Problems Concerning External Data Incorporation in Data Warehouses

(HS-IKI-MD-04-104)

Markus Niklasson (e00marni@student.his.se)
Institutionen för kommunikation och information
Högskolan i Skövde, Box 408
S-54128 Skövde, SWEDEN

Magisterarbete vårterminen 2004.
Handledare: Mattias Strand

[Problems Concerning External Data Incorporation in Data Warehouses]

Submitted by Markus Niklasson to Högskolan i Skövde as a dissertation for the degree of M.Sc., in the School of Humanities and Informatics.

[2004-06-11]

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

Signed: ____

Problems Concerning External Data Incorporation in Data Warehouses

Markus Niklasson (e00marni@student.his.se)

Abstract

Data warehouses (DWs) have become one of the largest investments in the past years for organisations, and incorporating external data into a DW can give organisations huge possibilities. Organisations that successfully manage to incorporate external data into a DW have an advantage over those who do not, but there are problems with incorporating data acquired from outside the organisation, and there is a lack of research aimed at these problems. The comprehensive aim of this dissertation is to characterise and categorise problems with incorporating external data. The available literature was scanned to find problems and an interview study was conducted to validate the problems found in the literature. Respondents from five well-known organisations in Sweden participated and the result is a list of problems backed up by both literature and empirical findings.

Keywords: data warehouse, external data, incorporation, problems

Acknowledgements

There are several persons that I would like to thank for the support throughout the work. First and foremost I would like to thank the five respondents that took part in the interviews. I also would like to thank my supervisor Mattias Strand for his valuable feedback, his guidance and his engagement in my work. I also would like to thank Henrik Engström and Mikael Berndtsson for valuable comments on my report. Last, but not least, I would like to thank my family and my girlfriend for their support. Without all of you, this would not have been possible.

Table of contents

1 Introduction

Organisations today experience a tougher climate and a more global market than before. Boundaries between countries and continents are fading away as new technology, like the Internet, makes physical distances less an issue for organisations willing to expand their market share. This means that organisations must not only be aware of the competitors in their own country or region, but must widen their perspective beyond borders to see the complete picture. Not only do organisations have to be conscious about their competitors, but also the customers that currently is gradually adapting to the new global market. Obviously, this all means that organisations must put a lot of work into observing the environment, and adapt to changes that occur. To take external factors into the decision making process are therefore more important than ever, and is a necessity to avoid making decisions that will put the company in an unwanted situation.

To make all this possible, organisations must somehow acquire data from outside the organisation, data that is here referred to as external data. In addition to that, organisations must also know how to use the external data, and how to incorporate the external data into the organisation. To be able to make the most of the external data, it also has to be integrated alongside the internal data. Damato (1999, p.5) describes external data as “wild and untamed”, and to incorporate external data is not a trivial task; external data is by definition something that the organisation does not have control of, and there are cases when organisations avoid incorporating external data simply because there are too many problems associated with the task. However, as Kelly (1996, p. 32-33) explains: “external data is the difference between operational and strategic decision making”, and by that clearly states that external data is of critical importance for strategic and top-level decision-making. Even if the several problems associated with incorporating external data can be hard to handle, organisations that manage to incorporate external data into the organisation experience a huge advantage over those organisations that, for some reason, choose not to incorporate external data.

This work is aimed at characterising and categorising problems associated with incorporating external data that are aimed to be used for decision support and integrated into a data warehouse. The problems described in the literature will be the base for the categorisation and characterisation, which then will be used as a base for an interview study that will be conducted among some of the largest companies in Sweden, in order to validate the problems. The findings from the interview study will then be used to improve the categorisation and characterisation of the problems. The results will shed some light into this rather undeveloped area, enhancing the understanding for the problems organisations have to deal with concerning incorporating external data into data warehouses.

1.1 Problem area

Organisations today need to be aware of the whereabouts in their surroundings to be able to handle an increased competition that many industry sectors experience. To be able to cope with this, organisations need to acquire data from outside the organisation to achieve a wider perspective of their surroundings and their own performance. This data, referred to as external data, is used by organisations to compare themselves with other organisations, or to attain information that the organisation could not get from solely internal data. Evidently, external data are increasingly incorporated into organisations and different systems. Salmeron (2002) recognises that the external data is increasingly being used in executive information systems (EIS), and he further believes that the usage will increase even more in the future. However, even though external data contributes in many positive ways, there are problems that have to be dealt with concerning the incorporation of external data.

Bischoff (1997) claims that the quality of the external data may be questionable, and a study made by the The Data Warehousing Institute showed that 34% of the data quality issues could be referred to external sources (Eckerson, 2002). In addition to that, Damato (1999) explains that organisations do not, by definition have any control over the content, consistency, completeness or correctness for the external data. Strand, Wangler and Olsson (2003) acknowledge that the integration of external data with internal data can be a problem both in terms of technical aspects as conceptual aspects. However, external data is of critical importance and Zhu and Buchmann (2002) explain that some external data may be as important for the organisation that data quality issues have to be neglected. Clearly there are a lot of different problem associated with the incorporation of external data and sometimes, in the worst cases, these problems may cause the organisations to avoid incorporating external data and by that missing out on the advantages that the external data gives.

A problem in the area is that literature covering external data problems in data warehouses is rather fragmented, problems are mentioned in various types of articles but there are no article aimed at gathering various types of problems associated with the external data incorporation. Another problem with the literature is that some problems are mentioned at a very basic level and some are discussed more in-depth. For example, there has been a rather in-depth study when it comes to Swedish financial organisations covered in Strand, Wangler and Lauren (2004a) and Strand, Wangler and Niklasson (2004b). In that study, several different problems came up and it is probable that these problems, even though they were mentioned by financial organisations, are valid for other industries as well. There is clearly a need for a study that is aimed at finding problems of various types and setting up a framework for problems that are associated with the external data incorporation in DWs. This need is supported in Strand and Wangler (2004) where it is claimed that organisations need some kind of support to be able to fully exploit the external data.

There is however another problem, how can one make sure that the problems mentioned in the literature really are valid for organisations today? Some literature may be a few years old and the problems may be out of date; some problems might only be valid or more significant for specific countries (e.g. United States). A large amount of the literature in the area is based on the situation in the United States, and the US is not only ahead of most countries when it comes to data warehouse technology, but also has a different setting when it comes to geographical and demographical issues. It is also reasonable to think that companies in the United States have larger data warehouses compared to other countries. In other words, it is important that the problems found in the literature go through a validation process. The study will be aimed at the Swedish market and it will be interesting to see which problems that are valid in this setting.

1.2 Aim and objectives

The aim is to categorise, characterise and validate problems concerning the incorporation of external data into data warehouses (DWs), with a particular focus on quantitative data.

The aim is divided into these objectives:

- To categorise problems concerning the incorporation of external data into DWs.
- To characterise problems concerning the incorporation of external data into DWs.

- To validate problems concerning the incorporation of external data into DWs.

The first two objectives will be met through a literature study, while the third objective will be met through conducting an interview study. Since the first two objectives will be handled the same way, these will jointly be discussed in the following section.

Objectives 1 and 2 – Characterise and categorise external data problems

The first two objectives are aimed at finding already known problems that are documented in various types of literature. The external data problems will be characterised and categorised to provide a better view of the different types of external data problems and the categorisation means that the different problems can be grouped and related to each other in a better way. The results from objective 1 and 2 will be a list of characterised and categorised problems with external data, and could be seen as a helpful guide on the problems that organisations may have to deal with when incorporating external data into a data warehouse. The categorisation will be based on the external data incorporation process developed by Strand and Wangler (2004).

Objective 3 – Validate the external data problems

In objective 3, the external data problems recognised by the literature will undergo a validation process with the help of an interview study. This validation process will give answer to whether these problems are problems in organisations as well. If the problems found in the literature are acknowledged by the organisations covered in the interview study, the problems can be said to be valid. If the problems are not acknowledged by the organisations participating in the study, it has to be analysed to see if there are any reasons behind this. During the interview study, problems that the literature has not mentioned may be brought up, which could mean that these problems are very specific for the organisations in the study or that the problems have not been mentioned in the literature because no one encountered and documented these problems before. Whatever results the interview study may bring, it will enhance the overall understanding of external data in data warehouses.

1.3 Research process

This section gives a presentation of how the study will be conducted. Figure 1 gives an overview of the process.

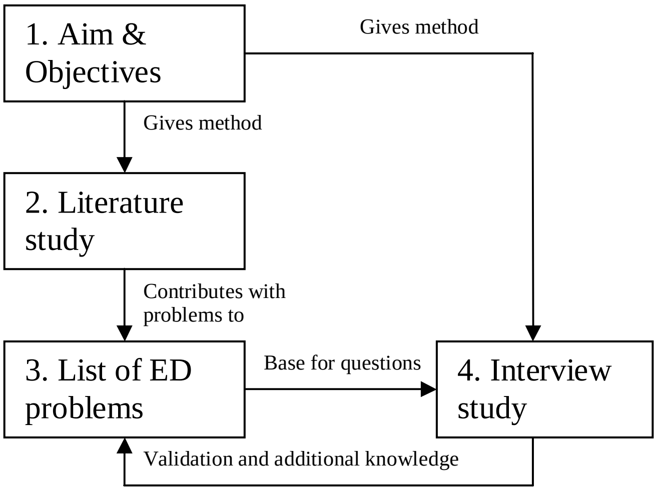

Figure 1 The research process

The aim and objectives (phase 1) will guide the rest of the process and give a help with choosing appropriate methods that can be used to fulfil the objectives.

The literature study (phase 2) gives a base to meet the first two objectives and will mostly be based on literature found in on-line databases. Data warehouse is a relatively new research area and most of the literature is believed to be found in articles. These various databases will be searched with specific keywords in order to systematically go through the literature. Articles are believed to be the most relevant for this project, whereas books usually do not cover the area of external data in data warehouses in a comprehensive way. There are also some web sites directed towards data warehouse researchers and practitioners, and these websites could be valuable sources for the study. According to Berndtsson, Hansson, Olsson and Lundell (2002) it is a problem to know when enough material is gathered for the literature study, and when to stop collecting new material. This will obviously be an issue during the project; however, the literature covering external data in data warehouses is not overwhelmingly huge which is indicated in Strand et al. (2004b). The articles that go deep into this particular issue are easy to count and a problem is that articles mention problems with external data at the same time as discussing a related issue. However, it is reasonable to believe that the problems mentioned in these articles are not very well described, on a rather high-level, and probably covered already in a more comprehensive way in other articles.

The results from the literature study will be presented as a categorisation and each problem will be characterised with the help of the literature (phase 3). The result will therefore be a list of characterised problems categorised in different categories. This categorisation and characterisation will then be the base for the questions used in the interview study.

The interview study (phase 4) will be aimed at the third objective and to reach the objective a number of respondents from different Swedish industries have to be found that agree to take part in the study. The companies in the study should be from different industry sectors as it is assumed that one organisation from a specific industry sector will be an indicator of what types of problems that are valid for organisations in that industry sector. The organisations should also be relatively big in their sector since it is reasonable to think that larger organisations use larger data warehouses and incorporate more external data. The respondents that agree to take part of the study will receive an introducing e-mail with information regarding the interview process and the project in general. The interviews will then be carried out over telephone and notes will be taken during the interviews. Directly after each interview a summary will be written containing the most important that was said during that interview. These summaries will be sent back to the respondents in order for them to approve them and to give them a chance to make any corrections or additions that they may find necessary. The interview study is primarily aimed at validating the problems in the literature, but also at enhancing the understanding of the nature of the problems. This means that the interview study will be a way to enhance the list of problems especially when it comes to the characterisation. There might be some contributions to the categorisation if the interview study shows that there are other problems than those described in literature. In either case, the categorisation and characterisation will be enhanced with empirical results after the interview study is conducted. This way the final version of the categorisation and characterisation will contain material from both the literature and the empirical study.

1.4 Main contributions

There will be several contributions from this project. First of all the resulting list of problems regarding the external data incorporation can be used by organisations to be better prepared when it comes to initiating new data warehouse projects or expanding a data warehouse by acquiring additional external data. Knowing what kind of problems that may arise during the data warehouse project would allow the organisations to be better prepared to handle these problems. Since there has not been found any other study focusing on problems with incorporating external data into DWs, this study will act as a summary of the fragmented literature in the area, with additional material from the interviews. The validation of the problems mentioned in the literature will also point out problems that for some reason are not actual problems in organisations today. The study will expand the over-all understanding of the area, especially since the study will be based on both empirical results and literature.

2 Background

This section is dedicated to describing important concepts that the work is based on. Both data warehouse and external data are defined and described.

2.1 Data warehouse

There are a few definitions of data warehouse in the literature, but the most frequently used is the one stated by Inmon (1996). His definition focus on the characteristics of the data that is located in the data warehouse. The characteristics Inmon (1996) mentions are subject-oriented, integrated, time-variant and non-volatile.

Subject-oriented means that data in a DW is focused on the main subjects of the organisation (e.g. customers, products, sales etc.). This implies that the data is not focused on functions or applications.

Integrated means that all data in a DW is integrated and unified, making all data a part of one big consistent data source. Data that is integrated into a DW includes summarized data, derived data, and historical data acquired from both external and internal sources.

Time-variant implies that the data in a DW is related to a specific time. Every record in the DW can be seen as a snapshot of the current state of the organisation for that specific time. Thereby a DW can be seen as a large collection of snapshots.

Non-volatile is an important characteristic for data in a DW. It gives that data stored in a DW is read-only and is not to be deleted or updated. This characteristic is important because if the data in the DW is updated as soon as something occurs in the organisation, it would mean that there would not be any historical data in the DW. As a substitute for updating the records, new records containing the new values are instead inserted as the DW are updated, and the old data is by that kept as historical data.

Even though Inmon's definition is the most well-known and widely used, it is also worth mentioning the definition set by Singh (1998, p.14):
"A data warehouse supports business analysis and decision making by creating an integrated database of consistent, subject-oriented, historical information. It integrates data from multiple, incompatible systems into one consolidated database. By transforming data into meaningful information, a data warehouse allows business managers to perform more substantive, accurate, and consistent analysis."

While Inmon concentrates mostly on the data in the DW, Singh have clearly more of an organisational focus in his definition. The focus is on the role of the DW in organisations, and explains that the main task for the DW is to support the decision makers to perform business analysis. These two definitions by Inmon (1996), and Singh (1998), together provide a decent base for understanding the DW, and the role that a data warehouse plays in an organisation.

2.2 External data

This section describes the definition of external data, the different types of external data, what sources it can be acquired from and the process of incorporating external data.

2.2.1 Definition

There are a few definitions of external data that can be found in literature. One of them that is often used and suits the aim of this work is the definition made by Devlin (1997, p.135):

"Business data (and its associated metadata), originating from one business, that may be used as part of either the operational or the informational processes of another business."

Kelly (1996, p.33) has a slightly different view of what external data is:

"External data is captured outside the enterprise and is, most often, made available at a cost by specialist information providers."

Kelly (1996) and Devlin (1997) both have the standpoint that external data is data acquired from outside the organisation. What is slightly different is that Kelly (1996) claims that organisations mostly have to pay for the data, something that also is in line with the definition from Kimball (1996) where he refers to "syndicate data", data that is bought from data suppliers. However, the definitions by Kimball (1996) and Kelly (1996) are too narrow for this project, and as Devlin's definition is more general, his definition will be adopted.

2.2.2 External data types

Data acquired from external sources can be of many types and may be used in various ways. Oglesby (1999) has found four rather broad categories of data that organisations may be able to acquire externally. The four categories are:

- Consumer demographics/psychographics
- Business profiles
- Address/phone verification
- Industry specific

Consumer demographics are according to Oglesby (1999), basic data about people such as age, income, education, marital status and so forth. *Psychographics* are what individuals like to do, such as hobbies. Various types of marketing information companies may provide these types of data, and according to Damato (1999) is consumer demographic data something that organisations often tend to use for decision-making. Kelly (1996) recognises what he refers to as econometric data as something organisations may incorporate into data warehouses. This type of data contains information about customers' income groups and consuming behaviour. Acxiom (2000) also recognises that companies selling to consumers, called business-to-consumer (B2C) companies, often use these types of data.

Business profiles are in some ways the same as consumer demographics, but with the modification that it is data about companies instead of individuals. Oglesby (1999) explains that this type of data could contain information such as number of employees, year established, annual revenue, description of the business, and different industry classification codes. Acxiom (2000) chose to call this type of data "firmographics", and Acxiom (2000) further indicates that this type of data could be used for organisations selling to other businesses, something that usually is referred to as business-to-business (B2B) companies.

Address and phone verification data is data that contains addresses and telephone numbers. It is often used by organisations to clean and update their data warehouses to make sure that the addresses and telephone numbers are valid. Often when people enter information like addresses in, for example, an ordering form on the Internet, they may make mistakes and enter incorrect information. Oglesby (1999) claims that people mistype information almost 10 percent of the time. This means that it is important to assure that the information is correct and organisations need to verify the information by acquiring address information from external sources. If addresses are not verified and updated, it will potentially harm the costumers' good will against the organisation, even though it actually might have been the customers fault in the first place.

Industry specific data is data that more or less varies from different industry sectors. An example: Strand and Olsson (2003) found in a study that industry codes are frequently incorporated into data warehouses in organisations. These codes make it possible for organisations to see what type of business a specific organisation is, and they can be a good way for B2B companies to categorise their customers. This specific example is related to business profiles, which points to the fact that some industry specific data also is related to other types of data.

2.2.3 External data sources and suppliers

According to Strand et al. (2003), there are a number of different types of organisations that can supply data to other organisations. These will be briefly described in this section. The different types of sources according to Strand et al. (2003) are:

- *Statistics institutes*: Organisations that are able to distribute various types of statistical data about individuals, organisations and so forth.
- *Syndicate data suppliers*: Organisations that provide economical company data that helps companies reducing credit risks, finding customers that are profitable, and manage vendors.
- *Industry organisations*: Information about specific industries can be acquired from these organisations.
- *County councils and municipalities*: Statistics regarding population, demographics and so on might be available without charge from these organisations.
- *The Internet*: The Internet is a rather enormous data source, but there is a problem in sorting out relevant data. Organisations can by scanning the web, for example, find out what competitors offer and to which prices.
- *Business partners*: Data from other businesses can give organisations a better overview of the business climate and their surroundings.
- *Bi-product data suppliers*: Organisations that have other core business activities but also sell some of their data to other organisations.

2.2.4 External data incorporation

This section briefly describes the external data incorporation process.

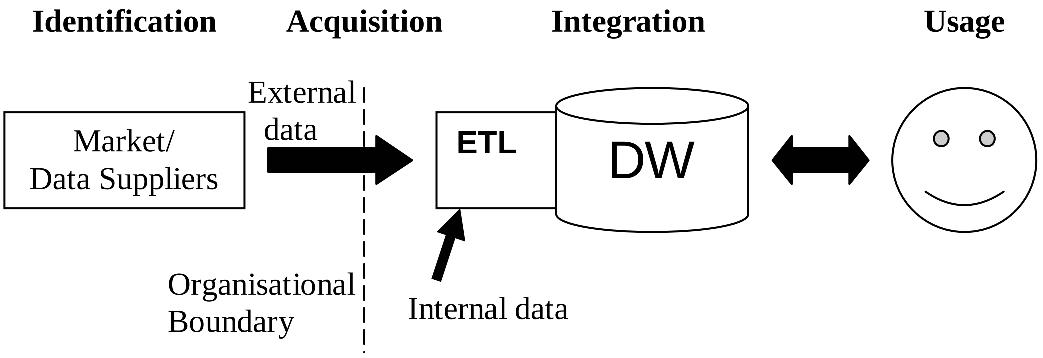

Figure 2 The external data incorporation process (based on Strand and Wangler, 2004, p.4).

The external data incorporation process consists of four phases according to Strand and Wangler (2004):

Identification
Identification is according to Strand and Wangler (2004) the activities that include finding and evaluating sources from where organisations can acquire data. This phase is important since some external data suppliers might be able to tailor data so that it becomes custom-made for acquiring organisations, making the later phases less problematic. This is something the organisations have to find out before they choose supplier. The reputation of data suppliers can also infect the willingness of organisations to acquire external data according to Strand and Wangler (2004).

Acquisition
Acquisition is according to Strand and Wangler (2004) the process of retrieving external data from a source, and distribute it to various internal systems. This is where the external data passes the organisational boundary. There are several different approaches that can be used to transfer data between organisations, and examples are compact discs, DVD-ROMs, magnetic tapes, or with the help of Internet (e.g. FTP). Strand and Wangler (2004) explain that if data is acquired in an appropriate way it can contribute to a less problematic integration.

Integration
Strand and Wangler (2004) clarify that the integration involves how the data is stored, how it is modelled and how the data is integrated into the data warehouse. According to the study by Strand and Wangler (2004), the integration phase is problematic for organisations in several ways. It is also important, they say, to acknowledge that the choice of integration approach depends on how the data is about to be used.

Usage
The usage phase includes, according to Strand and Wangler (2004), aspects such as how data is interpreted and the purpose of the data. The human involvement in the usage phase makes the usage problems difficult to solve. Strand and Wangler (2004) acknowledge that many of the usage problems can be solved in earlier phases.

3 The literature study

This section presents how the literature study was conducted.

3.1 Finding literature

When the aim and objectives (phase 1) were set up, the literature study (phase 2) was initiated. The first impression was that the literature is segmented when it comes to problems with external data. Not a single article focused on external data problems; instead, these problems often were found when discussing other aspects, such as application areas. To find relevant literature to support the study mostly articles were searched for, since articles are the most up-front and usually focus on a more narrowed area. Books in the data warehouse area usually do not mention external data in such a comprehensive way that they can be considered as an important source for this project. Internet was used to locate different article databases such as Springer Link, Science Direct, INSPEC, Cite Seer, and ESBCOHost Research Databases. Words used for searching included: “external data”, “data warehouse”, “data warehouses”, “problems”, and “issues”. These words were combined in different ways to make the search as efficient as possible. However, there are authors that do not use the exact phrase “external data” but instead something like “environmental data” or similar. However, the literature study was only aimed at finding articles that referred to “external data” or else the literature study would be an overwhelming process. Searches were also made at www.google.com even though results from such a search had to be treated with care and the source would sometimes not to be considered as high-quality or trustable. Lists of references located in the articles found were also a way to find more interesting articles in the area. However, these were often very hard to locate and it is hard to understand what the article really is about since only the title is known. This is not the first time I conduct a literature study in the area and the experience is that these articles are of lesser importance for a study like this. So actually not many new findings were made through looking in lists of references, but they were a good way to get a view of the literature in the area. Another useful source was two websites focusing on data warehouse technology, www.datawarehouse.com and www.dmreview.com.

A problem that arose during the literature study was that authors have different opinions on what external data is. Sometimes authors referred to external data as all data that is in the organisation but not integrated into the data warehouse. This was the case especially in more technical-oriented articles as opposed to organisation-oriented articles where external data mostly meant that the data was external for the organisation.

3.2 Handling the results from the literature study

The problems were written down with a description as they were found, and when the same problem was mentioned in another article the description of the problem was enhanced. When no new articles covering new problems were found in the area, the result was an unstructured list of several problems covering different aspects. The task was now to make categories to be able to place the different problems in the appropriate category. As a base for the categorisation, the external data incorporation process developed by Strand and Wangler (2004) was used. This process contains four steps (as can be seen in Figure 2, p.9); identification, acquisition, integration and usage, and these four steps are the main categories for the categorisation. The result became a list of characterised and categorised problems (phase 3) that mainly will be used in order to produce appropriate questions for the interview study (phase 4).

4 Analysis of the literature study

This section presents the external data incorporation problems that were found during the literature study, phase 2 in the research process (see 1.3). The problems are categorised to provide a structure and the problems are characterised to give a better understanding for each problem. The categorisation is based on the external data incorporation process developed by Strand and Wangler (2004). Worth mentioning is that some problems might seem irrelevant due to, for example, old literature sources. However, these problems were still taken into consideration during the literature study. If problems are found to be outdated according to the interview study, they will be left out in the final categorisation.

4.1 Identification

In this section, the problems that can be referred to the identification phase are presented.

Identify new suppliers
Strand et al. (2004b) acknowledge that it is a problem with identifying new external data suppliers that just entered the scene. They further state that there is a lack of formal routines that can be used for identifying external data suppliers, both new and old ones.

Establish relations
According to Oglesby (1999), it is time-consuming to establish relationships with external data suppliers. This is especially a problem for mid-size companies. Oglesby (1999) is not very clear when it comes to describing what he means by establishing relationships, but the assumption is that it includes activities that take place after a supplier is identified and before data is acquired from the supplier. Examples of activities could be agreeing on prices, setting up contracts, deciding how the data should be transferred to the organisation, and so forth.

Select suppliers
Strand et al. (2004b) identify that it is a problem to select the most appropriate suppliers for external data. There are often several suppliers that can provide external data that organisations need.

Identify relevant external data
Zhu and Buchmann (2002) claim that it is hard to identify relevant external data on the web. Web sites that e.g. are belonging to a competitor can contain data that is relevant for organisations, but to find the external data that can be relevant is difficult.

4.2 Acquisition

This section presents the problems that can be referred to the acquisition phase.

Time-consuming
Acxiom (2000) states that the process of updating internal data, with the help of external data is a problem mainly due to the transportation of the data, which is time-consuming. Acxiom (2000) gives an example concerning an organisation that updates addresses in the United States. Data is prepared and shipped to services outside the organisation, and the data is stored on tapes. When the data comes back, the reports created by the suppliers have to be studied. Obviously, this causes more problems such as the fact that the data is shipped means that the data will be old whenever it returns to the organisations. It also could be an economic problem since this shipping probably is quite expensive.

Acquire data
Strand et al. (2003) acknowledge the fact that it is hard to acquire external data from industry organisations. This is even more complicated if organisations try to acquire data from industry organisations that are functioning in another industry sector. Strand et al. (2003) got the impression that some industries do not share the internal industry data to external organisations due to some kind of agreement in the industry.

Large data sets
Oglesby (1999) indicates that large data sets may be a problem, because some data sets are too large to fit onto a compact disc, which is a common way for external data suppliers to deliver external data to organisations according to Oglesby (1999). However, Strand et al. (2004a) show in a study that 88% of the organisations use FTP for transferring data and acknowledge that DVD-ROMs can be a substitute for CDs. Oglesby (1999) further acknowledges that this issue regarding acquiring large data sets will become less important when Internet connections get faster.

Dynamic sources
Some types of sources may be very dynamic; this is specifically a problem for web sites. Data on the web is often updated frequently, new potential sources are brought in all the time, and some sources may disappear or undergo a drastic change either content-wise or representation-wise. Reasons for a change can be new upcoming developments, but some providers also would like to hinder other organisations from automatically extract their data and therefore change the representation frequently to cause problems for the companies that extract their data. All this makes a data warehouse based on web data more complex than a conventional data warehouse (Zhu, 1999; Zhu, Bornhövd, Sautner & Buchmann, 2000).

Source stability
Source stability is mainly an issue dealing with web sites. Zhu and Buchmann (2002) identify some characteristics that can be an issue concerning the stability:

- Availability; is the source up and running, is the response times good enough, is the links working.
- Accessibility; sometimes registration and passwords are required to access certain information, which makes the automatic extraction more complex.
- Durability; how long is the data available. If the source is updated often and the old data is removed in the process, there is a need to regularly extract the data to guarantee that nothing is missed. Often web sites do not keep historical data, which can be a problem.

Even though Zhu and Buchmann (2002) discuss this issue regarding web sites it is possible that this also is a problem for transactions by FTP.

Expensive
External data is rather expensive to acquire, especially from syndicate data suppliers (Strand et al., 2004a; Strand et al., 2004b). This is even more of a problem if the organisation needs to acquire the data on-demand in relation to acquiring it via subscriptions (Strand et al., 2004b). Oglesby (1999) however claims that the external data can be reasonably priced if the organisation has a large data warehouse and by that can overcome the minimum-order size that some external data suppliers apparently are using. However, Oglesby (1999) acknowledges that this is a problem for mid-size companies with a smaller data warehouse.

4.3 Integration

This section discusses problems associated with the integration of external data. The integration of external data with internal data is problematic in several ways; Damato (1999, p.5) refers to external data as “wild and untamed” and this implies that there is a problem to integrate external data with internal data. In a study conducted by Strand and Olsson (2003), they found that 60% of the companies experienced problems when mapping external and internal data. Strand et al. (2003) also claim that there are two aspects of the integration, a technical and a conceptual.

Time-consuming

To integrate external data into a data warehouse is very difficult and time-consuming according to Adelman (1998) and explains that it is because there are problems with the external data such as the data representation. Strand et al. (2004a) recognise the transformation process as very time-consuming and costly. This problem is related to all other integration problems really, since the more problems organisations have with different integration aspects the more time-consuming it will be.

Data representation and structure

External data is most likely not represented in an ideal way for the acquiring organisation and Adelman (1998) explains that external data usually does not follow the standards that the acquiring organisation has. Strand et al. (2004a) claim that the differences in data structure between the internal and external data are the most common problem when incorporating external data. Data warehouses are mostly based on either relational or multi-dimensional data models, which clearly can be referred to as structured data. However, external data, especially the data available on the web is mostly unstructured or semi-structured, which would cause problems according to Zhu et al. (2000). Zhu and Buchmann (2002) explain that to be able to use automatic extraction of data from the web only structured or semi-structured data is useful, semi-structured data on the web could for example be available as XML. However, HTML is currently the most used on the web and HTML is unstructured according to Zhu and Buchmann (2002). There are also, they say, other types of unstructured data and information available on the web in formats as pdf, ps, doc, or other formats that support documents, pictures or audio. To integrate unstructured data would have to be done manually according to Zhu and Buchmann (2002), and this is not only very time-consuming but there is also a high risk for transcription errors. Bischoff (1997) explains that there are a large amount of different data formats that are used and further explains that the external data format can change frequently which makes the integration process even more complex. Zhu et al. (2000) acknowledge that data providers on the web can drastically change the representation and content of their data. Changes can be made because of new developments, or to make automatic extraction of their data more complex.

Storage

Collett (2002) explains that external data, or basically data in general, consumes disc space, and of course this may be a problem even though storage devices nowadays are less expensive and can hold more data.

System consistency

Strand et al. (2004b) acknowledge a problem with assuring that all the systems in an organisation are updated when new external data is acquired. This may lead to inconsistency between the different systems as some systems are updated and some are not.

Data completeness

If organisations are unable to acquire all important data, problems arise in the integration process. If, for example, customer identifiers are not provided from the external data source, this cause significantly more work to make sure that the external customer data is accurately integrated with the internal customer data (Adelman, 1998). Strand et al. (2004b) also recognise that it is hard to match internal and external data due to incomplete data. This could result in additional problems when the data is to be used. Strand et al. (2004a) explain that organisations are not allowed to get all data about individuals that are not customers for the company, which makes this problem even tougher to deal with. Zhu et al. (2000) explain that default-values may be a possible solution to some of these problems. However, identifiers can not be replaced by default-values for obvious reasons.

Time-stamps

Damato (1999) claims that it is important to ensure that external data is valid at the same time as the related internal data. To make that possible the external data must contain time-stamps. However, problems arise as external data sometimes lacks time-stamps, something that Zhu and Buchmann (2002) mention is a problem with web data. Strand et al. (2004a) recognise another problem related to time-stamps; the data may have been time-stamped from an integration perspective and not from a real time perspective. Real-time time-stamped data is what organisations need to be able to map external data with internal data from the same time. Strand et al. (2004a) further explain that this is especially a problem when the source of the data is unknown.

Data overload

Collett (2002) explains that data overload causes many problems for organisations. To acquire too much external data means that it can become an overwhelming process to integrate the data. A lot of work and time is required to pick and integrate the data into the DW, and to clean up files that contain irrelevant data.

Conflicting data

Damato (1999) acknowledges the fact that using several different external sources may lead to a problem with conflicting data. This means that there is a problem when external data is integrating with internal data. Organisations must then figure out which source that is to be considered as the most reliable, and then integrate data from that source into the data warehouse.

Tools

The literature describes various problems concerning tool support for extracting, transforming and loading (ETL). First of all Strand et al. (2004a) mention that investing in a commercially developed ETL-tool is a large investment, which may be a problem for especially smaller organisations, and could force organisations into develop their ETL-tools themselves. Strand et al. (2004a) further point out that these commercial ETL-tools work like black-boxes. The external data is entered and mixed up with the internal data without any control of the data quality. This further indicates that in-house developed ETL-tools may be worth while. Damato (1999) also experiences a lack of support concerning extraction software used for extracting data from electronic on-line databases. To work around this problem, Damato explains that the external data could first be loaded into tables that the extraction software supports. He also adds that the technology keeps getting better and more efficient, but external data is still more difficult to access, clean and load into the data warehouse compared to internal data. This type of problem is relevant for both the acquisition and integration; however, it seems that the problems mostly occur during the integration.

Conceptual understanding
Strand et al. (2004a) explain that the conceptual meaning of the external data can be difficult to interpret. Damato (1999) explains that this may cause external data to be interpreted in an unintended way. Damato (1999) also states that it is difficult to know where the external data fits into the model that the data warehouse relies on. Zhu et al. (2000) claim that it is important to make sure that there is a semantic homogeneity among the internal and external sources before the data is loaded into the warehouse. This they say is important for web data; on the other hand, this could be a possible problem for all types of data. Zhu (1999) also clarifies that these semantic inconsistencies do not only exist between the external web data and the internal data, but also between different web data acquired from different sources. The problems concerning conceptual understanding are heavily dependent on the human ability. However, other things such as lack of metadata make the problems even more complex.

Metadata
In order to be able to understand all kinds of data, metadata is important. Especially for external data, since not much will be known about it otherwise. The lack of metadata for external data can become a problem according to Damato (1999). Sometimes the external data is stored somewhere without any existing metadata that describes the data and explains the relationships to other data located in other sources. Adelman (1998) recognises that external data often is not well documented. Zhu and Buchmann (2002) indicate that the lack of metadata may be a problem concerning external data from web sites, Zhu (1999) also claim that there could be problems with interpreting related metadata for external data acquired from the web. Even though these interpretation problems are mentioned in relation to web data it is reasonable to think that these problems also can be an issue concerning external data from other sources, even though the problems most likely are biggest dealing with web data. Adelman (1998) further explains that if the external data has been modelled it will be significantly easier to understand. However, Adelman (1998) realises that external data suppliers are unlikely to provide a model for the external data.

4.4 Usage

This section contains characterised problems from an end-user perspective.

Data correctness
Strand et al. (2004a) explain that if the external data is not correct, important decisions will be made on incorrect data or facts. Zhu and Buchmann (2002) also identify that correctness is important when it comes to external data. They further explain that data taken from web sites may not be as carefully reviewed or filtered as regularly information sources, and that there exist a lot of incorrect information on the web.

Data completeness
If data is missing or incomplete it will bring added complexity into decision making. Strand et al. (2004b) explain that because there sometimes are problems with matching external and internal data it will cause problems for organisations to, for example, see which people that already are customers. This is caused by the fact that they can not acquire all the data they need and that organisations might send out information to people that already are customers, causing the customers to get a negative impression of the organisation. Zhu and Buchmann (2002) explain that there exist a lot of incomplete facts on the Internet. They also

acknowledge that lack of units may be a problem with data acquired from the web. Some data completeness problems are according to Strand et al. (2004b) caused by regulating laws that hinder the organisations in acquiring certain data externally.

Data freshness

Data acquired from external sources may be old according to Strand et al. (2004a), which would mean that decisions are made on old data. However, Strand et al. (2004b) acknowledge the fact that external data can never be minute-fresh and data such as statistical data only is valid for a short time meaning that such data soon is to be old. Oglesby (1999) explains that the data on a CD is out of date the day after it is pressed. Bischoff (1997) explains that the external data must be documented and deployed as quickly as possible to reduce these problems. There is however also a cost issue involved. Zhu and Buchmann (2002) give an example of this; stock quotations have a price if an organisation needs to acquire it immediately, but are free to acquire if an organisation can cope with 15 minutes old data.

Data overload

Collett (2002) explains that if there is too much external data available, much of it may be left unanalysed. This problem can also be related to the cost issue, if an organisation acquires too much external data from e.g. a syndicate data supplier it also means that the organisation pays for a lot of data that in the end is not even used. Another problem with this might be that it could be hard to know for decision makers what specific data the decisions should be based on. In case the decision makers have to figure this out for themselves, it can cause problems and an increased workload for the decision makers.

Biased data

Zhu and Buchmann (2002) explain that external data may be biased which is an unwanted situation. They further explain that biased information could mean that some information has been left out (e.g. information that favour a competitor), and in that case the problem of biased data is also a problem of incomplete data since some parts have been left out. Zhu et al. (2000) further explain that the external data can be abnormal because of different assumptions that the organisation made, or different political and cultural contexts. They also mean that different intentions that an organisation have for the data can lead to biased data. Sadly Zhu et al. (2000) do not give any examples of different intensions that can influence the data.

Data reliability

According to Zhu and Buchmann (2002), the origin of the data may contribute to how reliable the data is perceived in the company. Strand et al. (2004a) agree and explain that if the origin of the data is unknown it is difficult to rely on.

Conflicting sources

According to Damato (1999), it may be a problem to use several different external sources, which may lead to conflicting data. This would mean that the end-user has to choose which data they trust the most, and base the decisions on the data they find most trustworthy and accurate. However, these situations require the users to have the knowledge and ability to estimate data quality, and different users may make different assumptions.

Exists without influencing the decisions

If external data exists in a company in some form, but is not integrated into the data warehouse it could mean that it will not be taken into consideration when it comes to making decisions (Damato, 1999).

Restricting laws

Strand et al. (2004b) recognise that companies are restricted by laws when it comes to what they are allowed to do with the external data. It is also an issue that laws can be rather vague when it comes to what organisations are allowed to do with the data. It is very important that the organisations take the time to understand the laws to make sure that they do not use the data in an illegal way.

Ethical aspects

To use the external data in certain ways may conflict with the ethical views that exist in e.g. an organisation (Strand et al., 2004b). People generally have their own ethical principles, and it is probable that their ethical views are influencing their work and decision-making.

4.5 The resulting categorisation

Presented in Table 1 is an overview of the problems that were found during the literature study.

Table 1 The categorisation after the literature study.

Identification
- Identify new suppliers - Establish relations - Select suppliers - Identify relevant data
Acquisition
- Time-consuming - Acquire data - Large data sets - Dynamic sources - Source stability - Expensive
Integration
- Time-consuming - Data representation and structure - Storage - System consistency - Data completeness - Time-stamps - Data overload - Conflicting data - Tools - Conceptual understanding - Metadata
Usage
- Data correctness - Data completeness - Data freshness - Data overload - Biased data - Data reliability - Conflicting sources - Exists without influencing the decisions - Restricting laws - Ethical aspects

Worth mentioning regarding the problems that are brought up in this chapter, is that most of them are not exclusive for external data but are also issues for internal data. External data however brings another dimension and makes the problems more complex compared to the related problems with the internal data. Some problems also arise in the mixing of external and internal data. Problems mentioned in identification and acquisition, are more valid for external data and probably not problems at all for internal data. This can be related to the definition of external data that is given in 2.2.1, since what is specific for external data is that it is taken from outside the organisation. Strand and Wangler (2004) explain that the organisational boundary lies in the acquisition phase. This means that the two first phases are the most specific for external data and the problems encountered in those phases can therefore be seen as rather specific for external data. The integration and usage problems on the other hand are probably issues for internal as well as external data.

5 The interview study

This section first describes how the process of conducting the interview study (phase 4 in the research process in 1.3) was carried out, and then the five respondents that took part of the study are described.

5.1 Setting up the interviews

The interview study began with identifying organisations that would be suitable for the study. To find suitable companies I brainstormed and wrote down what I considered to be the most well-known companies in Sweden. Other people were also asked the question and contributed with ideas on organisations that could be appropriate for the study. The yellow pages were also scanned to get any further ideas on appropriate organisations. Organisations were contacted by telephone and the project was initially introduced and the organisation was checked to see whether the organisation met the demands (having a data warehouse with external data) or not. If the organisation used a data warehouse and external data, the aim was to find the person who was responsible for that part in the organisation. Often the person that was first spoken to could direct me to a person that worked with the data warehouse. When I finally arrived at the appropriate person, the person was asked if he or she could take part of an interview study and was given a brief description of how the interview would be conducted. Needless to say, there were a lot of companies that were contacted that did not have a data warehouse or incorporated external data. Worth mentioning is also that one organisation had to refuse because the organisation's policies were saying that they were not allowed to speak about their data warehouse, which indicates that the data warehouse area is quite secretive and full of company secrets. Five persons agreed on participating in the interview study, and a time for the interview was scheduled for each participant. All of the participating persons then received an introducing e-mail where the project and the interview study were described. This introducing e-mail can be seen in Appendix 1.

5.2 Interview questions

The interview questions can be found in Appendix 2 and were mostly based on the categorisation in Chapter 4; every problem in the categorisation got a corresponding question in the interview. However, first of all there were some introducing questions with the intension of making a smooth start in the interviews and also to make the respondent think about the area in terms of what external sources they acquire data from, and what type of data they acquire from these sources. Lastly, the respondent was asked if there were any other problems regarding external data that he or she had experienced in the organisation, a question aimed at enhancing the list of problems. For each problem, there were a number of questions aimed at enhancing the characterisation of the problem. These were not always explicitly asked but rather a guide for asking follow-up questions when the respondent did not already cover these areas in the discussion.

5.3 Conducting the interviews

The five interviews were conducted over telephone, and notes were taken during the interviews. The interviews lasted for approximately 30-60 minutes each. The notes taken during the interviews were then the base for summaries that were written for each interview. The summaries were then sent back to the respondents in order for them to approve what has been written, and a chance for them to make corrections, additions or other changes. The final versions of the summaries can be found in Appendix 3. The summaries were then used to improve and enhance the original categorisation and characterisation, and together with some ideas and analysis from the author a new improved categorisation and characterisation emerged which are to be seen in section 6. Note that sometimes the respondent gave an answer to a question regarding a specific problem that is more suitable as an answer to another problem, most often a similar problem but in another phase of the external data incorporation process. In these cases the answers were still presented in the summaries under the problem that was discussed at the time and then in the analysis these issues are discussed where it is found most appropriate.

5.4 The respondents

This section gives a brief description of the five respondents that took part of the interview study. The information regarding the organisations has been gathered from the organisations' web sites.

Respondent 1
The first respondent is responsible for the development of Göteborgs-Posten's data warehouse (or customer database as the respondent sometimes refer to it). Göteborgs-Posten (GP) is the second largest morning paper in Sweden and the largest morning paper in Western Sweden.

Respondent 2
The second respondent is a system developer and involved in projects concerning business intelligence and data warehousing for Volvo IT in Skövde. Volvo is one of the most well-known car manufacturers not only in Sweden but also world-wide.

Respondent 3
The third respondent has the role "Manager IT Retail Sweden" for ICA AB. ICA AB is the largest retail trade organisation in the Nordic countries, and the main focus is groceries.

Respondent 4
The fourth respondent works as a database administrator for Preem. Preem is not only Sweden's largest oil manufacturer with two oil refineries, but the company also has petrol stations all around Sweden. Approximately half of the production goes on export, making Preem one of the biggest exporters in Sweden.

Respondent 5
The fifth respondent works for Astra Zeneca and is responsible for the business intelligence capability in the IS operation and business service section. Astra Zeneca is a world leading organisation in the pharmaceutical industry.

6 Analysis of the interview study

In this section each problem found in the literature study are analysed based on the findings in the interview study. The section focuses on the characterisation of the problems and connections to the previous results in the literature study are only discussed when it is considered to be interesting. Sometimes the organisations' names are used instead of the respondents' names. However, what is presented in this section are from the individual respondents' point of view, and is not necessarily a view that is shared in the entire organisation.

6.1 Identification

This section presents the findings from the interview study regarding the identification phase.

Identify new suppliers
During the interviews it was found that some organisations tend to have some problems with finding new suppliers. A big problem in the area, which both Astra Zeneca and ICA AB mentioned is that some data suppliers are in a monopoly situation which makes it hard for new suppliers to come up and compete against these big suppliers. Astra Zeneca mentioned that they are constantly looking for other possible solutions and that there are some other suppliers available, even though most suppliers can not deliver as fully-covered data as the organisation need. Volvo has some problems with finding the sub-contractors that are the most suitable for delivering the data Volvo needs. The respondent explained that it might be hard to know which sub-contractors that can deliver the appropriate data.

There are also companies that do not actively look for new suppliers, Preem explained that the organisation has the data they need for the moment and they know where to turn to acquire the data they need. However, Preem also acknowledge that there might be suppliers out there that can provide the data they need for a lower cost, but the respondent has not seen any advertisings or any direct marketing attempts from any other supplier. The same goes for Göteborgs-Posten who did not actively look for new suppliers.

Establish relations
Only one of the respondents, Volvo, experiences problems with establishing relations with suppliers, and it has to do with the fact that their sub-contractors not are as technological advanced as Volvo. The respondent said that Volvo has to have high demands on their sub-contractors but also said that they can not require that the sub-contractors have as high standards as Volvo. Volvo differs from the other organisations in the study since they acquire data from sub-contractors while the others mostly acquire data from syndicate data suppliers. It is reasonable to think that these syndicate data suppliers, who have a main focus in delivering data to organisations, have been working a lot with establishing relations and by that making it easier for the acquiring organisation as well. Since selling data is the syndicate data suppliers main focus, they have to be good at this, otherwise organisations would most likely change to another supplier, since there often is a selection of suppliers around. Astra Zeneca said that they experience some cultural differences since their supplier is based in United States, but Astra Zeneca does not see this as a problem. The main thing is that their supplier knows that they are in a monopoly situation and that they sometimes take advantage of this. Oglesby (1999) acknowledges that establishing relations is especially problematic for mid-size companies and the organisations in the study are rather large which might explain the result.

Select suppliers
Astra Zeneca and ICA AB both acquire data from a supplier in a monopoly situation and do not have any problems with selecting suppliers, since it is obvious that there are no other options. Only one organisation experiences problems regarding selecting suppliers and it is Volvo. Volvo requires the data and the supplier to uphold a certain quality and sometimes when the suppliers do not reach up to the required level Volvo has to change supplier. The respondent also said that the sub-contractors have to be of a certain size to be able to handle data deliveries to Volvo. Volvo wants suppliers that can deliver as fully-covering data as possible. Too small sub-contractors are more likely to not being able to cope with the data deliveries.

Most of the organisations do find it easy to select suppliers, and this is quite surprising. There are a lot of suppliers that are able to deliver the data the organisations need according to Strand et al. (2004b), but still when looking at the organisations acquiring data from syndicate data suppliers they do not think that selecting suppliers is a problem. A reason for this might be that some of these companies have a very product-oriented focus in their data warehouse. Groceries, oil and medicines are rather specific areas and there might not be as many suppliers in those areas, as for example suppliers that focus on data about individuals.

Identify relevant external data
Two out of the five respondents experience a problem with identifying relevant external data, Volvo and ICA AB. Volvo has to work out which data that they need that the suppliers can give them, and they have to screen out data that are not wanted. ICA AB said that this is a question about quality; they need to be confident that the quality is assured. The respondent at ICA also believes that this will become a bigger problem in the future.

Zhu and Buchmann (2002) acknowledge this problem and focus on the difficulties with identifying data on the web, but this study shows that there also can be problems to identify relevant data when it comes to other sources as well.

6.2 Acquisition

This section presents the findings from the interview study that can be referred to problems in the acquisition phase.

Time-consuming
Only two of the five respondents acknowledge that the acquisition is time-consuming, most of the organisations have methods that automatically acquire the data from the supplier. Volvo experiences some problems, and it was mainly due to the high security standards adapted by Volvo. Their sub-contractors might have to adapt to Volvo’s standards and methods to be able to communicate safely, which is a time-consuming task. ICA AB has some problems occasionally and it is because there are some changes in the flow on occasions.

Acxiom (2000) acknowledges the problem and explain that this is an issue if organisations have to ship e.g. tapes or compact discs between the supplier and the organisation. However, this is not how it is done these days; all of the organisations that took part of the interview study acquire data through FTP. ICA AB did receive the data on compact discs earlier and explained that this was more of a problem back then. Therefore, it is safe to say that this problem will be less important the more technology advances.

Acquire data

None of the respondents does really have any problems concerning acquiring data from specific types of organisations. Strand et al. (2003) claim that there can be problems to acquire data from industry organisations, but none of the organisations in the study have experienced these problems since no one has tried to acquire data from industry organisations. However, the respondent from Astra Zeneca said that acquiring statistical data from an industry organisation is probably better than purchasing it from a statistical data supplier. However, the respondent explained that it is a matter of maturity in the industry and he also believes that industry organisations have a better functionality on a national level than on an international level. Sometimes, as in Astra Zeneca's case there are no large industry organisations on a national level since the industry is not large enough in that specific country.

What did come up during the study was that there also can be a problem to acquire a certain type of external data. ICA AB experiences a problem with acquiring good statistical data regarding sales of meat products. This particular problem may be seen as very specific for ICA's business area but still shows that there are certain types of data that are difficult to acquire.

A problem that GP has is that they find it hard to acquire data from their customers. Some customers log in to GP's web site and add some information, and some customers may call them on the telephone, but most customers do not have any contact with GP at all which makes it hard for GP to find patterns within their customer registry. According to the definition by Devlin (1997) data directly from the customers can not be seen as external data so this problem is not completely aligned with the aim of the work. It is still interesting and worth mentioning, since it shows that there are problems with other types of data that are somewhere between internal and external data.

Large data sets

Only Volvo has some problems with large data sets, but these problems are more related to the integration process and not the actual acquisition. Oglesby (1999) explains that this was a problem since some data sets are too big to fit onto a compact disc. On the other hand compact discs are no longer a common way to transport data. Today most organisations use FTP, which is supported in the study by Strand et al. (2004a). Using FTP means that there is no size limit in that sense, making the problem less of an issue, and it is probable that this becomes even less of an issue in the future, as Internet connections become faster, something that Oglesby (1999) also acknowledges.

Dynamic sources

Dynamic sources can sometimes be a problem according to the respondents in the interview study. It is mainly on occasions when the source changes the representation and format of the data. Volvo experiences some problems concerning acquiring industrial data from machines. Sometimes these machines change their output and then all the systems using this data have to change as well. Some respondents did work closely together with the supplier to make sure that the data the organisation receives are in the appropriate format and sometimes even custom-made for the organisation. The suppliers have to be aware of the fact that if they are dynamic, e.g. changing format or representation of the data on a frequent basis, all the organisations and systems that use this data have to adapt to the changes, which are a lot of work for the organisations. It is therefore important that there is some kind of agreement between the parts on how the data shall be structured. Astra Zeneca explained that there are standards set by the United Nations controlling the format for this in the pharmaceutical industry, something that obviously makes it easier.

Zhu (1999) discusses this problem with the focus on web sources, but since none of the organisations in the study acquires data from the web this can not be further analysed. What is clear is that dynamic sources are a potential problem also regarding other sources than the web. Syndicate data suppliers are more likely to work together with the organisations to make the acquisition less problematic though it seems.

Source stability
Only one of the respondents has experienced problems with unstable sources. Volvo has experienced some problems with systems being down or terminating transferring processes. The respondent explained that it is important to uphold the quality, and to have protocols that can handle these situations with back-up processes and recreation of data. GP explained that internal systems might go down sometimes as the data warehouse is updated but have not experienced problems with external sources.

Once again, it is obvious that as Volvo acquires data from sub-contractors have to cope with more problems than those organisations that acquire data from syndicate data suppliers. Syndicate data suppliers have to make sure that their systems are up and stabile as organisations download data, and as this is the main focus for syndicate data suppliers this is probably a high priority for them. For bi-product data suppliers who have another focus of their businesses this is probably not as highly prioritised.

Expensive
All but one respondent consider the costs involved with purchasing data as a problem. Two of the respondents, GP and Volvo, explained that the data is a part of an agreement between the organisation and the supplier, and that several other services also is a part of the deal. Astra Zeneca feels that as the supplier they acquire the data from is in a monopoly situation the costs are even higher. The respondent at ICA AB, who also feels that their supplier are in a monopoly situation, thinks that the costs involved is a problem but also realises that there might be something ICA AB as a large organisation might be able to do to influence the price.

Even though the costs involved may be high, it is apparently worth the money; this is data organisations need and is willing to pay for. As new suppliers enter the scene prices might become lower, but as long as some suppliers are in a monopoly situation these suppliers will not make any efforts in lowering the prices.

6.3 Integration

This section presents findings concerning the integration phase.

Time-consuming
Three of the five respondents find the integration process time-consuming and thinks this is a problem. GP does not have any problems in the ongoing integration activities because it is all done automatically, but these methods for automatically integrate the data still took some time to develop. Almost all of the respondents find the integration time-consuming, even though not all of them perceive it as a problem. Volvo is the organisation that experiences most of the problems regarding this, and the respondent said that they were quite bad with coping with integration issues.

Data representation and structure

Three out of the five respondents have problems with how the data is structured and represented. The two organisations that do not have problems are Preem and GP. Preem just recently started using the external data in their data warehouse and has not yet experienced any changes regarding the data representation. GP do not have any problems because all the data they acquire externally is custom-made to fit into their systems. They decide themselves how the data is to be structured. ICA AB has experienced problems in the past. However, they now acquire the data broken up in different files instead of all data in one file, which makes the data easier to handle. Both Volvo and Preem have suppliers that cannot provide custom-made data to the organisation in the same sense as e.g. GP.

Storage

None of the organisations in the study consider storage to be a problem.

System consistency

ICA AB seems to be the organisation in the study that experiences most problems regarding keeping consistency among systems. It is on rare occasions that there are some changes in the flow and dates can be unsynchronized. Volvo also has some small problems, the respondent said that the organisation is very careful concerning this and that they work hard on their quality control. Preem also experiences some small problems, but it is mainly for the operational systems and not for the data warehouse. GP explained that they are using the data warehouse as the node that feeds all the other systems.

Data completeness

Two of the five respondents consider data completeness as a problem in their organisation. Volvo and Astra Zeneca both have problems with this, Volvo explained that they sometimes need to go back to the source and acquire more data and sometimes they aggregate the data to another level and work around the problem that way. Astra Zeneca experiences that they sometimes get holes in the structure that the OLAP-tools are using. Theses holes are caused by the fact that Astra Zeneca cannot acquire the data on all the dimensional levels they are using. For example, they might acquire data with the dimension levels country-district, but in their data structure they have an additional level such as: country-region-district. This requires a lot of work according to the respondent.

Time-stamps

Some various types of problems concerning time-stamps were brought up during the interviews. Volvo has some problems on occasions and they sometimes have to call the supplier to get more information about the time-stamps and so on. The respondent at Volvo said that it is important that this work well. ICA AB experiences problems at the turn of the year; this is caused by different periodical classification made by ICA AB and the supplier. The supplier bases their classification on 13 periods while ICA AB bases theirs on weeks and months. To solve the problems that arise every year ICA AB manually adapt the data and time-stamps to their system. Astra Zeneca also has some experiences regarding this issue, they acquire their data on a monthly basis but all the data they acquire at one time has the same time-stamp. Their supplier is apparently applying an integration time-stamp approach. This means that the reports they produce with the help of the external data can only be generated once a month. It would be impossible for them to change to a weekly or daily basis; something that the respondent thinks can be a possible task in the future. The respondent do believe that it is a possibility to acquire the data more frequently but it is also a cost issue involved. The interviews shows that the main problem is not the lack of time-stamps but rather problems that concerns different approaches on how time-stamps are used.

Data overload
Volvo has problems with data overload because they sometimes acquire data that is on a very low level, making the data set much larger than necessary. Volvo then has to raise the data to the level they want themselves. An example is that they sometimes get detailed specifications on products, which is too detailed for the data warehouse. The respondent said however that as long as they acquire the data they need, they are able to solve the problems with a more hands-on approach. This issue is more due to the fact that there is a large amount of the external data that are irrelevant for Volvo rather than the data set as such is too big to handle. A conclusion that can be made is that large data sets as such are not a problem for the integration process. However, if the data set contains irrelevant data it requires a lot of time to sort out the data that the organisation needs.

Conflicting data
Only two respondents have experienced problems regarding conflicting data, ICA AB and Volvo. For ICA AB, this is a very small problem and it has to do with the organisations that deliver data to their supplier. An example from the respondent was that ICA AB has sold more of another big retail trade organisation’s own label than this retail trade organisation has sold themselves. However, this is impossible since ICA AB does not even have groceries of that label in their stores. In this example, the external data was of course conflicting with ICA AB’s internal data. Volvo also has some problems with conflicting data and they use to examine the data to see what data they should use.

The thing about this problem is that there must be data that are even capable of conflicting with each other; otherwise obviously no problems can arise. Several of the companies in the study do not have any corresponding internal data and certainly not external data from two different sources covering the same area. This means of course that there can not be any problems, but it also would mean that it is harder to find errors. If you only have one value describing something then that value is all you got to base your decision on, even though there is a possibility that the value may be incorrect. If you have two values covering the same thing, and these two differ from each other, then it is obvious that one is wrong in some way. Then the problem is to choose what data to base your decision on. I do not claim that if an organisation only acquires data from only one source it will acquire faulty data; I am just saying that there is a higher possibility that the organisation will not find potential errors within the data.

Tools
Three out of the five respondents have experienced some problems with tools in their organisation. However, the organisations in the study experience quite different types of problems concerning the tools. Volvo has just recently begin to use ETL-tools and they do not have that much of experience, but the respondent said that the tools are at least expensive. GP however has some more experience and they think that the tool they are using is too generalizing and too complex. Things that lead to the fact that GP do not fully exploit the ETL-tool. ICA AB has another problem, they have developed an ETL-tool in-house, but the problem is that there are very few persons in the organisation that know how to use it and ICA AB now investigates to see if it is worth to change to a commercial ETL-tool instead.

In the study made by Strand et al. (2004a) it came up that there are aspects that push companies to develop their own ETL-tools, especially for smaller companies. In this study however, there are only larger organisations. The size may be a factor for why ICA AB maybe will switch from their in-house developed tool to a commercial tool. ICA AB may be too big and may find it beneficial to use a commercial tool, and maybe in-house developed tools works better in smaller organisations where the IT/IS departments are not that big.

Conceptual understanding
Three respondents experience some problems with the conceptual understanding of the external data. Both Astra Zeneca and Preem experience that these problems are biggest initially and that it becomes easier along the way. Astra Zeneca explained that the problem is to get use to another organisation's way of thinking, while Preem experiences some initial problems with matching the external data to the organisation's products. To solve this, help from experts within the company that previously has been working with this particular external data was required. Volvo also has some problems and the respondent explained that there is a demand for discussions between Volvo and their subcontractors to sort out what some data really means.

The conceptual understanding issue is probably influenced of what type of data that needs to be understood. If there is data that could be seen as central for an organisation's activities, it is probably easier to understand the data since the area is well-known, but it is also more important that there are no differences conceptually between the internal and external data since the data most likely is crucial for the organisation.

Metadata
Only two of the respondents have experienced problems regarding metadata. Volvo explained that the external data is not always perfectly described and that they sometimes have to contact the subcontractor to get more information about the data. GP do not acquire that large amount of data but still acknowledge that there can be some small problems with the metadata. Astra Zeneca explained that they do not have any problems and if for some reason metadata is missing, the organisation generates the metadata themselves. Preem explained that the data they acquire is so simple to understand that metadata is not an issue. All of the organisations in the study acquire external data that focus on their main business areas, and the need for metadata that fully explains the data is not that big since the organisations already knows what the data means and have experience in handling the data from before. It is reasonable to think; as the data warehouses in these organisations grow larger and cover more aspects, that metadata becomes more important.

6.4 Usage

This section presents the findings from the interview study regarding the problems in the usage phase.

Data correctness
All the organisations in the study except GP have experienced some problems with incorrect data and what is interesting is that the problems can be referred to different phases in the process. ICA AB explained that the problems they experience on occasions are due to faulty input to their supplier, and that this is caused by the organisations that deliver data to the supplier. These organisations are using wrong codes when they register their sales. Volvo also has some problems that are caused by mistakes. Preem experiences that some data may be incorrect, but what was interesting is that they can not see if the error is caused by internal

data, external data or the human factor. However, Preem discover the errors as the data enters the data warehouse, which means that the data warehouse acts as a validation to see if the data is correct. Volvo explains that their problems often are found by the end-users, since they are the ones that work with the data. Astra Zeneca has another type of problem related to incorrect data, they experience that different applications process the data differently (such as rounding-off figures). This causes reports to appear differently in different applications even though they logically should look the same. Astra Zeneca explained that this is not caused by the external data itself but rather an application issue.

External data, and data generally do contain errors and it is good that the organisations realise that this is the case. External data generally needs to be carefully examined and it is important for the decision makers to realise that there is no guarantee that the data is 100% correct, this is obviously also an issue for internal data. The impression during the interviews was that it often is hard to find the source of the problem and, as was exemplified above, there are many possible sources for these problems. If organisations are able to find the source of the problem, they might be able to take actions to make the problem smaller. Of course, if the problem lies outside the organisation, it will be harder to take actions but at least the organisation then knows what the source is and can be prepared for similar errors from that source.

Data completeness

The impression during the interviews was that these problems are solved during the integration phase, which means that these problems do not find the way to the usage phase.

Data freshness

Three out of the five respondents experience problems with data freshness. Volvo explained that sometimes there can be technical errors at the supplier that causes this type of problem. It can for example mean that Volvo do not get any updated data, but instead the same data they got last time. This usually works however and as soon as they discover the problem, the problem is solved as they contact the supplier. ICA AB thinks this is a rather huge problem, they acquire the data every fourth week but the respondent explained that he would like to see that they instead can acquire it once a week. Preem do not experience any problems with data freshness but rather sees the positive aspects in the issue. They acquire the data on a monthly basis, and they always update the data warehouse with the values for the month the 4th the following month. They see this rather positive since it is easier for them to find errors before the data enters the data warehouse. The reason for them to do it this way is that they want to be sure that all the external data for that specific month are available in the operative systems. Astra Zeneca explained that there are some problems with data freshness associated with acquiring external data and the respondent explained that this problem will always be there.

Of course it is unreasonable to think that the external data can be just as fresh as the internal data. Maybe as technology advances this problem might be smaller and organisations can update the external data more frequently, but there is also a cost issue involved. It is obvious that the costs are related to how often the data is acquired, and organisations have to somehow evaluate how important frequently updated external data is in relation to the costs involved. It is also reasonable to think that the update frequency is related to number of problems (such as data incorrectness). If an organisation updates the data warehouse more frequently it is likely that more errors will pass through to the DW. Therefore it is interesting to see that Preem rather waits a couple of days to make sure everything is correct before the data enters the data warehouse.

Data overload

None of the respondents considered data overload as a problem, except Volvo who had experienced some problems regarding this. However, they sorted out these problems in the integration phase, which means that there are no data overload problems left in the usage phase.

Biased data

Only ICA AB has some problems with biased data, and it is because ICA AB has made different assumptions than their supplier. ICA AB and their supplier has two different hierarchies of products they are using and this obviously causes problems. It was a bit unclear though if this really is a usage problem or if the problems are biggest in the integration phase.

Data reliability

All respondents feel that the organisation trusted the external data except Volvo. Volvo explained that if something in the data looks strange, they first of all looks at the data in the file they acquired and then contacts the supplier to straighten things out. Other faults, like aggregations on an inappropriate level and so on, are often discovered later in the process and then Volvo uses a more hands-on approach and redoes the aggregations.

Zhu and Buchmann (2002) and Strand et al. (2004a) both acknowledge data reliability as a problem when the source of the external data is unknown, but the organisations in this study do not acquire any data from unknown sources, which can be an explanation for why most of the organisations trust their data.

Conflicting sources

Only two respondents experience some problems related to conflicting sources. ICA AB find that there sometimes can be small differences between the internal and external data. Volvo explained that it happens that they as the respondent put it “compares apples with pears”. The respondent from Volvo explained that they try to eliminate this problem already in the integration phase, and that a lot of thinking is required to handle the problem in a good way. It is interesting to see that organisations are working to reduce problems in earlier phases and realise that the problems in different phases will be transferred to later phases as well if the problems are not solved. As was discussed in matter of conflicting sources in the integration phase, these problems are obviously dependent on data that are able to conflict with each other. Some organisations in the study acquire data that can not conflict with internal data.

Exists without influencing the decisions

Two of the organisations experience problems regarding external data that exists in the organisation without entering the data warehouse. Volvo explained that the external data sometimes is stuck in operative systems, and never enters the data warehouse, sometimes this is caused by faults in e.g. the aggregations in which case they have to go back and see how the aggregations are set up. The respondent also explained that it is the requirements from the end-users that control what external data that is integrated in the data warehouse and which data that stays in the operative environment. Astra Zeneca has external data that they would like to integrate into the data warehouse, but in their case this is not possible due to regulations in the contract between the organisation and the supplier, which means that Astra Zeneca can not integrate as detailed data in their data warehouse as they would like to.

Restricting laws

There are some problems concerning laws that came up during the interviews. GP, who are using credit card information believes that the integrity of the customers is outmost important, and the problem the respondent sees related to laws is that it is quite time-consuming and sometimes complex to understand the laws. Volvo on the other hand thinks that most of the problems with laws are related to the internal personnel data, the external data are the data suppliers' responsibility according to the respondent. Astra Zeneca does not experience any problems with laws but instead with the data supplier contracts, that hinders them from using the data as they would like to. It was not made clear during the interview but the contracts may look the way they are because there are restricting laws regarding this data and the data supplier has to include it in the contracts. ICA AB do not have any problems; the only thing the respondent came up with was that they are not allowed to use data on a shop level when negotiating with suppliers. The respondent from ICA AB also said that if you see the laws as a problem you are somewhat criminal minded. Preem do not experience any problems as today, but they will be integrating data about customers in the future and the respondent believes that the problems will come at that point.

Laws are something organisations have to deal with and the laws will always be there. It is up to the organisations to do their lesson and get an understanding for the laws. What is a problem though is that laws often are perceived as rather fuzzy and hard to understand, and this is a problem that the authorities have to work harder on.

Ethical aspects

None of the respondents consider ethical aspects as a problem for using external data, but then again it is hard for an organisation to see their own ethical principles as a problem.

7 Conclusions and reflections

This section first gives a concluding categorisation that according to both the literature study and the interview study are actual problems for organisations today, and then some general reflections are presented.

7.1 The updated categorisation

This section gives an overview of the results from the study and in Table 2 the problems that were found valid during the interview study are presented. The problems that were excluded from the categorisation are over-stroked in the list.

Table 2 The updated categorisation

Identification
- Identify new suppliers - Establish relations - Select suppliers - Identify relevant data
Acquisition
- Time-consuming - Acquire data ~~- Large data sets~~ - Dynamic sources - Source stability - Expensive
Integration
- Time-consuming - Data representation and structure ~~- Storage~~ - System consistency - Data completeness - Time-stamps - Data overload - Conflicting data - Tools - Conceptual understanding - Metadata
Usage
- Data correctness ~~- Data completeness~~ - Data freshness ~~- Data overload~~ - Biased data - Data reliability - Conflicting sources - Exists without influencing the decisions - Restricting laws ~~- Ethical aspects~~

Compared to the list presented in Table 1 there are some changes made due to the results of the interview study. Some problems are excluded because they are seen as outdated and some are excluded because organisations do not consider them to be problems. Large data sets in acquisition, storage in integration, and data overload in usage, are all excluded because none of the respondents consider them to be problems. The data completeness problem is also excluded from the usage phase because organisations seem to handle these issues in earlier phases (e.g. the integration). Ethical aspects are also left out of the final categorisation because the organisations in the study do not consider them to be a problem for the external data usage. Notable is that no problems are added to the categorisation since no new problems were brought up during the interviews. This shows that the literature study was successful.

7.2 General reflections

In this section some general conclusions are drawn, mostly focusing on the results from the interview study. Most of the results are aligned with the results from the study among financial organisations in Sweden (Strand et al., 2004a, Strand et al., 2004b). However, the impression is that financial organisations such as banks have bigger problems with the laws, but that is not strange considering the types of data financial organisations acquire. The banks acquire a lot of data concerning individuals, such as credit information, and data on an individual level is probably the data that is controlled by laws the most (e.g. PUL, a Swedish law regulating data concerning individuals). Another thing that is related to laws is that the banks have more problems with matching external and internal data due to incomplete data. The reason for this problem is that organisations are not allowed to acquire social security numbers on persons that are not customers, and social security numbers often are used as identifiers. However, in this study no organisation experiences these problems concerning lack of identifiers, which can be explained by the fact that none of the organisations in the interview study acquire data on an individual level. However, at least one organisation has the intension of expanding their data warehouse with data about their customers as well, which most likely would mean that problems about laws and incomplete data will become an issue.

It is interesting to see that some of the problems mentioned in the literature turned out to be no problem in reality. Some problems concerning large data sets, data overload and storage were expected to be not considered as problems in the interviews and that expectation was met. Huge quantities of data may have been a problem a few years ago but nowadays the technology has improved enough to eliminate problems concerning storage and data quantity.

Another interesting thing is that when it came to solutions of problems, the manual or hands-on approach is still often used. Maybe this is because there are no tools that are flexible enough to deal with all types of problems. To manually edit data is very time-consuming, but the person who edits the data manually has full control over the data and is able to alter the data in whatever way the person wants. In my world, tools are good as long as everything work as it should, if some problems arise, the easiest and most effective way is often a hands-on approach. At least when the data set is of a reasonable size.

The impression from some interviews is that some suppliers control the market and more or less force the organisations to adjust to the suppliers' way of thinking. The impression from the interviews is that especially the suppliers in a monopoly situation can do what they like, and the organisations have to follow since there are basically no other options available.

Another thing that was made clear during the study is that the phases in the incorporation process are highly linked together in terms of problems. If a problem is not solved in one phase then it will carry on causing problems in the following phases as well. The organisations in the study seem to realise this, which was exemplified by several respondents and lead to the fact that the incorrect data problem was excluded in the updated list of problems. This might also be an explanation for why almost all the problems regarding large data sets and storage are found not to be an issue, because if the problems are solved in one phase (e.g. the acquisition) this specific issue will not cause as severe problems in the following phases.

8 Discussion

In this section, discussions are presented. First, reflections on the project are presented, then the dissertation is discussed in a wider context and lastly some ideas for future work in the area are put forward.

8.1 Reflections on the project

This section is aimed at reflecting upon the project. What did go well was that the interviews were conducted in a relaxed form with many discussions. The follow-up questions concerning each problem worked well and gave some more subjects for discussion. An experience from the interview process was that it was hard to get in contact with persons on organisations, and when you finally got the name of the person that is the most appropriate, it was hard to get in touch with him or her since these persons often are very busy. The selection of organisations and industry sectors for the interviews was well constituted and gave a good base for the analysis. Maybe it would have been better to have found some kind of list over the largest organisations in Sweden, but it would not have been possible to get in contact with e.g. the five largest anyway. It often took days and a number of telephone calls in order to get in contact with the persons. This meant that several organisations were contacted every day, and that these contacts were contacted again during the coming days if it was necessary. The problem with reaching persons was larger for bigger companies, while it was easier to get in touch with smaller companies. Something that was a bit of a failure was that the problem ratings, which each respondent answered for every problem, could not be used in the analysis. It was too hard to make sure that the respondents based their ratings on the same basis, and as certain aspects were discussed more in-depth it was hard to know what the respondents really based their ratings on. Ratings probably work better when conducting a strict survey-based study.

8.2 The dissertation in a wider context

The dissertation hopefully gives a base for further work in this rather immature area. The resulting list of problems can be used by companies to get a feeling for what types of problems that may come up when incorporating external data in their data warehouse, and especially organisations that do not incorporate external data but plan on doing so, would take benefits in acquiring knowledge about these problems. However, organisations that already incorporates external data and have problems with it, may get some ideas on solutions as they get a chance to read other organisations' views of the problem. The study can also be seen as an indicator on how far the development in the area has reached when it comes to the Swedish market. It was for example shown in the study that problems regarding storage and such are no longer an issue which points to the fact that technological advancements has made impact in the area. Another interesting aspect is whether the results are aligned with the literature originating from the United States or not. I would say that the results from this study would not be considered as surprising for researchers in the United States. The only thing that can be said is that problems related to data overload seem to be more of a problem in the United States compared to Sweden. However, the literature claiming that data overload is a problem may be out-of-date. To conclude this section, it is important to say that the results from this study can be seen relevant for the Swedish market, but it is obviously difficult to claim that these problems are universal.

8.3 Ideas for future work

Since the area is rather immature there are a lot of aspects that needs to be looked upon, presented here are some ideas for further work. The ideas that are presented in this section are:

- Survey-based study
- Focus on suppliers
- Case study
- Focus on problem solution
- Generality of the problems

Since this study only covered five organisations, it would be interesting to have a larger survey-based study that involves more organisations, obviously such a study could not have as in-depth questions as this study but it would give an enhanced picture of the status of the area.

A study with a focus on the suppliers with the aim of studying what kind of problems they experience would also be interesting. The results from such a study could be compared to this study and conclusions can be drawn to see in what aspects the suppliers and the organisations can work together to make the problems smaller.

Another possible study would be to go in-depth in one or more companies and base the study on a case study. This would give a further in-depth perspective and problems can be described in a more comprehensive way. The problem is that such a case study would not possibly cover all types of problems, so a case study should instead focus on a particular type of problem.

A big part of this study was aimed at characterising problems, where solutions were one part of this, but it would be interesting to see a study that is completely focused on finding solutions rather than problems. Such a study can use the list of problems from this study as a base for investigating what solutions there might be to these problems.

Another way the list of problems can be used is to take it and try to see if the problems are valid also for other types of systems, such as executive information systems (EIS) or decision support systems in general. Such a study would give answers to how general these problems are. It is reasonable to think that many of the problems presented in this work are not specific for data warehouses but can be problems in other types of systems as well.

To sum up, this study gives a good base for further research in the area, but there are still many issues in this rather immature area that require more research.

References

Acxiom Corporation (2000) *Transform customer knowledge into strategic advantage.* Available from Internet: http://www.dmreview.com/master.cfm?NavID=61&WhitePaperID=33 [Accessed 03.02.21].

Adelman, S. (1998) *Estimating a data warehouse pilot project?* DM Direct Newsletter. Available from Internet: http://www.dmreview.com/editorial/newsletter_article.cfm?nl=dmdirect&articleId=937&issue=1371 [Accessed 04.03.21].

Berndtsson, M., Hansson, J., Olsson B. & Lundell, B. (2002) *Planning and implementing your final year project with success!* London: Springer Verlag.

Bischoff, J. (1997) Physical design. In: J. Bischoff & T. Alexander (eds.), *Practical advice from the experts* (p. 177-198). New Jersey: Prentice Hall.

Collett, S. (2002) *Incoming!* Computerworld, Vol. 36, Issue 16, p. 34.

Damato, G. M. (1999) *Strategic information from external sources: a broader picture of business reality for the data warehouse*. Available from Internet: http://www.dwway.com/file/20020726170552_get_ext_data.pdp [Accessed 03.02.20].

Devlin, B. (1997) *Data warehouse: from architecture to implementation.* Harlow: Addison Wesley Longmann.

Eckerson, W. W. (2002) *Data quality and the bottom line – Achieving business success through a commitment to high quality data*. The Data Warehousing Institute. Available from Internet: http://www.dataflux.com/data/dqreport.pdf [Accessed 04.05.05].

Inmon, W. H. (1996) *Building the data warehouse, 2nd edition.* New York: John Wiley & Sons.

Inmon, W. H. (1999) *Integrating internal and external data.* The Bill Inmon.com Library LLC. Available from Internet: http://www.billinmon.com/library/articles/intext.asp [Accessed 04.05.18].

Kelly, S. (1996) *Data warehousing: the route to mass customization.* New York: John Wiley & Sons.

Oglesby, W. E. (1999) *Using external data sources and warehouses to enhance your direct marketing effort*. DM Direct Newsletter. Available from Internet: http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=1743 [Accessed 03.02.21].

Salmeron, J. L. (2001) EIS data: findings from an evolutionary study. *Journal of systems and software,* Vol.64, Issue 2, 87-172.

Singh, H. S. (1998) *Data warehousing: concepts, technologies, implementations, and management*. New Jersey: Prentice Hall.

Strand, M. & Olsson, M. (2003) The hamlet dilemma on external data in data warehouses. In Proceedings of *the 5th International Conference on Enterprise Information Systems (ICEIS) – Part 1*, 23-26 April, 2003, Angers, France, pp.570-573.

Strand, M. & Wangler, B. (2004) Incorporating external data into data warehouses - problems identified and contextualized. Presented at *the 7th International Conference on Information Fusion*, 18 June – 1 July, Stockholm, Sweden. (To appear)

Strand, M., Wangler, B. & Lauren C-F. (2004a) Acquiring and integrating external data into data warehouses: Are you familiar with the most common process? Presented at *the 6th International Conference on Enterprise Information Systems (ICEIS'04)*, April 14-17, Porto, Portugal. (To appear)

Strand, M., Wangler, B. & Niklasson, M. (2004b) External data incorporation into data warehouses: an exploratory study of identification and usage practices in banking organizations. Presented at *the CAiSE Forum*, 7-11 June, 2004, Riga, Latvia. (To appear)

Strand, M., Wangler, B. & Olsson, M. (2003) Incorporating external data into data warehouses: characterizing and categorizing suppliers and types of external data. In Proceedings of *the Americas Conference on Information Systems (AMCIS'03),* 4-6 August, 2003, Tampa, Florida, USA, pp.2460-2468.

Zhu, Y. (1999) A framework for warehousing the web contents. In Proceedings of *the 5th International Computer Science Conference on Internet Applications (ICSC99),* December 1999, China.

Zhu, Y., Bornhövd, C., Sautner, D. & Buchmann, A. P. (2000) Materializing web data for OLAP and DSS, In Proceedings of *the 1st International Conference on Web-Age Information Management (WAIM00),* June 2000, China.

Zhu, Y. & Buchmann, A. (2002) Evaluating and selecting web sources as external information resources of a data warehouse. In Proceedings of *the 3rd International Conference on Web Information Systems Engineering (WISE2002),* December 2002, Singapore.

Appendix 1 – Introducing e-mail

Hej.

Först av allt vill jag tacka Er för att Ni ställer upp i intervjustudien och för visat intresse. Detta e-mail ska introducera mitt projekt och förklara hur intervjustudien kommer att gå till.

Projektet utförs som ett magisterarbete (D-uppsats) för Högskolan i Skövde och syftar till att undersöka vilka problem det finns med extern data i data warehouse (datalager).

Definitionen på extern data som används i arbetet är:
"Business data (and its associated metadata), originating from one business, that may be used as part of either the operational or the informational processes of another business."

För att vara tydlig så är det alltså den externa data som på något sätt integreras och används i data warehouset som är intressant.

Intervjun kommer att gå till på följande sätt:
Jag anger ett problem som jag identifierat i litteraturen och Ni svarar på om detta är ett problem som även ni stött på i Er organisation någon gång. Anser Ni att detta är ett problem så ber jag er gradera problemet mellan 1-6 där då 1 är ett relativt litet problem medan 6 är ett stort problem. Jag ber er även ge exempel på när problemet uppstår och även hur ni brukar hantera problemet. Det kan även vara intressant för mig att få reda vilken data och vilka källor som orsakar problemet. Intervjun förväntas pågå i cirka 60 minuter.

Problemen kommer vara uppdelade i 4 kategorier:
- Identifiering av källor/leverantörer
- Inhämtning av data från källorna
- Integrering av den externa datan i data warehouset
- Användning

Det bör väl tilläggas att vissa problem kan vara av en viss karaktär som gör att dessa problem inte är aktuella för ert företag.

Efter intervjun kommer jag att skriva ihop en sammanfattning på engelska av det som blivit sagt under intervjun. Denna sammanfattning skickar jag sedan över e-mail till Er för att Ni ska godkänna det och även få chansen att korrigera eventuella missuppfattningar och dylikt.

Namn på personer och företag kommer om Ni vill, inte att användas i rapporten. Oavsett så kommer en kort beskrivning av företaget dock att finnas med i rapporten tillsammans med en kort beskrivning av den roll som respondenten (d.v.s. Ni) innehar i företaget. Men som sagt, namn utelämnas på begäran.

Hör gärna av dig om det är några frågor eller funderingar.

Med vänliga hälsningar,
Markus Niklasson

E-post: <e-post adress>
Telefon: <telefonnummer>

Appendix 2 – The interview questions

Inledande frågor

Namn?

Företag?

Roll i organisationen?

Vilka externa källor/leverantörer hämtar ni data ifrån?

Vilken typ av data hämtas från dessa källor?

Frågor rörande problem

För varje problem:

- Är det ett problem för er organisation? (ja/nej)
- Till vilken grad? (1-6)
- Kan ni ge exempel på när problemet uppstår?
- Hur hanterar ni problemet?
- Vilken data orsakar problemen?

A. Identifiering av källor/leverantörer

A.1 Har ni upplevt det svårt att identifiera nya leverantörer av extern data som nyss kommit in på markanden?

A.2 Har ni upplevt att det finns svårigheter med att upprätta relationer med leverantörer av extern data?

A.3 Har ni problem med att välja vilka leverantörer ni ska använda?

A.4 Har ni problem med att identifiera vilken av den data som de externa källorna erbjuder som är relevant för er?

B. Inhämtning av data från källorna

B.1 Anser ni att inhämtningen av data är ett tidskrävande projekt och att detta skulle vara ett problem?

B.2 Anser ni att det är svårt att få tag i data från vissa organisationer?

B.3 Är stora mängder data ett problem för er process att hämta in data?

B.4 Har ni upplevt problem med att källorna ni hämtar från är dynamiska?

B.5 Har ni upplevt problem med stabiliteten på källorna?

B.6 Anser ni att det är ett problem att externa datan är dyr att hämta in?

C. Integrering av den externa datan i data warehouset

C.1 Upplever ni att det är ett problem att integrering av extern data är en tidskrävande process?

C.2 Har ni upplev problem med hur externa datan är representerad och strukturerad?

C.3 Anser ni att lagringsutrymmet för ED i DW är ett problem?

C.4 Har ni upplevt att det är ett problem att se till att alla era system är uppdaterade?

C.5 Har ni upplevt att inkomplett data är ett problem när man ska integrera den externa datan med den interna?

C.6 Har ni upplevt problem med att den externa datan inte kan relateras till den interna datan på grund av att datan inte är tidsstämplad?

C.7 Har ni upplevt att mängden extern datan är så pass omfattande att det uppstått problem i integrationsprocessen?

C.8 Har ni upplevt att data från olika externa källor säger olika saker, alltså motsäger varandra och att detta skulle vara ett problem då ni ska integrera denna externa data med den interna?

C.9 Har ni upplevt problem med några datorbaserade verktyg som används i samband med den externa datan?

C.10 Har ni haft problem med att förstå vad den externa datan betyder och bestämma vart datan ska in i data warehouset?

C.11 Har ni upplevt några problem angående metadata för den externa datan?

D. Användning

D.1 Har ni upplevt att inkorrekt data orsakat problem med användningen?

D.2 Har ni upplevt att inkomplett extern data är ett problem för användningen av data warehouset?

D.3 Har ni upplevt att färskheten på datan orsakat problem med användningen?

D.4 Har ni upplevt att mängden extern data är så pass omfattande att det uppstått problem i användningen?

D.5 Har ni upplevt att vinklad data varit ett problem för användningen?

D.6 Har ni upplevt att trovärdighet för externa datan är låg och att detta orsakat problem när ni ska använda den?

D.7 Har ni upplevt att data från olika källor säger olika saker, alltså motsäger varandra och att detta skulle vara ett problem då ni ska använda datan?

D.8 Har ni upplevt problem med att extern data existerar i er organisation utan att påverka beslut?

D.9 Upplever ni att lagar hindrar er från att använda extern datan till vad ni vill?

D.10 Upplever ni att etiska aspekter på något sätt hindrat er till att använda den externa datan till vad ni vill?

Avslutande frågor

Har ni upplevt några övriga problem som inte redan nämnts?

Vill ni ta del av materialet som tas fram?

Tack för din medverkan!

Appendix 3 – The interview summaries

Respondent 1

Name: Hanna Konyi

Organisation: Göteborgs-Posten (GP)

Role: Responsible for the development of the customer database.

Sources/suppliers:

- MasterCard transactions. The organisation acquires data from their business partners based on MasterCard transactions. They have a travel theme for their selection of partners. This data are brought to GP from one of the biggest banks in Sweden with the help of a middle company. The data is aimed at creating added value for their customers by providing them with discounts and offers from their business partners. This data is acquired through FTP.
- MOSAIC – MOSAIC-codes are for categorizing customers into different segments based on where they live and this data is acquired from a company named Marknadsanalys. GP sends the data over to the company that provides MOSAIC and then gets updated data back and the whole process is done once a month through FTP. However, they plan to use a tool to allow them to acquire the data themselves, more or less on-demand, and do the mapping themselves. They also have used the MOSAIC codes as a basis for their analysis to for example find areas that are similar, "twin areas" as the respondent called it. If they identify areas that are profitable for the company they would like to find "twin areas" to these areas. The organisation also has made five segments of their customers that are specific for GP, which for example means that there are five different invoices that a customer might get depending on what segment you belong in. The MOSAIC codes are solely used to categorise individuals.
- Internet – They acquire some information from the customers that chose to register and give information on the Internet. Such information includes interests, however the organisation has not yet began to use this kind of information.

A. Identification

A.1 Identify new suppliers
No. This is not a problem for the organisation; they have what they need for the moment.

A.2 Establish relations
No. The respondent claimed that since they are a business in the media branch they are very careful with how this is handled. They know what bad press is, and they have spent a lot of time to check laws and rules. They have good people who handle the relation and good partners.

A.3 Select suppliers
No. Since they have this travel theme it was easy to select partners. Before they acquired the MOSAIC codes themselves they went to see how another company used it and thought at that

point that it would be something that they could find useful, and that was the way they choose that they should acquire the MOSAIC codes. The respondent said that there could be more suppliers of MOSAIC codes or the like and that this could be something that they had to choose. The organisation has an agreement with the supplier of MOSAIC codes spanning over 3 years, which is soon about to end and then it may be worth, according to the respondent to see if there is other suppliers of the same type of data available.

A.4 Identify relevant data
Yes, this is a problem according to the respondent. It is hard to acquire data from customers, since many people never call or get in contact with the organisation in any way. The respondent said that some people that have been a customer for more than 10 years never have had any contact with the company. This makes it hard to find patterns within the customer database. However, the respondent said that the journey has only begun and that they hope to make more use of the data and the data warehouse. The respondent also said that the way of thinking has been changed the last years, and that the competitive forces are stronger now than before. Problem rating: 4.

B. Acquisition

B.1 Time-consuming
No. They use FTP-scripts that automatically acquire the external data. It only takes 15 minutes to update the data. However, since the scripts are made in-house it might have been a bit time-consuming to implement them.

B.2 Acquire data
No. They have never tried to acquire data from different branch organisations.

B.3 Large data sets
No. The organisation does not acquire that large amount of external data.

B.4 Dynamic sources
No. Since the organisation gets custom-made data from their sources based on the file-specification the organisation provided the sources, the sources never change this.

B.5 Source stability
No. Sometimes the internal systems may be down, but then a simple restart solves the problem. They have not had these problems with the external data suppliers though.

B.6 Expensive
Yes. The MOSAIC codes costs 500.000 Skr every year for the organisation. The MasterCard transaction data is part of an agreement that the organisation has with a large bank, and this agreement itself is expensive. However, the MasterCard transaction data is not influencing the price in that sense; it is more of a package-deal. Problem rating: 3.

C. Integration

C.1 Time-consuming
Well, it is not a problem since it goes automatically. However, the scripts and so on take some time to develop. Problem rating: 2.

C.2 Data representation and structure
No. Since the data is custom-made to fit the needs of the organisation this is not a problem. The organisation provided a file specification to the supplier to tell them exactly how they wanted the file to look like.

C.3 Storage
No. There is not a huge quantity of external data, and the systems are of high performance.

C.4 System consistency
No. The customer database (the data warehouse) is the system that feeds all other systems that uses the external data, for example a tool used for analysis.

C.5 Data completeness
No.

C.6 Time-stamps
No.

C.7 Data overload
No.

C.8 Conflicting data
No. The different external data does not cover the same areas.

C.9 Tools
The organisation used their own developed FTP-scripts to acquire the data and they also used ETL-tools to some extent. However, they did not fully exploit the ETL-tools because they thought that the tool they had from Oracle were too generalized and too complex to use. They had this Oracle tool available since they have a license agreement with Oracle which is why they used that particular ETL-tool. Problem rating: 3.

C.10 Conceptual understanding
No. Before the organisation developed their data model they interviewed 40 of the employees to find out what needs they had, and these interviews was the base for the model. All of the external data is located in separate tables in the database but they never experienced any problems in understanding what the external data meant or where the external data should fit into the data warehouse.

C.11 Metadata
The organisation does not acquire that much data externally which reduces the problem. The respondent thought that they got good metadata from the suppliers. Problem rating: 2.

D. Usage

D.1 Data correctness
No. The respondent has not got any indications from the any of the end-users that this is a problem. If there is any problems it is because of the queries to the database are put wrong or something like that.

D.2 Data completeness

No.

D.3 Data freshness
No.

D.4 Data overload
No.

D.5 Biased data
No.

D.6 Data reliability
No, the organisation does not incorporate such critical data that this is an issue.

D.7 Conflicting sources
No.

D.8 Exists without influencing the decisions
No, the organisation has full control of where their external data is located.

D.9 Restricting laws
Well, some MasterCard information could be sensitive information on an individual level. However, the organisation has spent lots of time to understand the laws and PUL and the respondent thinks that it has been more of a problem and time-consuming to fully understand the laws. Problem rating: 2.

D.10 Ethical aspects
The organisation has very strong ethical principles and is very careful of their customers' integrity. The respondent claimed however that it is hard to make such a judgment when they do not know anything else.

Other problems
No.

Respondent 2

Name: Göran Frank

Organisation: Volvo IT

Role: Systems developer, involved in projects concerning business intelligence and data warehouse.

Sources/suppliers:

- Subcontractors. Volvo acquires external data from their subcontractors. This data concerns mainly products and is at first integrated into the operative systems at Volvo before entering the data warehouse. No external data enters the data warehouse directly. Sometimes this data is acquired through a file that the subcontractor sends to

Volvo, and sometimes Volvo enters the subcontractors system and extracts the data they need that way.

- Address updates. Volvo IT in Skövde also acquires address updates from a central repository in Gothenburg. This central repository most likely uses some supplier of address updates. The respondent did not have a good knowledge about that specific part.

A. Identification

A.1 Identify new suppliers
Yes, it could be a problem in finding appropriate subcontractors that can provide Volvo with good data. Problem rating: 4.

A.2 Establish relations
Yes, Volvo is a highly developed organisation working a lot with business-to-business (B2B). However, not all of the subcontractors have gotten that far in their developments. Therefore, Volvo has to have high demands on the subcontractors. However, the respondent also explained that you have to deal with the fact that most subcontractors can not uphold the same high standards as Volvo. Problem rating: 3.

A.3 Select suppliers
Yes, this could be a problem. It is mostly about data quality, sometimes the subcontractors can not provide Volvo with the high quality data that they need, which means that Volvo change and use another supplier instead. It is also a matter of size of the company; Volvo ultimately wants to acquire data from a supplier that can provide data from many areas and as fully-covering as possible. If the supplier is too small they might not be able to handle the delivery of the data to Volvo. Problem rating: 2.

A.4 Identify relevant data
Yes, this could be a problem. Volvo often has to sort out what data they find relevant, and this is often done in cooperation with the subcontractor. Data such as quantity and product number is often relevant. Problem rating: 3.

B. Acquisition

B.1 Time-consuming
Yes, this could be a problem. However, it is more that the process is time-consuming than it is complex. Volvo has very high security standards regarding this process, and sometimes the subcontractors have to adopt Volvo's methods to handle this. This is all about communication according to the respondent, and sometimes it does not work well. Problem rating: 4.

B.2 Acquire data
Well, the respondent explained that it is hard to extract data from SAP system that Volvo use. However, this system is internal and the respondent could not think about any problem they have had where it has been hard to acquire data from a specific organisation. The respondent adds though that they acquire some data from branch organisations indirectly from the head-office, but he could not answer if this was a problem or not.

B.3 Large data sets

Yes, sometimes the files are too big, occasionally due to the fact that much of the data they acquire might include data that is not relevant for Volvo. This means that they have to sort out the relevant data, and this process is rather time-consuming. Problem rating: 4.

B.4 Dynamic sources
Yes, this could happen. Sometimes they extract data from machines and so on, data that the respondent referred to as industrial data. And sometimes these machines change their output, which means that the other systems that want to integrate the data also need to be changed. Problem rating: 3.

B.5 Source stability
Yes, this could be a problem which makes it more important to have highly developed quality control, such as back-up, good protocols, and the chance to recreate data. Basically techniques and routines that make it more secure if an external source goes down during the process. Problem rating: 2.

B.6 Expensive
Yes, since they have high demands on the subcontractors it could be rather expensive. However, it is more that the resources that are being used during the process than the cost of the actual data that makes it expensive. The data is part of an agreement between Volvo and the subcontractor. Problem rating: 3.

C. Integration

C.1 Time-consuming
Yes, the respondent claimed that Volvo IT is quite bad in this part, which demands a lot of time and resources. The respondent would like to see the external data more closely integrated to the internal data and thus make it easier to compare the external and internal data. Problem rating: 5.

C.2 Data representation and structure
Yes, the data Volvo acquires is on a low level and Volvo has to aggregate the data themselves to an appropriate level, so the data acquired from the suppliers is not custom-made for Volvo's systems. An example is that they sometimes get detailed specifications for products which they have to aggregate up to the level they want. However, the respondent said that as long as they get the data they need, it is no problem for them to aggregate the data, even though the process can take up some time. Problem rating: 2.

C.3 Storage
Volvo has very large storage devices so this is not a problem.

C.4 System consistency
Volvo is very careful in this part and have as said before high quality concerns. There is a lot of work with this quality control however. Problem rating: 2.

C.5 Data completeness
Yes sometimes Volvo does not get all the data they need. The aim is to find a common identifier to allow them to integrate internal and external data, and sometimes they have to aggregate the data to another level to work around the problem. However, Volvo sometimes

also discusses with the supplier and acquires more data from them to allow them to solve the problem. Problem rating: 3.

C.6 Time-stamps
Yes this is very important that it works. Sometimes Volvo has to contact the subcontractor to be able to work out any problems with time-stamps. Problem rating: 4.

C.7 Data overload
Yes, this could be a problem. Volvo is however used to deal with large data sets, but this was a huge problem in the beginning though. Problem rating: 2.

C.8 Conflicting data
Yes, this is sometimes a problem. To work around it, Volvo looks at the data and tries to figure out which to use. Problem rating: 3.

C.9 Tools
Volvo just started to work with ETL-tools, so they did not have any experience of it yet. The respondent did however acknowledge the fact that they are very expensive and the rating reflects that. Problem rating: 5.

C.10 Conceptual understanding
Yes, this is a problem and to find a solution Volvo has to have a discussion with the subcontractors. For example, when they acquire prices, in which time intervals is the price valid? Problem rating: 4.

C.11 Metadata
Yes the respondent experience that some of the metadata they acquire from external sources is not perfectly described. An example is the format for how attributes are described, some use numbers and some use characters, but Volvo needs more help to figure out how these two relate to each other and to do that they often need to contact the subcontractor that provided the data to get a better explanation. Problem rating: 2.

D. Usage

D.1 Data correctness
Yes, but this is more based on mistakes. The respondent also said that these could be hard to find, but the end-users are often the ones who find these problems due to abnormal results and so on. Problem rating: 4.

D.2 Data completeness
Yes, it could be a problem sometimes, but Volvo sees to it that the data is complete when the data enters the warehouse. He explained that they take the blast regarding this problem earlier in the process. Problem rating: 1.

D.3 Data freshness
Not a big problem, it usually is fresh data that enters the warehouse. However, on occasion some problems arise at the suppliers which mean that they acquire the same data over and over. As soon as they discover this it usually sorts out however. Another problem could be that the server goes down which means that they can not get any updated data. Problem rating: 2.

D.4 Data overload
No, not a big problem in the usage phase, they aggregate the data and sort out the data they find relevant. Problem rating: 1.

D.5 Biased data
No.

D.6 Data reliability
Yes, this may cause trouble, but if something looks abnormal they look at the files they acquired from the subcontractors or they might discuss with the subcontractors. There could also be a problem due to the fact that they aggregated to the wrong level. Problem rating: 3.

D.7 Conflicting sources
Yes this is a problem. It is easy to compare apples and pears according to the respondent. Volvo tries to solve these problems before the data is used. There are a lot of mental activities involved to be able to cope with this. Problem rating: 5.

D.8 Exists without influencing the decisions
Yes, sometimes the external data stays in the operative systems and never enters the data warehouse. However, this is often based on the end-users requirements, if they need something they look into the operative systems to see if they have it available. Other types of problems concerning this could be that an aggregation sorts out data that is supposed to enter the warehouse, so they have to look at the aggregations to find errors. Problem rating: 4.

D.9 Restricting laws
Well, the respondent explained that the subcontractor has most of the responsibility in this case; they have to look up what they are allowed to send and so on. However, the respondent explained that maybe some problems lie within how the internal personal register is to be used. Problem rating: 1.

D.10 Ethical aspects
Sometimes the subcontractors do not share some of the data due to company principles. So the respondent said that this is more of an issue that concerns out-put companies that provides data to other organisations. Problem rating: 2.

Other problems
Sometimes the suppliers can not give the data that we need, if they only provided the data to us they would have reduced the number of questions that we instead have to ask them.

Respondent 3

Name: Martin Randler

Organisation: ICA AB

Role: Manager IT Retail Sweden

Sources/suppliers:

- AC Nielsen. ICA acquires statistic data on an article level concerning sales from competitors. This data includes business ratio for products. They acquire this data through FTP.

A. Identification

A.1 Identify new suppliers
Yes, this is a problem. There are no alternatives because AC Nielsen is in a monopoly position in this area. There have been some attempts from other companies but no one has of yet come up to AC Nielsen's standard. ICA has an advantage in that they are a big company. Problem rating: 2.

A.2 Establish relations
No, this is not a problem. AC Nielsen is very caring and also able to make changes to adapt to ICAs demands.

A.3 Select suppliers
No, since AC Nielsen is in a monopoly position it was not a difficult choice.

A.4 Identify relevant data
Yes and ICA needs to be confident about the quality of the data. Statistical deviation is something that they would like to acquire. The respondent does believe that this will cause more problems later on. Problem rating: 4.

B. Acquisition

B.1 Time-consuming
No, this acquisition is made automatically through FTP. Sometimes there is a change in the flow for some reason, but that has only happened once or twice in two years. To adapt to a flow change some work is required. Before they started using FTP they acquired the data from a CD which was a little more work. Problem rating: 3.

B.2 Acquire data
Yes, there is a problem in acquiring good statistics about for example sales of meat products. Problem rating: 5.

B.3 Large data sets
No.

B.4 Dynamic sources
No, the respondent said that it was because of the fact that they have thought about it from the start, and adapted the database tables and so on to be able to deal with these types of situations. AC Nielsen now sends three different files and not only 1. These 3 files contain different information. For example, one of them contains facts, and one contains business ratios. ICA and AC Nielsen has a worked together to solve this, and both did their part to make it work.

B.5 Source stability
No.

B.6 Expensive

Yes, this could be seen as a problem. ICA is a big company and might be able to do some actions to make the problem smaller. Right now they have an agreement with AC Nielsen and pay a fixed amount of money for this. Problem rating: 3.

C. Integration

C.1 Time-consuming
No, they are prepared and it pretty much goes automatically. However, to start up and making this work was rather time-consuming.

C.2 Data representation and structure
Yes, this is a problem. ICA has however made it easier by acquiring three files instead of just one from AC Nielsen. They also mentioned that it may be a problem to handle historical changes. Problem rating: 5.

C.3 Storage
No.

C.4 System consistency
Yes, this might be a problem on occasions. It is related to if something is changed in the systems and for example dates may appear unsynchronized. It does not happen often but when it does, it is a big problem. Problem rating: 6.

C.5 Data completeness
No.

C.6 Time-stamps
Yes, this is a problem at the turn of the year. AC Nielsen uses another type of periodical classification than ICA. ICAs classification is based on weeks and months while AC Nielsen's is based on 13 periods. ICA solves this problem manually. Problem rating: 2.

C.7 Data overload
No, this is not a problem but it requires a certain amount of experience to be able to normalise the data, set the index correctly and so on. It is basically about having skilful database people.

C.8 Conflicting data
No.

C.9 Tools
ICA uses in-house developed ETL-tools but the problem is that there are few persons that know how to handle it. They will look into this issue and see what they will do about it, maybe they will chose a product instead. Problem rating: 6.

C.10 Conceptual understanding
No.

C.11 Metadata
No.

D. Usage

D.1 Data correctness
Yes, sometimes. An example is data that are saying that ICA has sold more of COOPs own trademark than COOP themselves. Obviously this is false since ICA does not even have these products in their stores. This is due to faulty input to AC Nielsen, in other words some suppliers that made mistakes with code representation. Problem rating: 1.

D.2 Data completeness
No. The data they receive is not incomplete. However, some product groups are missing.

D.3 Data freshness
Yes, this is a problem. ICA receives fresh data once every fourth week, and they would like to have it once a week which they might change to in time. Problem rating: 5.

D.4 Data overload
No.

D.5 Biased data
Yes regarding product hierarchies. ICA has their version, and AC Nielsen has their version, and these two does not match. Problem rating: 6.

D.6 Data reliability
No.

D.7 Conflicting sources
Yes, sometimes there can be some kind of small discrepancy between the external statistic data and the internal data. Problem rating: 1.

D.8 Exists without influencing the decisions
No.

D.9 Restricting laws
No. Well, they are for example not allowed to use information on a shop level during negotiations with suppliers, but to see the laws as a problem is to have somewhat a criminal’s mind.

D.10 Ethical aspects
No. The only thing could be that ICA is supporting a monopoly situation, but that is not a problem for the organisation like that.

Other problems
No.

Respondent 4

Name: Björn Jäderberg

Organisation: PREEM

Role: Database administer

Sources/suppliers:

- Oil prices. PREEM acquires price notations regarding oil once a day at 5 AM automatically. They acquire this data both from New York and from IPE in Chicago. PREEM acquires the data through FTP from a supplier named Plats where PREEM has their own directory where they can log in and acquire the data. The oil prices enter the operational systems at first, and then once a month this data enters the data warehouse. This data is stored in separate tables in the data warehouse and is used by PREEM to allow them to see what the costs are for purchasing oil at a specific time.
- Exchange rates. PREEM acquires currency exchange rates from SEB regarding dollars, Euros and pounds. This data is acquired through FTP as well, but in this case SEB sends the data to PREEM. The data is also stored in operative systems at first and then is entered into the DW in separate tables. There are three tables that handle this type of external data, and one of them is updated as new values are entered.

A. Identification

A.1 Identify new suppliers
No. PREEM does not really check for new suppliers. And they know where to look regarding acquiring the data they need. It is easy because they acquire data that basically is focused on their main business area. When it comes to new suppliers in the market there might be suppliers that are cheaper and so on, but the respondent had not seen any advertising regarding this.

A.2 Establish relations
No.

A.3 Select suppliers
No. It has been clear from the start.

A.4 Identify relevant data
No.

B. Acquisition

B.1 Time-consuming
No.

B.2 Acquire data
No. The respondent said that they have not tried to acquire data from business organisations.

B.3 Large data sets
No. The data they acquire is not that large.

B.4 Dynamic sources
No. Ever since they started to acquire the data through FTP in the turn of the year they have not experienced any problems.

B.5 Source stability

No.

B.6 Expensive
The respondent did not know the actual costs; however he did not believe that this was an issue.

C. Integration

C.1 Time-consuming
No.

C.2 Data representation and structure
No. It has not as of yet been any changes regarding identifiers and so on from the external sources.

C.3 Storage
No. The external data is a very small part of all the data that is in the data warehouse.

C.4 System consistency
Well, sometimes this is a problem but not for the data warehouse. It is more of an issue in the "ordinary systems", the operative systems. This is mainly due to the manual insertions to the databases. Problem rating: 1.

C.5 Data completeness
No.

C.6 Time-stamps
No. The external data is always time-stamped so this is not a problem.

C.7 Data overload
No.

C.8 Conflicting data
No.

C.9 Tools
No. They use Microsoft SQL Server which is one of the least expensive in the market. They did an evaluation before they decided to use it.

C.10 Conceptual understanding
When the data warehouse started up it were some problems regarding matching external notations and internal products. Experts in the area were able to help; since notations always have been used in the organisation, knowledge about them was available among experts. Problem rating: 2.

C.11 Metadata
No. The data is so simple that this is not an issue.

D. Usage

D.1 Data correctness
Sometimes problems like this can arise. The respondent did not know if it was due to the external data or due to human errors as faulty inputs. It is an easy mistake to put 30 instead of 3.0 into the system. However, this faults are often noticed when the data enters the data warehouse. Problem rating: 3.

D.2 Data completeness
No.

D.3 Data freshness
No. Four days into every month the data warehouse is updated with fresh data. This is to make sure that all of the data from the external sources are available in the operative systems, and this approach also makes it easier to find errors in the data before it enters the data warehouse.

D.4 Data overload
No.

D.5 Biased data
No.

D.6 Data reliability
No.

D.7 Conflicting sources
No.

D.8 Exists without influencing the decisions
No. There are however data that the respondent would like to see in the data warehouse that may influence decisions and so on. The next step for PREEM is to integrate customer data into the DW, using Upplysningscentralen (UC), branch codes and so on. The respondent do believe that they may experience more problem in the future as the data warehouse grows.

D.9 Restricting laws
No. Not any such problems today with the data warehouse. Especially since they do not have any customer data in the DW, maybe problems will arise in the future regarding PUL etc. when customer data is included in the DW.

D.10 Ethical aspects
No.

Other problems
No.

Respondent 5

Name: Tor Björkman

Organisation: Astra Zeneca

Role: Responsible for business intelligence capability in the IS operation and business service section.

Sources/suppliers:
- IMS. The company acquire market information regarding the pharmaceutical industry. This data covers the sales for all medicines. A contract is upheld between the two parts. The company also acquires data from a third-part company that first acquires the data from IMS, and then modify the data in various ways, such as cleaning, to obtain a higher quality. Not all of this data enters the data warehouse because they are not allowed to spread the data in the organisation in the way they would like to, as this is controlled by the contract. The respondent was unsure of how the data was transferred between the supplier and the organisation.

A. Identification

A.1 Identify new suppliers
Yes, this is a constant work where the organisation looks at possibilities. IMS is in a monopoly situation at the moment, but there are attempts by other companies to enter the market. The organisation would like to have more flexible contracts with the suppliers. Problem rating: 4.

A.2 Establish relations
No. The only thing could be that there are some cultural differences between Sweden and United States. Most of the contacts at the suppliers are based in United States and one thing is for example that IMS are aware of the monopoly situation and may take advantage of it sometimes.

A.3 Select suppliers
No.

A.4 Identify relevant data
No.

B. Acquisition

B.1 Time-consuming
No. They have a good established relation with the suppliers.

B.2 Acquire data
It is a matter of maturity in the industry. The organisation would rather go to an industry organisation to acquire the data than buying it from IMS, but that is not possible as today. The industry organisations work best at a national level. However, the pharmaceutical industry is not very large in Sweden, and the industry organisations are international. It is also a matter of trust and economical issues. Problem rating: 3.

B.3 Large data sets
No.

B.4 Dynamic sources

Some structures are stabile because they are regulated by UN standards. However, it is a tough work to keep track of all the new medicines, packaging and medicine strengths. Other things that could make it problematic are to know what dates the medicines are active and generic manufacturing and parallel trade caused by expired patents. Problem rating: 4.

B.5 Source stability
No. They acquire the data on a monthly basis. As soon as the data is updated at the supplier the organisation gets a message.

B.6 Expensive
Yes, there are very few actors in the area and little competition which lead to high costs. Problem rating: 6.

C. Integration

C.1 Expensive and time-consuming
Yes, this is a time-consuming process and it is also quite complex. You have to be accurate when you deal with the integration to make the mapping correct. It is a problem since the organisation has their data model with dimensions and then another organisation thinks in a different way. Problem rating: 5.

C.2 Data representation and structure
Yes, the supplier has their format and the organisation does not have that much influence on the format. It was quite recently the organisation started to integrate the external data into their data warehouse and by that had to adjust the external data to fit into the data warehouse model. Until recently they had only worked with the standard that the data are in when they acquire it and adjusted the systems to that standard. Problem rating: 4.

C.3 Storage
No.

C.4 System consistency
No.

C.5 Data completeness
Yes, this is a problem if you for example work with OLAP-tools. There are some holes in the structure because they do not acquire data from all the dimension levels they use themselves. An example could be that they acquire data with the levels country-district and then they have the levels country-region-district. In this case there are no data that can be referred to the region level in the dimension. There is a lot of work with this. Problem rating: 2.

C.6 Time-stamps
The company buy the data once a month and all of the data they acquire that month has the same time-stamp. This means that the reports the organisation uses has to adapt to this and they can not make these reports on a weekly or daily basis, but have to stick to a monthly basis. This could be a problem further on if the organisation would like to generate new reports. However, the respondent did not perceive this as a problem as today. The organisation might be able to acquire the data more often to be able to generate these reports in the future, but it is also a cost issue.

C.7 Data overload
No.

C.8 Conflicting data
No.

C.9 Tools
No. The company uses two different ETL-tools but did not experience any problems regarding the tools.

C.10 Conceptual understanding
Yes. It takes some time to get used to another organisation’s way of thinking. This is mainly a problem initially and becomes less of a problem when you get used to it. Problem rating: 2.

C.11 Metadata
No. On occasions when metadata has been missing the organisation has generated it themselves, but this was not a problem.

D. Usage

D.1 Data correctness
Yes, the data itself is not a problem. However, different applications handles the data differently, an example could be rounding-off problems concerning currency. This means that the same report can appear different when generated in two different applications. Problem rating: potentially 6.

D.2 Data completeness
No. There is a department in the organisation that is responsible for generating the reports based on external data. They have knowledge about the problems and adjust to the data quality.

D.3 Data freshness
Yes, since the data is acquired on a monthly basis there are some falling behind associated with this. However, there will always be this kind of problem and of course the organisation would like to acquire the external data concerning sales at other organisations at the same time as they can acquire the internal sales data but that is impossible. Problem rating: 1.

D.4 Data overload
No.

D.5 Biased data
No.

D.6 Data reliability
No.

D.7 Conflicting sources
No.

D.8 Exists without influencing the decisions

Yes, the organisation would like to fully integrate all of the external data from IMS. There are some problems concerning contracts though and the data that enters the data warehouse is on a quite high level, and not so detailed. Problem rating: 3.

D.9 Restricting laws
No. Laws are generally not a problem, except for parallel trade. Restrictions in the contracts may be a problem though.

D.10 Ethical aspects
No.