object function of the Data Alignment through Logistics (DAL) method. Since there are 72 possible scenarios, we could start with defining 72 decision parameters X_1 to X_{72}, and then describe the remaining cost function per scenario. However, we know that the object function depends on the 5 decision parameters g, h, I, j and k. Therefore, we would like to develop an object function based on only these 5 parameters.

If we create an object function with only the five decision parameters, it would mean that the object function consists of the terms of the five cost functions where for each term the combination of decision parameters determines if the term is included yes or no. To do this, it is important to determine per cost function, namely $C_{transappl}$, $C_{checkappl}$, C_{test}, $C_{profappl}$, C_{adm}, C_{ass}, C_{check} and C_{prof}, on which decision parameters they depend. From Section 9.3 we know that all cost functions depend on 1, 2 or maximally 3 parameters where each decision parameter consists of 2 or maximally 3 values. Next, we have to determine a formula, which makes sure that only those cost terms that are relevant for a certain scenario are included. For instance, the cost of creating a translation application ($C_{transappl}$) depends on one decision parameter: where administration (g). Since g has only two values (0 and 1), which means centralized or local translation, the cost term for centralized translation ($= 2 \cdot S \cdot anr_s \cdot c_{transappl}$) should be selected in case g=1, and $C_{transappl} = (R+S) \cdot c_{transappl}$ in case g=0.

More generally, if we have a cost function that depends on only one decision parameter x with possible values {0, 1}, then there are two possible cost terms A, and B. Our objective is to define a formula that should return 1 for cost term A and 0 for cost term B in case of x=1. In case of x=0, the formula should return 0 for cost term A and 1 for cost term B. A formula with this property is shown in the table below.

Appendix C

x	x	(1-x)
1	1	0
0	0	1

The formula in this table is applicable for the cost of developing an online translation application ($C_{transappl}$) and the cost of administration (C_{adm}).

For a decision parameter z, consisting of 3 values, we developed the following formula:

z	1-z	z	z-1
0	1	0	-1
1	0	1	0
2	-1	2	1

This formula is applicable for the cost of building a profile maintenance application ($C_{profappl}$), and the cost of online profile maintenance (C_{prof}). As we can see from the table, the formula also returns –1, and 2 as possible values. This is no problem, when we arrange that the outcome of –1, and 2 are excluded. This can be done through a constraint, which determines that all formula terms should lie between 0 and 1.

For two decision parameters x, y with each two values {0,1}, we developed the following formula:

x	y	1-x-y	x-y	y-x	x+y-1
0	0	1	0	0	-1
1	0	0	1	-1	0
0	1	0	-1	1	0
1	1	-1	0	0	1

This formula is applicable for the cost of testing, ($C_{testing}$), and the cost of assigning C_{ass}).

For three decision parameters, 2 consisting of two values {0,1} and 1 of 3 values {0,1,2}, we have developed the following formula:

g	h	i	0,0,0	0,0,1	0,0,2	0,1,0	0,1,1	0,1,2	1,0,0	1,0,1	1,0,2	1,1,0	1,1,1	1,1,2
x	y	z	1-x-y-z	2x-4y+z	z-x-y-1	y-x-z	y-+2z-2	y-x+z-2	x-y-z	x-3y+2z-2	x-y+z-2	x+y-z-1	x+y+3z-4	x+y+z-3
0	0	0	1	0	-1	0	-2	-2	0	-2	-2	-1	-4	-3
0	0	1	0	1	0	-1	0	-1	-1	0	-1	-2	-1	-2
0	0	2	-1	2	1	-2	2	0	-2	2	0	-3	2	-1
0	1	0	0	-4	-2	1	-1	-1	-1	-5	-3	0	-3	-2
0	1	1	-1	-3	-1	0	1	0	-2	-3	-2	-1	0	-1
0	1	2	-2	-2	0	-1	3	1	-3	-1	-1	-2	3	0
1	0	0	0	2	-2	-1	-5	-3	1	-1	-1	0	-3	-2
1	0	1	-1	3	-1	-2	-3	-2	0	1	0	-1	0	-1
1	0	2	-2	4	0	-3	-1	-1	-1	3	1	-2	3	0
1	1	0	-1	-2	-3	0	-4	-2	0	-4	-2	1	-2	-1
1	1	1	-2	-1	-2	-1	-2	-1	-1	-2	-1	0	1	0
1	1	2	-3	0	-1	-2	0	0	-2	0	0	-1	4	1

This formula is applicable for the cost of building a checkapplication ($C_{checkappl}$), and the cost of online checking (C_{check}).

Appendix C

Bibliography

AC.1 (1996) *Interim Report,* TRADE/WP.4/GE.1/R.1189

AKEN J.E. VAN (1994:1) "Bedrijfskunde als Ontwerpwetenschap: de Regulatieve en Reflectieve circel", *Bedrijfskunde* (66:1) pp. 16-26

AKEN J.E. VAN (1994:2) "Developing scientific knowledge for Management Professionals from a Players' Perspective: The Role of Design Models and Heuristics", *Management & Organisatie* (4) pp. 388-404

ALEXANDER C. (1979) *The Timeless Way of Building,* Oxford University Press, New York

ANDERSEN N, KENSING F., LUNDIN J., MATHIASSEN L., MUNK-MADSEN A., RASBECH M., and SORGAARD P. (1990) *Professional Systems Development: Experience, Ideas and Action,* Prentice Hall International, Hemel Hempstead

ANVARI M. (1992) "Electronic Data Interchange and Inventories", *International Journal of Production Economics*, vol. 26, pp. 135-143

BACHARACH S.B. (1989) "Organizational Theories: Some Criteria for Evaluation", *Academy of Management Review,* (14:4) pp. 496-515

BAKKENIST MANAGEMENT CONSULTANTS (1996) *Haalbaarheidsonderzoek CBL classificatiesysteem,* EAN Nederland, Amsterdam, The Netherlands

BATINI C., LENZERINI M. and NAVATHE S.B. (1986) A Comparative Analysis of Methodologies for Database Schema Integration, *ACM Computing Surveys,* (18:4) pp. 323-364

BATINI C., CERI S. and NAVATHE S.B. (1992) *Conceptual Database Design: An Entity-Relationship Approach,* The Benjamin/Cummings Publishing Company, Redwood City, CA

BENBASAT I. and WEBER R. (1996) "Research Commentary: Rethinking "Diversity" in Information Systems Research", *Information Systems Research* (7:4) pp. 389-399

BENJAMIN R.I., D.W. DE LONG and SCOTT MORTON M.S. (1990) "Electronic Data Interchange: How Much Competitive Advantage?", *Long Range Planning* (23:1), pp. 29-40

BERTRAND J.W.M. and WORTMANN J.C. (1990) *Productiebeheersing en Material Management*, Stenfert Kroese, Leiden The Netherlands

BOLAND R. (1985) "Phenomenology: A preferred Approach to Research in Information Systems", In: *Research Methods in Information Systems,* Mumford E., Hirschheim R.A. and Wood-Harper A.T. (eds), NorthHolland Publishers, Amsterdam pp. 193-201

BOLAND R. (1991) "Information System Use as a Hermeneutic Process", in: *Information Systems Research: Contemporary Approaches and Emergent Traditions,* Nissen H.E., Klein H.K. and Hirschheim R.A. (eds), North-Holland, Amsterdam, pp. 439-464

BOOCH G. (1991) *Object Oriented Design: with applications,* Benjamin Cummings, Amsterdam

BROWNE J., SACKET P.J. and WORTMANN J.C. (1995) "Future Manufacturing Systems: Towards the Extended Enterprise"*, Computers in Industry*, (25:3) pp. 235-254

BRYNJOLFSSON E. and HITT L.M. (1998) "Beyond the Productivity Paradox", *Communications of the ACM*, (41:8), pp. 49-55

BSR PROJ (1995) *BSR Workshop on Rules and Guidelines for the Specification of BSUs and Bridges: Draft Workshop Output Document,* BSR PROJ N0021 rev.1

CARTER J.R. (1990) "The Dollars and Sense of Electronic Data Interchange", *Production & Inventory Management Journal*, 31:2, pp. 22-26

CASH J.I. and KONSYNSKI B.R. (1985) "IS redraws competitive boundaries", Harvard Business Review (64:2), march-april 1985, pp. 134-142

CASH J.I. and LAWRENCE P.R. (1989) *The Information Systems Research Challenge: Qualitative Research Methods,* Harvard Business School, Boston MA

CHANDRA A., KROVI R. and RAJAGOPOLAN B. (1998) "Flow Parameters and Quality in Accounting Information Systems", In: *Proceedings of the 1998 conference on IQ*, Chengular-Smith I, Pipino L.L. (eds) Cambridge MA pp. 194 - 201

CHECKLAND P. (1981) *Systems Thinking, Systems Practice,* Wiley, Chichester

CHUA W.F. (1986) "Radical Developments in Accounting Thought", *The Accounting Review,* (61) pp. 601-632

CHURCHMAN C.W. (1971) *The Design of Inquiring Systems: Basic Concepts of Systems and Organization,* Basic Books, New York

CLARK T.H. and STODDARD D.B. (1996) "Interorganizational Business Process Redesign: Merging Technological and Process Innovation", *Journal of Management Information Systems* (13:2) pp. 9-28

COENJAERTS M. and VERMEER B.H.P.J. (1995) *De Groothandel is dood, Lang leve de Groothandel!,* Report EUT/BDK/70, University of Technology Eindhoven, Eindhoven, The Netherlands

COMMITTEE ON AUDITING PROCEDURE (1949) American Institute of Accountants, *Internal Control*, American Institute of Certified Public Accountants, New York

CUSHING B.E. (1982) *Accounting Information Systems and Business Organizations,* Addison-Wesley, Reading MA

DATE (1990) *An Introduction to Database Systems – Volume 1*, 5th edition, Addison-Wesley Publishing Company, Reading MA

DAVENPORT T.H. (1993) *Process Innovation, Reengineering work through Information Technology*, Harvard Business Press, Boston

DEMING (1986) *Out of the Crisis.* Cambridge center for advanced engineering study, MIT Cambridge, MA

DORST K. (1994) *Describing Design: a Comparison of Paradigms,* Thesis University of Technology Delft, Delft, The Netherlands

EEMA/EDI working group (1998) *Setting up a UN Repository for XML-EDI,* CEFACT/TMWG, N071, http://www.harbinger.com/resource/klaus/tmwg/documentlist.html, see also: http://www.eema.org/

EISENHARDT K.M. (1989) "Building Theories from Case Study Research", *Academy of Management Review* (14:4) pp. 532-550

ELMASRI R. and NAVATHE S.B. (1994) *Fundamentals of Database Systems,* Benjamin Cummings Publishing Company, Redwood City, CA

EMMELHAINZ M.A. (1993) *EDI: A Total Management Guide*, Van Nostrand Reinhold, USA

ERIKSSON H-E, and M. PENKER (1998) *UML Toolkit,* Wiley Computer Publishing, New York

FINANCIAL ACCOUNTING STANDARDS BOARD (FASB) (1980) "a Hierarchy of Accounting Qualities", published in: Statement of Financial Accounting, Concepts No. 2*, Qualitative Characteristics of Accounting Information*

FOWLER M. and SCOTT K. (1999) *UML Distilled,* Addison-Wesley Object Technology Series, Reading (MA)

FORESTER J. (1992) "Critical Ethnography: On Field Work in an Habermasian Way", in: *Critical Management Studies,* Alvesson M. and Willmott H. (eds.) Sage Publications, London, pp. 46-65

FOX C., LEVITIN A. and REDMAN T. (1993) "The Notion of Data and Its Quality Dimensions", *Information Processing & Management* (30,1), pp. 9-19

GADAMER H-G (1976) *Philosophical Hermeneutics,* University of California Press, Berkely, CA

GALBRAITH J. (1973) *Designing Complex Organizations,* Addison-Wesley, Reading MA

GLASER B. and STRAUS A. (1967) *The Discovery of Grounded Theory: Strategies of Qualitative Research,* Wiedenfeld & Nicholson, London

GOH C.H., MADNICK S.E. and SIEGEL M.D. (1994) Context Interchange: Overcoming the challenges of large-scale interoperable database systems in a dynamic environment. *in Proc. 3rd International Conference on Information and Knowledge Management (CIKM-94),* Gaithersburg, Md

GOH C.H., BRESSAN S., MADNICK S.E. and SIEGEL M.D. (1999) Context Interchange: New Features and Formalisms for the Intelligent Integration of Information *ACM Transactions on Information Systems*, Forthcoming July 1999

GOLDRATT E.M. (1997) *Critical Chain,* North River Press, Great Barrington

GOODHUE D.L., WYBO M.D. and KIRSCH L.J. (1992) The impact of data integration on the costs and benefits of information systems. *MIS Quarterly,* (16:3), pp. 293-311

GOOR A.R. van, PLOOS van AMSTEL M.J. and PLOOS van AMSTEL W. (1989) *Fysieke Distributie, Denken in Toegevoegde Waarde*, Stenfert Kroese, Leiden The Netherlands

GOOSENAERTS J. (1999) *Web oriented Modeling in XML*, lecture notes University of Technology Eindhoven, Section technology management Information & Technology, Fall 1999

GUMMESSON E. (1991) *Qualitative Methods in Management Research,* Sage Publications, London

HABERMAS J. (1968) *Knowledge and Human Interests*, Beacon Press, Boston, (1971 transl. J.J. Shapiro)

HABERMAS J. (1984) *The Theory of Communicative Action: Reason and the Rationalization of Society,* Vol 1, Beacon Press, Boston, (1971 transl. T. McCarthy)

HAMILTON S. and IVES B. (1982) "MIS Research Strategies", *Information & Management,* Vol 5, pp. 339-347

HARTOG F. DEN and SLUIJS E. VAN (1995) *Onderzoek in Bedrijven: Een Methodologische Reisgids,* Van Gorcum & Comp B.V., Assen, The Netherlands

HEIMBIGNER D. and MCLEOD D. (1985) A Federated Architecture for Information Management, *ACM Transactions on Office Information Systems* (3:3) pp. 253-278

HIRSCHHEIM R. KLEIN H.K. and LYYTINEN K. (1995) *Information Systems Development and Data Modeling, Conceptual and Philosophical Foundations,* Cambridge University Press, Cambridge

HOOGEWEEGEN M.R. and WAGENAAR R. (1996) "A Method to Assess Expected Net Benefits of EDI Investments", *International Journal of Electronic Commerce (1:1), pp. 73-94*

HUSSERL E. (1982) *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy,* Kluwer, Boston

HUTCHINSON N.E. (1987) *An Integrated Approach to Logistics Management*, Prentice Hall, Englewood Cliffs, England

HUANG K. (1998) *Organizational Aspects of EDI: a Norm-oriented Approach,* Thesis Enschede, University of Enschede, The Netherlands

INSTALNET (1994) *Handleiding EANCOM Artikel Bericht voor de Installatie Sector, versie 2.01,* Stichting Instalnet, 's-Hertogenbosch

ISO (1996) *Open-EDI,* ISO/IEC JTC 1/SC30, CD 14662

IVES B., HAMILTON S., and DAVIS G. (1980) "A Framework for Research in Computer Based Management Information Systems", *Management Science* (26:9), pp. 910-934

JACOBSON I. (1995) *Object-Oriented Software Engineering: a Use Case Driven Approach,* ACM Press Books, New York

JANSEN M.H., VERMEER B.H.P.J. and JAGDEV H.S. (1997) "Towards a Typology of Electronic Product Information Distribution"*, Computers In Industry* vol. 33, pp. 395-409

JELASSI T. and O. FIGON (1994) "Competing through EDI at Brun Passot: Achievements in France and Ambitions for the Single European Market", *MIS Quarterly*, (18:4), pp. 337-352

JONES J.C. (1977) "How my Thoughts about Design Methods have Changed over the Years", In: Cross NG (ed) *Developments in Design Methodology,* Wiley, Chichester

JÖNSSON S. (1991) "Action Research", in: *Information Systems Research: Contemporary Approaches and Emergent Traditions,* Nissen H.E., Klein H.K. and Hirschheim R.A. (eds), North-Holland, Amsterdam, pp. 371-396

JURAN J.M., GRYNA F.M.J. and BINGHAM R.S. (1974) *Quality Control Handbook* (3rd ed.), McGraw-Hill Book Co, New York NY

KEKRE S. and T. MUKHOPADYAY (1992) "Impact of Electronic Data Interchange Technology on Quality Improvement and Inventory Reduction Programs: A field Study", *International Journal of Production Economics* (28:3), December, pp. 265-282

KLOOSTER, A.J. van 't, and OONINCX J.A.M. (1978) *Leerboek Informatiesystemen,* Samson, Alphen aan den Rijn

KLEIN H.K. and MYERS M.D. (1999) "A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems", *MIS Quarterly (23:1)* pp. 67-91

KLEIN H.K. and Truex III, D.P (1995) "Discourse Analysis: A Semiotic Approach to the Investigation of Organizational Emergence" in: *The Semiotics of the Workplace*, P.B. Andersen and B. Holmqvist (eds.), Walter De Gruyter, Berlin

KRALJIC, P. (1983) "Purchasing must become Supply Management", *Harvard Business Review,* Sept-okt, pp. 109-117

KUIJER M. (1991) "Technische Automatisering is van Groot Belang voor de ET Sector", *Elektro Magazine Installatie*
KUIJER M. (1992) "Technische Automatisering ET Sector in Implementatie Fase", *Elektro Magazine Installatie*
KURT SALMON ASSOCIATES (1993) *Efficient Consumer Response: Enhancing Customer Value in the Grocery Industry*, Food Marketing Institute, Washington, DC
LAUDON K. (1986) "Data Quality and Due Process in Large Interorganizational Record Systems", *Communications of the ACM,* (29), pp. 4-11
LEE A.S. (1989) "A Scientific Methodology for MIS Case Studies", *MIS Quarterly* (13:1), pp.33-50
LEE A.S. (1994) "Electronic Mail as a Medium for Rich Communication: An Empirical Investigation Using Hermeneutic Interpretation", *MIS Quarterly* (18:2), pp.143-157
LITWIN W., MARK L. and ROUSSOPOULOS N. (1990) Interoperability of Multiple Autonomous Databases, *ACM Computing Surveys* (22:3) PP. 267-293
LOCKHEED MARTIN ADVANCED CONCEPTS CENTER and RATIONAL SOFTWARE CORPORATION (1996) *Succeeding with the Booch and OMT methods: a Practical Approach,* Addison Wesley, Amsterdam
LYYTINEN K. and KLEIN H. (1985) "The Critical Social Theory of Jürgen Habermas (CST) as a Basic for a Theory of Information Systems", in: *Research Methods in Information Systems,* Mumford E., Hirschheim R.A. and Wood-Harper A.T. (eds), NorthHolland Publishers, Amsterdam pp. 219-232
MACKAY D.R. (1993) "The Impact of EDI on the Components Sector of the Australian Automotive Industry*", Journal of Strategic Information Systems*, 2:3, pp. 243-263
MADNICK S.E. (1995) "Integrating Information From Global Systems: Dealing With the "On- and Off Ramps" of the Information Superhighway", *Journal of Organizational Computing,* (5,2), pp. 69-82
MARKUS M.L. (1994) "Electronic Mail as the Medium of Managerial Choice", *Organization Science (5:4),* pp. 502-527
MCCUSKER T. (1994) "How to Get More Value from EDI", *Datamation* (40:9), May, pp. 56-60
MILES M.B. and HUBERMAN A.M. (1984) *Qualitative Data Analysis: A Sourcebook of New Methods,* Sage Publications, Newbury Park
MILES R.E. and SNOW C.C. (1992) "Causes of Failure in Network Organizations", *California Management Review,* Summer 1992, pp. 53-72
MONTFORT L. van. (1997:1) *Herontwerp van het Orderproces van Croon B.V.* Graduation Report University of Eindhoven, Eindhoven
MONTFORT L. van. (1997:2) *Herontwerp van het Orderproces van Croon B.V.: Bijlage*, Graduation Report University of Eindhoven, Eindhoven
MOURITS M.and EVERS J.J.M.(1995) "Distribution Network Design, an Integrated Planning Support Framework", *International Journal of Physical Distribution & Logistics Management*, Vol. 25 Np. 5., pp. 43-57
MUKHOPADHYAY T., KEKRE SW. and KALATHUR S. (1995) "Business Value of Information Technology: a Study of Electronic Data Interchange", *MIS Quarterly*, (19:2), pp. 137-155
MYERS M.D. (1994) "A Disaster for Everyone to See: An Interpretive Analysis of a Failed IS project", *Accounting, Management and Information Technologies* (4:4) pp. 185-201
MYERS M.D. (1997) "Qualitative Research in Information Systems", *MIS Quarterly* (21:2) pp. 241-242. MISQ Discovery, archival version, June 1997,

http://www.misq.org/misqd961/isworld. MISQ Discovery, updated version, Februari 22, 1999, http://www.auckland.ac.nz/msis/isworld/

NGWENYAMA O.K. (1991) "The Critical Social Theory Approach to Information Systems: Problems and Challenges", in: *Information Systems Research: Contemporary Approaches and Emergent Traditions,* Nissen H.E., Klein H.K. and Hirschheim R.A. (eds), North-Holland, Amsterdam, pp. 267-280

NGWENYAMA O.K. and LEE A.S. (1997) "Communication Richness in Electronic Mail: Critical Social Theory and the Contextuality of Meaning", *MIS Quarterly (21:2)* pp. 145-167

NISSEN H.E., KLEIN H.K. and HIRSCHHEIM R.A. (eds) (1991) *Information Systems Research: Contemporary Approaches and Emergent Traditions,* North-Holland, Amsterdam

ORLIKOWSKI W.J. and BAROUDI J.J. (1991) "Studying Information Technology in Organizations: Research Approaches and Assumptions", *Information Systems Reseearch* (2:1) pp. 1-28

OWEN J. (1993) *STEP: An Introduction,* Information Geometers, Winchester, UK

PEIRCE C.S. (1960), *Collected Papers,* Hartborne C. and Weiss P. (eds), Belknap Press of Harvard University Press, Cambridge, 1931-1935

PELS H.J. (1988) *Geïntegreerde Informatiebases, Modulair Ontwerp van het Conceptuele Schema*. Doctoral Thesis, H.E. Stenfert Kroese B.V., Leiden-Antwerpen, The Netherlands

PERCY T. (1986) "My Data, Right or Wrong", *Datamation* (32,11), pp. 123-128

PIJL, G.J. van der (1993) *Kwaliteit van Informatie, In theorie en Praktijk,* Dissertation, Katholieke Universiteit Brabant, Tilburg

REDMAN T.C. (1995) "Improve Data Quality for Competitive Advantage", *Sloan Management Review,* Winter, pp. 99-106

REEKERS N. and S. SMITHSON (1996) "The Role of EDI in Interorganizational Coordination in the European Automotive Industry", *European Journal of Information Systems* (5), 1996, pp. 120-130.

REIJSWOUD V. VAN (1996). *The Structure of Business Communication: Theory, Method, and Application*, Delft University, Ph.D. Thesis

RIBBERS P.M. (1995) "Purchasing Through EDI- The Case of Technische Unie in The Netherlands", in: *EDI in Europe*, H. Krcmar, N. Bjorn-Andersen and R. O'Callaghan (eds), John Wiley & Sons, Chichester

RICARDO C. (1990) *Database Systems: Principles, Design, and Implementation,* MacMillan Publishing Company, New York NY

RIGGINS T. and T. MUKHOPADHYAY (1994) "Interdependent Benefits from Interorganizational Systems: Opportunities for Business Partner Reengineering"*, Journal of Management Information Systems*, (11:2), pp. 37-57

ROBEY D. (1996) "Research Commentary: Diversity in Information Systems Research: Threat, Promise, and Responsibility", *Information Systems Research* (7:4) pp. 400-408

SCHÖN D.A. (1983) *The Reflective Practitioner,* Basic Books, New York

SCHÖN D.A. (1987) *Educating the Reflective Practitioner,* Basic Books, New York

SHANNON C.E. and WEAVER W. (1949) *The Mathematical Theory of Communication,* University of Illinois Press, Urbana

SHETH A.P. and LARSON J.A. (1990) Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, *ACM Computing Surveys, (22:3*) pp. 183-236

SIMON H.A. (1967, 1996) *The Sciences of the Artificial,* The MIT Press, Cambridge MA

SLIM-WORKING GROUP (1997) *Van Zand tot Klant: Ketenlogistiek werkt!* CIP-DATA Library University of Eindhoven, The Netherlands

STAMPER R.K. (1992) "Signs, Organisations, Norms and and Information Systems", in: *Proceedings of the third Australian Conference on Information Systems*, University of Wollongong, Australia

STARREVELD R.W., DE MARE H.B. and JOËLS E.J. (1976) *Bestuurlijke Informatieverzoring, Deel 1: Algemene Grondslagen,* Samson Bedrijfsinformatie, Alphen aan den Rijn, (first print)

STARREVELD R.W., DE MARE H.B. and JOËLS E.J. (1994) *Bestuurlijke Informatieverzoring, Deel 1: Algemene Grondslagen,* Samson Bedrijfsinformatie, Alphen aan den Rijn, (third print)

STEEL K. (1996:1) The users guide to the BSR, *Standards Australia,* University of Melbourne, Australia

STEEL K. (1996:2) *The Basic Semantic Repository,* http://www.cs.mu.oz.au/research/icaris/bsr.html

STRAUS A. and CORBIN J. (1990) *Basics of Qualitative Research: Grounded Theory Procedures and Techniques,* Sage Publications, Newbury Park

STRONG D.M. (1997) "IT Process Designs for Improving Information Quality and Reducing Exception Handling: A Simulation Experiment", *Information & Management,* vol 31, pp. 251-263

STRONG D.M., LEE Y.W. and WANG R.Y. (1997) "Data Quality in Context", *Communications of the ACM*, (40,5), pp. 103-110

SUH N.P. (1990) *The Principles of Design* Oxford University Press, New York

TANENBAUM A.S. (1986) *Computer Networks,* Prentice Hall International, London, 1989

TAYLOR R. (1986) *Value Added Processes in Information Systems*, Albex Publishing, New York

TAYLOR D.A. (1998) *Object-Oriented Technology: A Manager's Guide*, 2nd edition, Addison-Wesley Publishing Company, Amsterdam

TAYLOR S.J. and BOGDAN R. (1984) *Introduction to Qualitative Research Methods,* John Wiley, New York

TMWG (1998:1) *Reference Guide: The Next Generation of UN/EDIFACT - An Open-EDI approach using UML models and OOT Revision 12,* Techniques and Methodologies Working Group N010R1, http://www.harbinger.com/resource/klaus/tmwg/documentlist.html

TMWG (1998:2) *Explanation of TMWG Decision to use Unified Modeling Language (UML) for Business and Information Modeling,* Techniques and Methodologies Working Group N065R1, http://www.harbinger.com/resource/klaus/tmwg/documentlist.html

TMWG (1998:3) *An Open-EDI Prototype Based on UML, CORBA and Java,* Techniques and Methodologies Working Group N069, http://www.harbinger.com/resource/klaus/tmwg/documentlist.html

TMWG/XML taskgroup (1999) *TMWG's position on XML, recommendations to UN/CEFACT*, Techniques and Methodologies Working Group N089R1, http://www.harbinger.com/resource/klaus/tmwg/documentlist.html

TUSHMAN M. and NADLER D. (1978) Information processing as an integrating concept in organizational design. *Academy of Management Review,* (3:3) pp. 613-624

UN/ECE (1996) *UN/EDIFACT: A Strategy for the Next Phase,* TRADE/WP.4/CRP.123/ Appendix 2

UNETO (1993) *Haalbaarheidsonderzoek Centrale Artikel Database,* Uneto, Zoetermeer The Netherlands
UNETO (1994) *Centrale Artikel Database voor de Electrotechnische Sector,* Brochure Uneto, Zoetermeer The Netherlands
VENKATRAMAN N. (1994) "IT-enabled Business Transformation: from Automation to Business Scope Redefinition", *Sloan Management Review,* Winter 1994, pp. 73-87
VERMEER B.H.P.J. (1995) *Aantekeningen Begotel Interview,* University of Eindhoven
VERMEER B.H.P.J. (1996:1) *Het ene Artikel is het Andere Niet! Een onderzoek naar de problemen met slechte datakwaliteit van artikelgegevens in de Levensmiddelensector*, Report EUT/BDK/77, Eindhoven 1996, The Netherlands
VERMEER B.H.P.J. (1996:2) *Het Informatie Distributie Centrum: Een distributiemechanisme voor basis artikelinformatie in de pharmaceutische sector*, Neprofarm, Utrecht, The Netherlands (In Dutch)
VERMEER B.H.P.J. (1998:1*) Impact van Gegevenskwaliteit op het Orderprocess bij E. Schwartz B.V.,*case study report, Eindhoven 1996, The Netherlands (In Dutch)
VERMEER B.H.P.J. (1998:2) *Data Alignment: van Papier naar Bilateraal of Centraal?*, Final Report of the EAN-DAS case, University of Technology, Eindhoven (In Dutch)
VERWIJMEREN A.A.P. (1998) *Networked Inventory Management by Distributed Object Technology,*Thesis, Technische Universiteit Eindhoven, Eindhoven, the Netherlands
VETH A.F.L. (1996) "Ketens, Logistiek in het kwadraat", *VLO-Magazine,* (96:1), pp. 14-17
VETH A.F.L. (2000) *E-motion in E-business; de praktijkervaringen met e-commerce in - Business to business- relaties*. CEBRA, H&J publishers and AS Marketing Network, Eindhoven (In Dutch)
VLIST P. van der (1991) *EDI in de Handel,* Samson Bedrijfsinformatie, Alphen aan den Rijn (In Dutch)
VREEDE G.J. (1995) *Facilitating Organizational Change: The Participative Application of Dynamic Modeling,* Thesis University of Delft, Delft, The Netherlands
WALSHAM G. (1993) *Interpreting Information Systems in Organizations.* Wiley, Chichester
WALSHAM G. (1995) "Interpretive Case Studies in IS Research: Nature and Method", *European Journal of Information Systems (4:2),* pp. 74-81
WAND Y. and WANG R. (1996) "Anchoring Data Quality Dimensions in Ontological Foundations", *Communications of the ACM* (39), pp. 86-95
WANG R.Y.and STRONG D.M. (1996) "Beyond Data Accuracy: What Data Quality Means to Data Consumers", *Journal of Management Information Systems*, (12:4), pp. 5-34
WEBER R. (1999) *Information Systems Control and Audit*, Prentice Hall, Upper Saddle River
WEELE A.J. VAN (1992) *Supply Management:Four Basic Strategies,* Series: Developments and ideas
WETTEN D.A. (1989) "What Constitutes a Theoretical Contribution?", *Academy of Management Review,* (14:4) pp. 490-495
WEIGAND H., MOOR A. de, HEUVEL W.J. van den (2000) "Supporting the Evolution of Workflow Patterns for Virtual Communities", in: Proceedings of the 33rd annual Hawaii International Conference on System Sciences, IEEE Computer Society Press
WINOGRAD T. and FLORES F. (1987) *Understanding Computers and Cognition: A New Foundation for Design,* Addison-Wesley Publishing Co. Inc., New York
WRIGHLY C.D., R.W. WAGENAAR, and R. CLARKE (1994) "Electronic Data Interchange in the International Trade: Frameworks for the Strategic Analysis of Ocean Port Communities", *Journal of Strategic Information Systems*, 3:3, pp. 211-234

YIN R.K. (1984, 1994) *Case Study Research: Design and Methods,* Sage Publications Thousand Beverly Hills CA

ZMUD R. (1978) "An Empirical Investigation of the Dimensionality of the Concept of Information", *Decision Sciences* (9), pp. 187-195

Bibliography

Summary

Introduction

Integrating computer systems in business chains is a difficult task. Already since the seventies (1970) companies are attempting to integrate their computer systems using Electronic Data Interchange (EDI), with varying success. Today, E-business is promising to change all that, because web technology enables companies to present their company information to users all over the world. However, there is difference between presenting information and the ability to act on it. Since the human interpretation mechanism is very sophisticated, it is relatively easy for humans to understand and use the presented information, which might explain the recent success of web technology in Business-to-Business (B2B) networks. Computer systems do not have a sophisticated interpretation mechanism. Therefore, when we integrate computer systems, we must be very precise about the meaning of the presented data. In other words, the data must have an exceptional high degree of quality. Otherwise, errors in the communication will occur. To achieve this high degree of data quality, it is necessary to align the data between computer systems. That is the central issue of this thesis.

Problem statement

When computer systems in business chains are interconnected using EDI, an increasing number of data quality problems occur. In a field study in the Dutch food sector, we found the following examples: orders with outdated product numbers, missing numbers in product lookup tables, reception of different goods than ordered and price differences in invoices.

To solve these problems we asked ourselves the following three questions:

1. What causes data quality problems, when computer systems are interconnected using EDI?
2. What is the key to solve these data quality problems?
3. How should we construct a method to do this?

What causes data quality problems when EDI is introduced?

To understand data quality, and how it is affected in electronic communication, we studied the data quality and communication literature. From this literature, we found three views on data quality:

- The reliability view, which defines data quality as conformance to reality;
- The relevance view, which defines data quality as conformance to the user (= 'fitness for use');
- The context view, which defines data quality as conformance to the shared context of the communication process. Here, the context of the communication process is defined as the shared norm system of a sender and a receiver that is used to interpret the information in a message. Specifically, this norm system consists of shared rules, definitions and data.

Using these three views, we are able to construct a new model that explains the relationship between an EDI initiative and data quality. According to this model, there is a direct and an indirect relationship between EDI and data quality. The direct relation between EDI and data quality is that EDI increases the data quality of an EDI message, because it increases both the reliability and relevance of the information in the message. The indirect relation between EDI and data quality is explained as follows. An EDI initiative results in higher requirements for the quality of the context between sender and receiver. Since EDI implementation eliminates the human interpretation mechanism in the transaction communication between business processes, this means that the quality of the context must be improved. If the realized context quality matches the required context quality, then the high quality of the context positively influences the data quality of the EDI message, and hence has a positive effect on the business process. However, when the context quality is not improved when EDI is implemented, then the required context quality exceeds the realized context quality. Since context quality positively influences the data quality of the EDI message, the insufficient realized quality of the context will result in a decreased quality of the data in the EDI message, and hence will negatively influence the business process.

Using this indirect view, we are able to explain the data quality problems in the EDI messages, when computer systems are interconnected:

> *If EDI is introduced, and the quality of the context is not improved, then this context quality will decrease the direct positive effects of EDI on the business process.*

We tested this proposition in two cases studies, where this relationship was confirmed.

What is the key to solve these problems?

As we can see from our proposition, we have to improve the quality of the context to solve the data quality problems between interconnected computer systems. To know how to do this, we need to examine the communication process between interconnected computer systems more closely. When computer systems are interconnected, they exchange orders, delivery notifications, invoices, and many more messages, which we refer to as *transaction communicatio*n. The messages in the transaction communication contain all kinds of references to other information in the context of the communication process, such as product information, pricing schemas, delivery locations, and more. These facts are stored in the respective databases of sender and receiver. When changes in these facts occur, the sender will send a fact

update to the receiver, to notify the receiver of that change. We will refer to this type of communication, which is independent of the transaction, as ***data alignment***. Since the exchange of fact updates will improve the quality of the information in the databases of sender and receiver, the quality of the context is improved. Thus, data alignment is the key to improve the quality of the context.

How should we construct a method to do this?

The final question is, how we should construct a data alignment method to improve the quality of the context, and therefore, the quality of the transaction communication.

Basically, two problems arise when data is aligned in multiple database situations: the translation problem, and the distribution problem. The *translation* problem arises because the same fact may be differently structured at different locations in the network. Therefore, database schema translation is necessary to map the structure of the source schema to the structure of the manufacturer's schema. This results in a mapping schema between the source schema and the receiver's schema that is used every time a fact in the source database is updated. The *distribution* problem arises because each fact update is first translated and then transported over a network to a limited set of users, where it is finally interpreted and stored in the receiver's database. During translation and interpretation, mapping errors may occur, which results in loss of data quality. During transportation, the data may get delayed, damaged, or delivered to the wrong recipient, resulting in inconsistencies among different locations.

Studying the data integration and data distribution literature, we constructed a new method to align the data in large interorganizational networks, which we named the Data Alignment through Logistics method (DAL method). Realizing that the real challenge in data alignment is to distribute data from a source database to many receivers while maintaining a certain quality level, the idea behind this method is to apply principles from the field of logistics to the data alignment problem.

How DAL solves the translation problem

In an evaluation of several data integration and distribution methods, we found that the best way to solve the *translation* problem in large interorganizational situations is the definition of a central schema per specific user group, where local users define their own mappings between local and central schema. Complete data integration of all database schemas in the network is not feasible, since this is very labor intensive, and ultimately results in a very inflexible integrated schema. The central schema with local mappings approach provides a lot of flexibility, since each local user decides for himself what information is relevant to map to the central schema. However, this comes at the cost of having to check for data quality, because all local users make their own interpretation of the central schema, which results in mapping errors.

Therefore, in the DAL method, the translation problem is solved through the definition of an Information Product (IP). An IP is a central schema, which describes all definitions and agreements for a specific *application* in a specific user community. Viewing the central schema as a product, we provide focus in the definitions, so that a minimal set of mutual agreements is established. An example for the food retail community is the *Product Master data* IP, which describes the product hierarchy (how do consumer units relate to trade units and to transport units). This IP is used in logistical applications.

How DAL solves the distribution problem
From the evaluation of data integration and distribution approaches we found that no effective distribution mechanism between multiple, heterogeneous databases existed. Therefore, we modeled the distribution part of the DAL method after physical distribution networks. From logistics theory, we learned that physical distribution networks basically consist of two parts: a static part that determines the structure of the network, and a dynamic part, that controls the goods flows in the network. Normally, the complexity of designing physical distribution networks is reduced through modeling the static and control structure separately. We used this principle from logistics to design a fact update distribution network. In two case studies, where we applied and improved the DAL method, we learned that the static structure of a fact update network depends on the placement of the following five functions:

- The administration function, which is concerned with administering the community schema.
- The assignment function, which is concerned with entering new facts in a database.
- The checking function, which is concerned with performing manual sample checks and defining business rules for a checking application to check for mapping errors, and assignment errors.
- The testing function, which is concerned with testing a message link.
- The profile maintenance function, which is concerned with administering which information receivers would like to receive.

In the DAL method, the structure of the fact update distribution network is the result of deciding where to locate each of the five functions: centrally in the network, or locally at each bilateral link in the network.

Conclusion

In this thesis we studied how data quality problems can be prevented when computer systems are interconnected. Using the DAL method, both E-business agencies and individual companies in B2B networks have the means to align their data, thus preventing data quality problems in their operational business processes.

Samenvatting

Inleiding

Integreratie van computersystemen in bedrijfsketens is niet eenvoudig. Al sinds de zeventiger jaren proberen bedrijven met EDI hun computersystemen te integreren, met wisselend success. E-business is de belofte die dit waar moet maken, doordat bedrijven via het web in staat zijn om hun gegevens aan gebruikers over de gehele wereld te presenteren. Echter, er is een verschil tussen het presenteren van informatie en het gebruik kunnen maken van die informatie. Vanwege de geraffineerdheid van het menselijke interpretatie mechanisme, is het voor mensen vrij eevouding om de gepresenteerde informatie te begrijpen en te gebruiken, wat een verklaring zou kunnen zijn voor het recente succes van web technologie in B2B netwerken. Echter, computer systemen beschikken niet over zo'n geraffineerd mechanisme. Daarom moeten we bij het integreren van computer systemen zeer precies vastleggen wat de gepresenteerde informatie betekent. Met andere woorden: de gegevens moeten een zeer hoge gegevenskwaliteit hebben. Anders ontstaan er fouten in de gegevensuitwisseling tussen de computersystemen. Om dit hoge kwaliteitsniveau te bereiken, moeten de gegevens tussen computersystemen op elkaar worden afgestemd. Dit is het centrale onderwerp in dit proefschrift.

Het probleem

Als we computersystemen in een bedrijfsketen met elkaar integreren met EDI ontstaan een aantal problemen met de gegevenskwaliteit. In een studie in de nederlandse voedingsmiddelensector vonden we de volgende typische voorbeelden: orders met verlopen productnummers, kassatabellen waarin productnummers ontbreken, levering van andere goederen dan oorspronkelijk waren besteld en prijsverschillen in facturen.

Om deze problemen op te lossen hebben we onszelf de volgende drie vragen gesteld:

1. Wat veroorzaakt de geconstateerde gegevenskwaliteitsproblemen die optreden bij het aansluiten van computersystemen met EDI?

2. Wat is de sleutel tot een oplossing voor deze problemen?
3. Hoe moet een methode eruit zien die dit probleem aanpakt?

Wat veroorzaakt de geconstateerde gegevenskwaliteitsproblemen?

Om te begrijpen wat gegevenskwaliteit is en hoe het wordt beïnvloed tijdens elektronische communicatie, hebben we de gegevenskwaliteit- en communicatieliteratuur bestudeerd. Hieruit hebben we drie manier om gegevenskwaliteit te beschouwen gedestilleerd:

- De juistheidsvisie, welke gegevenskwaliteit beschouwt als de mate van overeenstemming met de werkelijkheid;
- De relevantievisie, die gegevenskwaliteit beschouwt als geschiktheid voor gebruik;
- De contextvisie, die gegevenskwaliteit beschouwt als de mate van overeenstemming met de context van het communicatieproces. Hierin is de context van het communicatieproces gedefinieerd als het gezamenlijke normensysteem dat bestaat tussen een zender en ontvanger, en dat wordt gebruikt om de informatie in een bericht te interpreteren. Meer specifiek bestaat dit normensystemen uit gezamenlijke regels, definities en data.

Met behulp van deze drie verschillende visies op gegevenskwaliteit, hebben we een nieuw model kunnen maken, dat het verband tussen een EDI initiatief en gegevenskwaliteit verklaart. Volgens dit model is er een direct en een indirect verband tussen een EDI initiatief en gegevenskwaliteit. Het directe verband houdt in dat EDI de gegevenskwaliteit van een EDI bericht verbetert, omdat zowel de juistheid als de relevantie van de informatie in een EDI bericht is verhoogd. Het indirecte verband tussen EDI en gegevenskwaliteit luidt als volgt. Een EDI initiatief resulteert in hogere eisen ten aanzien van de kwaliteit van de context tussen zender en ontvanger. Aangezien EDI het menselijke interpretatiemechanisme in de transactiecommunicatie tussen de bedrijfsprocessen elimineert, heeft dit tot gevolg dat de kwaliteit van de context moet worden verhoogd. Indien de gerealiseerde contextkwaliteit overeenkomt met de vereiste contextkwaliteit, dan leidt de dan gerealiseerde hoge contextkwaliteit tot een positieve invloed op de gegevenskwaliteit van het EDI bericht, en dus tot een positieve invloed op het bedrijfsproces. Echter, indien de contextkwaliteit niet is verbeterd bij de invoering van EDI, dan leidt dit ertoe dat de vereiste contextkwaliteit de gerealiseerde contextkwaliteit overtreft. Omdat de contextkwaliteit een positieve invloed heeft op de gegevenskwaliteit in het EDI bericht, leidt dit in geval van een gebrekkige gerealiseerde contextkwaliteit tot een verlaging van de kwaliteit van het EDI bericht. Dit zal op zijn beurt leiden tot een negatieve invloed op het bedrijfsproces.

Door gebruik te maken van dit indirecte verband, zijn we in staat om de gegevenskwaliteitsproblemen in EDI berichten bij het op elkaar aansluiten van computersystemen te verklaren:

> *Als we EDI introduceren, en we verbeteren niet tegelijkertijd de kwaliteit van de context, dan leidt deze contextkwaliteit tot een verlaging van de positieve effecten van EDI op een bedrijfsproces.*

Deze propositie is in twee casestudies getest, en bevestigd.

Wat is de sleutel tot een oplossing voor deze problemen?

Volgens bovengenoemde propositie, zullen we dus de contextkwaliteit moeten verbeteren om de problemen met gegevenskwaliteit, die optreden bij het op elkaar aansluiten van computersystemen, op te lossen. Om te weten hoe we dit moeten doen, zullen we eerst het

communicatieproces tussen electronisch verbonden computersystemen beter moeten bekijken. Als computersystemen elektronisch met elkaar worden verbonden, wisselen ze normaliter orders, pakbonnen, facturen en nog vele andere berichten met elkaar uit. Dit type van communicatie noemen we transactiecommunicatie. De berichten in de transactiecommunicatie bevatten allerlei verwijzingen naar andere informatie in de context van het communicatieproces, zoals product informatie, prijsafspraakschema's, afleveradressen en nog meer. Deze *feiten* worden opgeslagen in de respectievelijke databases van de zender en de ontvanger. Telkens wanneer er veranderingen in deze feiten optreden, zal de zender deze nieuwe feiteninformatie opsturen naar de ontvanger, om de ontvanger hiervan op de hoogte te stellen. We zullen dit type van communiceren, dat plaatsvindt onafhankelijk van de transactie, aanduiden met gegevensafstemming. Omdat het uitwisselen van informatie over feiten de kwaliteit van de informatie in de respectievelijke databases van zender en ontvanger zal verbeteren, leidt dit tot een verbetering van de contextkwaliteit. Dus, gegevensafstemming is de sleutel tot het verbeteren van de contextkwaliteit.

Hoe moet een methode eruit zien die dit probleem aanpakt?

Uiteindelijk is de vraag hoe we een gegevensafstemmingsmethode moeten construeren, die de contextkwaliteit verbetert, en uiteindelijk dus de gegevenskwaliteit in de transactie-communicatie.

Er treden twee problemen op wanneer gegevens in multidatabase situaties op elkaar worden afgestemd: het vertaalprobleem, en het distributieprobleem. Het *vertaalprobleem* ontstaat doordat hetzelfde feit op verschillende locaties in het netwerk versschillend kan zijn gestructureerd. Daarom is database schemavertaling noodzakelijk om de structuur van het bronschema te mappen naar de structuur van het schema van de ontvanger. Dit resulteert in een mappingschema tussen het schema van de bron en het schema van de ontvanger, wat telkens opnieuw wordt gebruikt als een nieuw feit in de brondatabase wordt ingevoerd. Het *distributieprobleem* ontstaat doordat elk nieuw feit eerst moet worden vertaald, en vervolgens moet worden getransporteerd naar een beperkt aantal gebruikers, waar het uiteindelijk wordt geïnterpreteerd en opgeslagen in de database van de ontvanger. Tijdens de vertaling en intepretatie kunnen mappingfouten optreden, wat resulteert in een lagere gegevenskwaliteit. Tijdens het transport kunnen de gegevens worden vertraagd, beschadigd, of afgeleverd aan de verkeerde gebruiker, wat resulteert in inconsistenties in de feiten tussen verschillende locaties.

Door het bestuderen van de data integratie- en distributieliteratuur hebben we een nieuwe methode ontwikkeld, die gegevensafstemming in grote interorganisatorische netwerken mogelijk maakt. Deze methode hebben we Data Afstemming door Logistiek (DAL) methode genoemd. Het idee achter de methode is namelijk het toepassen van logistieke principes op het gegevensafstemmingsprobeem. Wij kwamen tot dit inzicht omdat wij ons realiseerden dat de echte uitdaging in gegevensafstemming het distribueren van gegevens van een brondatabase naar vele ontvangers is, waarbij een zeker kwaliteitsniveau wordt nagestreefd.

Hoe DAL het vertaalprobleem oplost

Om het vertaalprobleem op te lossen hebben we een evaluatie van verschillende data integratie- en distributiemethoden uitgevoerd. Tijdens deze evaluatie constateerden we dat de beste manier om het vertaalprobleem op te lossen het definiëren van een centraal schema per specifieke gebruikersgroep is, waarbij de locale gebruikers hun eigen mappings tussen hun locale en het centrale schema definiëren. Volledige gegevensintegratie van alle database schema's in het

netwerk is niet haalbaar, omdat dit zeer arbeidsintensief is, wat uiteindelijk resulteert in een zeer inflexibel geïntegreerd schema. De centrale-schema-met-locale-mapping benadering voorziet in een grote hoeveelheid flexibiliteit, omdat ieder locale gebruiker zelf bepaalt welke informatie relevant is om naar het centrale schema te mappen. Het nadeel aan deze benadering is dat er op gegevenskwaliteit moet worden gecontroleerd, omdat locale gebruikers zelf het centrale schema moeten interpreteren, wat leidt tot mappingfouten.

Daarom wordt in de DAL methode het vertaalprobleem opgelost door het definiëren van een Informatie Product (IP). Een IP is een centraal schema, wat alle definities en afspraken bevat ten aanzien van een specifieke toepassing voor een specifieke gebruikersgroep. Omdat we dit centrale schema beschouwen als een product, zijn we in staat om de definities en afspraken te focusseren, zodat een minimale set van gemeenschappelijke afspraken wordt verkregen. Een voorbeeld voor de voedingsmiddelensector is het informatieproduct Basisgegevens, wat de product-hiërarchie in de voedingsmiddelensector beschrijft (hoe verhouden consumenteneeheden zich tot handelseenheden, en transporteenheden). Dit IP wordt met name gebruikt voor logistieke toepassingen.

Hoe DAL het distributieprobleem oplost
Uit het evaluatieonderzoek van data integratie- en distributiebenaderingen kwam verder naar voren dat er geen effectief distributiemechanisme voor afstemming tussen meerdere, heterogene database situaties beschikbaar is. Daarom hebben we het distributiegedeelte van de DAL methode gemodelleerd naar het voorbeeld van fysieke distributienetwerken. Vanuit de logistieke theorie weten we dat fysieke distributienetwerken normaliter bestaan uit twee delen: een statisch deel wat de structuur van het netwerk bepaalt, en een dynamisch deel, wat de coördinatie van de goederenstromen regelt. Om de complexiteit van het ontwerpen van een fysiek distributienetwerk te reduceren, worden het statische, en dynamische deel normaliter separaat gemodelleerd. Wij hebben dit principe vanuit de logistiek gebruikt om een distributienetwerk voor feitinformatie te ontwerpen. Tijdens twee case studies, waar we de DAL methode hebben toegepast en verbeterd, hebben we geleerd dat de statische structuur van een feitinformatie-distributienetwerk afhangt van de locatie van de volgende 5 functies:

- De administratiefunctie, die zich bezighoudt met het administreren van het schema van de volledige gebruikersgroep.
- De toekenningsfunctie, die zich bezighoudt met invoeren van nieuwe feiten in de database.
- De controlefunctie, die zich bezighoudt met het uitvoeren van handmatige steekproeven en het definiëren van controleregels, die een controleapplicatie gebruikt om te controleren op mapping fouten of toekenningsfouten.
- De testingfunctie, die zich bezighoudt met het voor de eerste keer testen van de berichtuitwisseling tussen 2 communicatiepartners.
- De profieladministratiefunctie, die zich bezighoudt met het administeren van welke informatie ontvangers wel en niet willen ontvangen.

In de DAL methode is de structuur van het feitinformatie-distributienetwerk het gevolg van het vaststellen waar elk van de vijf functies moet worden geplaatst: centraal in het netwerk, of locaal bij elke bilaterale link in het netwerk.

Conclusie

In dit proefschrift hebben we gekeken hoe gegevenskwaliteitsproblemen kunnen worden voorkomen wanneer computersystemen met elkaar worden verbonden. Met behulp van de DAL methode beschikken zowel E-business branchegroeperingen als individuele bedrijven in B2B netwerken over een middel om hun gegevens op elkaar af te stemmen, wat uiteindelijk leidt tot het voorkomen van gegevenkwaliteitsproblemen in operationele bedrijfsprocessen.

Samenvatting

Curriculum Vitae

Bas Vermeer (1968) received his M.Eng. in Industrial Engineering and Management Science from Eindhoven University of Technology in 1994. He joined Bakkenist Management Consultants (now Deloitte & Touche Bakkenist), where he was a consultant in Logistics and Information Management. There, he worked on several projects ranging from supply chain integration in the food, pharma and technical supply sector, to planning and control systems for administrative processes in the financial sector. From 1998 to 1999 he was project manager of several data warehousing and management information systems projects in the financial sector. In 2000, he started working as a program manager for Cebra, the Center for Electronic Business Research and Application at the University of Eindhoven. There, he is responsible for the Virtue project, which aims to develop the university campus as an experimental garden for E-commerce. In this project, he coordinates an E-commerce program of 20 commercial companies, who use the broadband infrastructure of the university to experiment with 3D technology for online interaction and orientation, online videoconferencing, software agents, and identification and authentication using PKI.

Next to his work at Deloitte & Touche Bakkenist, Bas became a PhD student at the department of Information & Technology in Eindhoven on the subject of system integration in 1995. He published in Computers and Industry, and presented papers at several international conferences such as the Data Quality conference at MIT (1998, 1999) and the Hawaiian International Conference on System Sciences (2000). In 2001 he finished his dissertation.