Incorporating External Data into Data Warehouses – Problems Identified and Contextualized

Mattias Strand
Information Systems Engineering Research Group,
University of Skövde, Box 408
SE- 541 28 Skövde
Sweden
mattias.strand@ida.his.se

Benkt Wangler
Information Systems Engineering Research Group,
University of Skövde, Box 408
SE- 541 28 Skövde
Sweden
benkt.wangler@ida.his.se

Abstract - *Incorporating external data into data warehouses is a rather unexplored area and deserves more attention, especially if considering the resources organizations spend on acquiring data from specialized data suppliers. Therefore, this paper gives the results of an interview study partly aimed at evaluating, from a general perspective, the outline of such a process and related problems. The results show that trust and cost issues are utmost important characteristics to consider when acquiring external data. In addition, the results also show that the organizations incorporating external data need some type of hands-on support, for being able to fully exploit the potential thereof. Finally, the results also gave the outline of the external data incorporation process.*

Keywords: Data warehouse, external data, external data incorporation process.

1 Introduction

For most organizations, the whereabouts in the environment is nowadays equally (or more) important as the internal performance and therefore, they need systems that may assist in keeping them updated about their environment. One such system, which may assist in the tasks of understanding the environment, is the data warehouse (DW) ([9], [12], [4]). A DW is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decision ([9], p 33). DWs are designed to support the aggregation and integration of data from various internal, as well as external, data sources.

Still, although many authors highlight the importance of and abilities to incorporate external data into DWs (e.g. [9], [10], [5]), foremost due to the new demand stated above, the literature describing such incorporation is limited. In addition, the little amount of literature does not give the problems of such incorporation. Instead, there is a tendency to only address the topic from an opportunistic point of view. Therefore, in this paper, partly results of an explorative interview study among DW consultants are presented. Other results of the study are found in [16] and [17]. The part of the interview study accounted for in this paper was aimed at exploring how external data is incorporated and also to identify and describe the most common problems. Furthermore, as the problems identified ought to be put in a context and the fact that literature lacks details on the process of incorporating external data into DWs; the external data incorporation process is outlined.

The result of the interview studies showed that the process of incorporation external data may be divided into the following four activities; 1) identification, 2) acquisition, 3) integration, and 4) usage. The activities chosen are not solely unique for external data. However, since the external origin of the data implies issues such as less control of the data, difficulties in ensuring the quality of the data, the associated costs when purchasing the data, integration problems due to disparate data structures, and problems of conceptually mapping the external data with the internal data, it deserves more attention.

The paper is outlined as follows. In Sec. 2, related works are presented and external data is defined. Details of the interview study are given in Sec. 3. Sec. 4 includes the evaluation of the process, from a problem-perspective. In Sec. 5, the problems are put into context. The paper is concluded with a discussion in Sec. 6.

2 Related works

As indicated in previous section, the literature covering external data incorporation into DWs is rather fragmented. From a general DW evaluation perspective, there are much more literature to be found and many different topics are covered, for example, view selection ([18]), distributed DWs ([2]), and evaluations of DWs initiatives in particular contexts ([14]). Unfortunately, evaluation of external data seems to be an ignored topic, since no references were found that makes such an effort. This is somewhat surprising, since organizations are spending a lot of money on incorporating external data, but the natural explanation may be that such data still is considered as a means to increase competitiveness and therefore nothing that is to be presented in public. As will be shown later on, this also affected the selection of respondents for the interview study.

With respect to outlining external data incorporation, Damato [5] gives the most complete description, as he elaborates upon important aspects related to external data. However, Damato [5] has excluded some aspects that will be shown as important later on in this paper. Foremost, he does not give any problems related to the incorporation of

external data and thereby contributes to the general opportunistic view on such incorporation.

According to Singh [15], one of the clear goals with a data warehouse is to free the data that is locked up in the internal, operational databases and to mix it with data from external sources. Singh [15] further advocates that organizations should increase their acquisition of data originating from outside their own system boundaries. Generally, this includes some level of market-share information and could for example be information that includes economic forecasts, political information, consumer demographics, and competitive and purchasing trends. Inmon [9]) is more precise on the contribution of integrating external data into data warehouse as he claims that even though external data doesn't say anything directly about a company, it may still give a lot of valuable information about the universe that the company must work and compete in. When comparing internal data to external data one of the most useful things is the ability to perform comparisons over a period of time. The comparison allows management to "see the forest for the trees" ([9], p. 272).

What is then external data and what makes it external? For definition purposes, two notions will be presented and discussed. Firstly, Kimball [12] refers to syndicate data which are data purchased from data suppliers, such as A.C. Nielsen, IRI, and IMS. Secondly, Devlin ([6], p. 135) suggests the following definition of external data:

> *"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."*

If comparing the definitions, it is shown that the definition given by Devlin is more general and his definition may also include data acquired from e.g. business partners and from the Internet. Kimball's description is narrower and only relates to those commercial organizations specialized in compiling and selling data. Since the results of the interview study showed that most of the data incorporated was bought from commercial organizations, i.e. external data suppliers, the focus of this work will be directed towards the incorporation of such data.

Of course, some organizations may apply for more than one supplier type and there may also be organizations which are hybrids or combinations of these suppliers. In literature (e.g. [12]) and from the results of this study presented in [17] the following sources have been identified: statistics institutes, syndicate data suppliers, industry organizations, county councils and municipalities, the Internet, business partners, and bi-product data suppliers.

3 The interview study

The material for the evaluation was collected through interviews. The following steps were used to guide the preparation and accomplishment of the interviews:

1) *Selecting the respondents*. Consultants were considered as most appropriate for an exploratory study such as this, since consultants have often been working in many different projects and therefore have a broader and more general knowledge on the topic, as opposed to personnel in specific organization applying the external data in their daily work. In addition, as indicated previously, external data is still a private matter. By interviewing consultants, problems were identified without any particular organizational reference. The respondents were selected by searching for consultant companies on the Internet. Varying company size, geographical distribution, and a long experience were considered as important parameters for us to be able to acquire as general knowledge as possible. This resulted in a respondent group with at least 3 years of DW experience, which was geographically distributed and representing both large and internationally well-known consulting firms and national companies with 5-10 employees.

2) *Constructing the interview questions*. The set of interview questions was split into three groups. The first group was introductory and aiming to collecting background material, and to get the interview started. Furthermore, this group included questions aimed at verifying that the respondent shared the definitions of external data and DWs chosen for this work. The second group was the main part of the interview and aimed at collecting material for the analysis. Finally, the last questions aimed at letting the respondents introduce additional ideas and aspects not covered by the main questions. The analysis accounted for in this paper is based on material from group 2 questions.

3) *Initiating the interviews*. The interview questions were sent to the respondent in advance so as to let them read through the questions and reflect upon them before the actual interview. In this way, it was assumed that the respondents were better prepared when the interview was done. This approach also had another advantage, related to the discussion on how to define important concepts. Both external data and DWs are given various definitions in literature. Since this may also be the case among practitioners, it was considered as important to certify that the respondents shared the definitions chosen for this work. If the respondents were having different definitions the interview results might not have been comparable and not possible to use in the analysis. A personalized cover letter accompanied the interview questions, the aim of which was to explain the purpose of the study and to guarantee the confidentiality of the collected data, and to explain how the material was to be compiled and validated.

4) *Conducting the interviews*. Every interview lasted for approximately 45 minutes. After the interviews, the answers were written down. The follow good academic practice and to validate the answers, the compiled material was sent back to the respondent for reviewing and authorization. In this way, errors in the material were avoided and misinterpretations were corrected. The result of the interviews gave answers from 12 respondents. The final response rate from the scheduled interviews was 86% (12 out of 14). The missing two respondents were scheduled, but at the time for interviewing them, they

were unavailable. After repeated, but unsuccessful, attempts to contact them, they were excluded.

4 The problems of external data incorporation into DWs

In this section, the empirically identified problems of external data incorporation will be given. The interviews gave rather detailed information on the problems of incorporating external data into DWs. Below; the most common problems identified are listed:

- Ensuring the quality of the external data
- Making the users trust the external data
- Physically integrating the external data with the internal data
- Conceptually map the external data with the internal data
- Identifying possible external data sources/suppliers
- The external data is expensive

The problem which was most frequently mentioned was the difficulty to ensure the quality of the external data. The consultants claimed that the data incorporated from the syndicate data suppliers were most trustworthy. The high degree of trust put on these suppliers very mostly based on the fact that these organizations had been around for quite some time and the consultants new by experience that the data incorporated from them are in alignment with contracts or other agreements. Most of the consultants also mentioned the Internet as a possible source, but the quality of data incorporated from the Internet were considered as highly questionable and the consultants therefore considered the resources spent on acquiring such data as wasted money. The problem of ensuring the quality of the external data acquired also gave transitive problems. Since the data was not considered as trustworthy, the users in the organizations very not very keen on applying the external data as a basis for decision support. As a result, organizations spent rather large amount of money on buying and acquiring data that were not used.

In conjunction, most of the consultants claimed that the problems of integrating external data with the internal data were one of the key reasons why organizations hesitated when incorporating external data. The consultants gave the impression that it was difficult to establish approaches for how to integrate the external data with the internal data, if it was to be integrated at all. Some consultants claimed that it was not advisable to integrate the external data on a schema level, due to the quality issues discussed above. Instead, they kept the external data separated from the internal data and only "integrated" the data on a spread-sheet level. This is also in alignment with the ideas of Damato [5], as he claims that the "spread-sheet integration" is an applicable starting point when wanting to be able to compare the internal data with external data.

On a usage level, the consultants claimed that it is very difficult to conceptually relate the external data to the internal data. This problem also made it difficult to achieve a common understanding that external data incorporation is important. Especially since some of the consultant claimed that the users hesitated on the data quality issue from the start, and in addition they questioned how the external comparison data was calculated. Often, different organizations have different approaches on how to calculate and apply key measures, making it difficult and not appropriate to compare these numbers. Still, the conceptual mapping problem is only relevant when the user is to compare internal data with external data. Other applications for external data were also exemplified upon. For example, some of the consultants claimed that they had worked in projects, in which external data is used to update customer data, e.g. addresses. In such case, the internal data may be seen as the dirty laundry, the warehouse as the washing-machine, and the external data as the soap.

The consultants also claimed that it were difficult to identify external data sources. If excluding the providers which Kimball [14] refers to as syndicate data suppliers, it was considered as difficult to identify the suppliers. In conjunction, even if considering the syndicate data suppliers, the consultants gave the impression that there are some well-known companies which most consultants are familiar with and from which most of the external data are bought.

The external data were also considered as rather expensive Many DW projects are consuming rather large amounts of resources already on internal data matters, making it even harder to justify an additional inclusion of external data that is rather costly and may not be trusted. Still, all consultants claimed that the incorporation of external data is becoming more and more important and that the amounts of that data will be incorporated are growing exponentially, since it gives the organizations totally new ways of viewing the business and to compare the business to its environment and the actors constituting the environment. Finally, all consultants also claimed that the organizations trying to incorporate external data into their data warehouses needs some type of support, for being able to make better use of the external data incorporated and for being able to fully exploit the potential of such incorporation.

5 The external data incorporation process

In order to become successful with the incorporation initiatives, the problems given in previous section must be dealt with. However, for being able to deal with them and to be able to suggest guidelines on how to resolve the problems, one must understand where in the process those problems are situated. This is utterly important since many of the problems (and their guidelines) are interlinked and may be considered as chains of problems with related guidelines. For example, if you are having problems with using the external data in the way that you want, it may be dependent on how it is integrate into the star-schema. However, if that data is impossible to integrate (may depend on several reasons e.g. format, data types, or meaning) in any other way than what is currently being made, you may have to consider to acquire some similar data from the same supplier, that doesnot cause these problems. If that suppliers is not possible to supply with any similar data, thatn you may have to look for another

supplier. In way, a problem that was related to usage has transferred you all the way via the incorporation process to the identification activity. Therefore, in this section the process of incorporating is firstly outlined and described. Secondly, the activities of the process are described in more detail and the problems introduced earlier are contextualized and related to the activities of the process.

5.1 The outline of the external data incorporation process

The result of the interview study showed that the incorporation of external data may be divided into the following four activities; 1) identification, 2) acquisition, 3) integration, and 4) usage (Fig. 1). From now on, this process will be referred to as the *external data incorporation process*. In Fig. 1, the activities are given in a strict sequential order, starting with identification However, it should be mentioned that the starting point of the process may differ, depending on which of the underlying activities that functions as the initiator of incorporating external data. For example, the process may become initiated from a usage perspective, in which a decision-maker wants to be able to perform a specific task, e.g. marketing campaign follow-up, and therefore starts in the other end of the sequence given by the subobjectives. The decision maker then starts at a usage perspective of external data incorporation and from that starts to identify possible sources or suppliers of relevant data. On the other hand, the DW administrator may find a certain data very difficult or time-consuming to integrate into the DW and therefore needs to acquire the data in another format (from the same supplier). Thereby, the process becomes initiated from an integration perspective.

The activities are not unique for external data. The activities may be considered as rather generic, since they are included in many methods related to information systems development in general (e.g. [19], [11]) and data warehouse development in particular (e.g. [8], [7], [5]). Still, in Fig.1, the process of including external data have deliberately been distinguished from the more common process of incorporating internal data, (e.g. [3], [1], by pinpointing the importance of a market. The concept market clearly shows that there is data available, but the concept also indicates that you most often will have to pay for it. The cost associated with external data is also high lightened by Kelly ([10], p.32), who states that: "*external data is captured outside of the enterprise and is, most often, made available at a cost by specialist information providers*." (See. Kimball's syndicate data suppliers).

To conclude the description of the external data incorporation process and its constituting activities, maintenance will be discussed. From the description of the process, it is clearly shown that maintenance is not considered as a separate activity. Instead, one may argue that maintenance concerns both acquisition and integration. How the external data is maintained, ones it has been integrated into the DW depend on several aspects. Two extremes will be presented as a start for the elaboration. Firstly, if the external data acquired is highly tailored to the needs of the consumer, the maintenance is rather straight forward from an integration perspective, since transformations or recalculations are unnecessary. Notice however that the example is not intended to trivialize on the problems of selecting maintenance policies, maintenance intervals, or identifying low-load time windows, but from an incorporation perspective, most of the maintenance efforts relates to the acquisition activity. Secondly, in such cases were the data is acquired from the Internet, without relevant metadata and staleness information, much effort will be spent on transforming the data into formats suitable for the DW environment. Under such conditions, the acquisition is rather straightforward, whereas the integration activity consumes a lot of resources. Since these two extremes shows that maintenance may be considered as belonging to one of the two activities, or mostly likely, a combination of both, it is well motivated to exclude maintenance as a separate activity.

5.2 The problems contextualized

In the following, the essence of these four activities will be described and the problems described in Sec. 4 are related to the appropriate activity.

Firstly, identification refers to the activity of finding and evaluating available sources to acquire external data from. Examples of such types of sources are given in Sec. 2. This activity is utmost important, since the consultants claimed that it is difficult to identify relevant sources and that the knowledge of available sources seemed to be

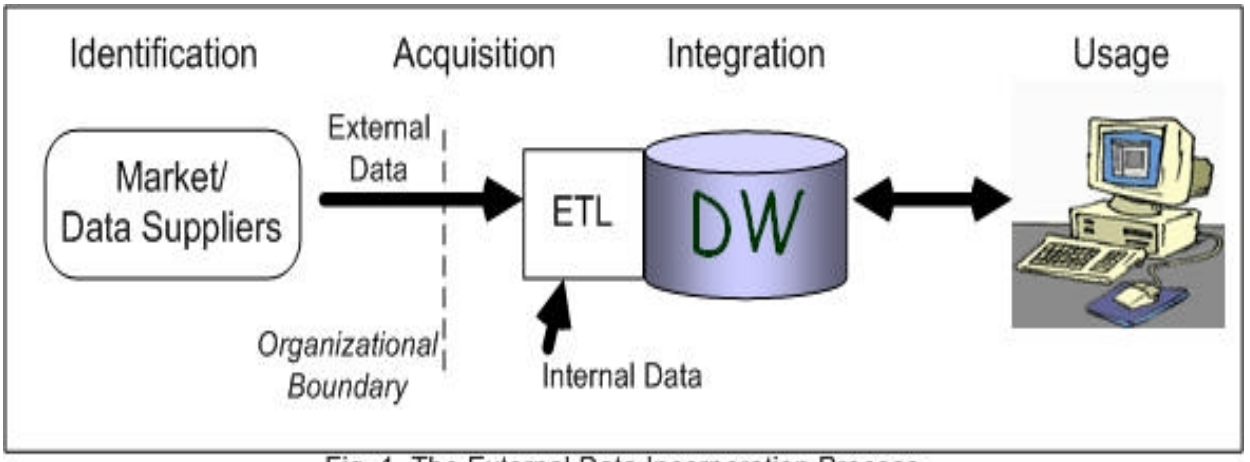

Fig. 1. The External Data Incorporation Process

spread on a mouth-to-mouth basis. If comparing with internal data identification, it is truly so that the external data sources are more difficult to identify, since they are not inside the organizations systems boundary.

Furthermore, the difficult to ensure the quality of the external data may partly be handled during the identification activity. Some of the consultants claimed that the external data suppliers are able to, with varying degrees, to tailor the data according to the needs of the organizations. By tailoring the external data so that it is structured and calculated according to the internal preferences, the data quality problem may partly be managed. Another data quality aspect to pay attention to during the identification activity is to choose data suppliers (if the data is acquired from syndicate data suppliers) with a good reputation, since if the supplier have a good reputation, that may increase the users willingness to include the external data as a basis for decision support.

Secondly, with acquisition it is referred to the process of acquiring the data and distributing it to the internal systems. Relevant aspects are approaches for retrieving the data, different distribution channels, and subscription policies. The interviews indicate a strong emphasis of applying web technology to make the data available and to distribute it. The respondents also brought up two distinct perspectives on the application of the Internet, which ought to be mentioned: 1) The Internet may be considered from a data source perspective, in which the data is extracted from different web pages and to which there are no contractual relationships to there owners/creators. Thereby, all the information stored on the Internet is considered as possible external data. 2) Many of the external data suppliers use the Internet and its underlying technology to present and distribute the data to their customers (web-hotel). In this sense, the Internet may be considered from a data container perspective, in which the data is made available to established customers for downloading or distribution, according to contractual legislations. Other alternatives would have been to send the data to the customers on other storage media, such DVD-ROM or magnetic tapes. Obviously, there are hybrids combining the characteristics of both these extremes. For example, there are a lot of organizations which publish different statistics on web pages, without any commercial undertones, which may be acquired and downloaded by anyone interested. Moreover, the acquisition activity may be alleviating the problem of integrating the external data with the internal data, since if the data are acquired in a proper manner, the integration may be rather straightforward, especially if the external data are strongly tailored according to the needs of the organizations. Kimball [14] reflects upon the problem of acquiring external data from a target problem. He claims that it is important to acquire the data directly to the DW environment, since there are organizations that have bought external data for other purposes than DW incorporation and when they understand that the data may also be contributory in the decision making environment, massive manual, and thereby costly, transformations are needed before the data may be integrated into the DW system.

Thirdly, the integration of external data involves e.g. aspects on how the external data is integrated into the data warehouse, how it is modelled in the underlying schemas, how the quality is assured, and how the external data is stored. As previously mentioned, most of the consultants claimed that many organizations are having problems with the actual integration of the external data acquired into the DW. First and foremost, one must understand that the approach for integrating the external data with the internal data is *purpose-driven*, i.e. depending on how you plan to use the external data, you must integrate it accordingly. In literature, one may only find brief examples on different approaches for how to integrate the external data. For example, Damato [5] introduces three general approaches on how to model the external data and which may be used as guidance when trying to solve the physical data integration problem: 1) The external data may be stored into separate dimensions, if the data is not directly associated to any of the internal data dimension. Thereby, the external data becomes integrated into the DW and its underlying e.g. star-schemas, without being mixed with the own internal data. This approach may be beneficial if the users are suspicious to the external data or if it is unclear to what degree the external data is of a high or accepted quality. Naturally, it may also be the case that the external data is of such a type that is should be in a separate dimension. For example the customer stock of a business partner may be interesting to acquire and such data may be stored as a separate dimension, allowing cross-analysis between the own customer data and the customer data of the business partner. 2) The data may be integrated into existing dimensions as added attributes. Damato [5] exemplifies on gross sales and credit ratings of customers. For example, if you acquire liquidity data about companies and wants to add the liquidity of those that are customers to you, it seems rather intuitive to add such an attribute in the customer dimension of the star-schema 3) The external data may be integrated as cross dimensional attributes, meaning that the data is related to the intersection of two dimensions. Damato [5] claims that weather is a perfect example of such cross dimensional attribute, since it is related to both a time- and a geographical dimension.

Finally, the usage activity of the external data incorporation process includes e.g. how the data is interpreted, for what purposes it is used, and how the external data is conceptually mapped with the internal data. The consultants claimed that the problems associated with using external data were very difficult to solve, due to the human involvement. The usage of the external data may be very contributory, but it also needs that the issues mentioned in Sec. 4 are handled. The only solutions indicated by the consultants was to: 1) Be very careful when selecting the suppliers of the external data, since data acquired from well-known suppliers with a good reputation are more likely to become trusted by the users, than data that is acquired from e.g. the Internet, since the quality of such data is difficult to assure. 2) Tailor the data so that it suites the purposes of the decision-makers, since

if they are unable to use the data for the intended purposes, it is very likely that they well develop a *not-invented-here* syndrome and thereby to use the external data in alignment of the purpose of its acquisition. In addition, if the data is not being use, it will never produce any return on investment. Than it will become axiomatic that incorporation initiatives fail. 3) Be very clear towards the intended users on the benefits of the data. Otherwise you may end up in the same problems described for the previous solution. 4) Initially use external data that is rather straightforward. If trying to immediately introduce complex ratings or metrics, which are hard (or even impossible) for the user to interpret and understand, the data will not become used or maybe even worse, critical decisions that are being made are based upon guesses of the meaning of the data. For example, if you introduce an attribute giving customer rating of the most profitable customers, if must be without any doubt that the rating are calculated correctly, since we do not want to end up in a situation in which we have abandoned the 5 percentage of our customers that generated the lowest profit or no profit at all and that it later shows that the ranking was partly calculated on misinterpreted external data or solely based on external data. This may seem like a rather hypothetic example, but there are external data suppliers that sell rankings of organizations and companies and which the consultants have had some severe problems to really understand or even penetrate and see what actual data that are used when calculating these rankings. One of the consultants even claimed that:" it was a hocus-pocus ranking, which the supplier could not explain the origin of" (translated from Swedish).

To conclude, the list of problems given in this paper is not to consider as complete. Still, already these rather few problems clearly show that incorporating external data into DWs is not a trivial undertaking and there are a lot of problems to be aware of and which one must be prepared to handle or solve. Naturally, some of the problems, especially those with human involvement, are very difficult to solve, but in this paper we suggest some solutions or guidelines that at least may be used to alert an organization that there are a lot of problems, and not only benefits (as opposed to most other related literature).

6 Discussions

Incorporating external data into DWs is not an easy task, since it involves a lot of different parameters. The phenomenon may be considered from a multitude of angels, such as contribution, process design, tool support, data quality and appropriateness. However, in this paper, a general view on the incorporation process was address, aimed at evaluating the incorporation process from a problem perspective. As indicated previously, current literature fails in giving a balanced view on external data, by only claiming its advantages. Therefore, the general outline was considered as most urgent to describe, as there is no idea to solve minor e.g. technological issues, if the overall process is so restricting and problematic, making it almost impossible to achieve any benefits at all.

The tendency of an opportunistic viewpoint on external data incorporation was also somewhat encountered during the interviews, as all consultants claimed that external data is important and that the incorporation must, and will, increase in the future. (A more detailed discussion related to this may be found in [16]). Still, the respondents gave a balanced view on the ins and outs of external data incorporation, by contributing with problems that needs to be handled. Therefore, the author would like to claim that the problems identified and contextualized are contributory and highlights some very important problems that need to be dealt with, before the incorporation of external data may start to really pay off.

Still, more work is to be done, in order to fully understand the problems and benefits of incorporating external data into DW. As indicated earlier, the process of incorporating external data is somewhat similar to the process of integrating data from solely internal sources. However, there are some issues that makes the externally oriented process different compared to the internally oriented process. For example, the sources are more difficult to identify, you have no control of the sources (e.g. you have limited, if any, documentation on the format of the data, which data types it is stored as, and associated metadata may be missing), and the users are often more sceptic towards the data originating from another organization.

Therefore, it is important to further study this multi-faceted concept. A similar study ought to be directed towards the investigation of external data incorporation among the consumers/users of external data. From the consumers' point of view, one may imagine that there are a lot of issues to be resolved, such as technical and conceptual integration of the external data with internal figures and numbers. Especially if considering that the consultants claimed that the organizations need some type of support, for being able to fully exploit the potential of the data being incorporated. In conjunction, the experiences of the suppliers of external data, such as syndicate data suppliers and semi-syndicate data suppliers, would also be beneficial to investigate and acquire, since their view on external data incorporation may bring new perspectives on important issues and challenges. In other words, contrasting the experiences of the consumers with the experiences of the data suppliers could be synergetic, since it is not unrealistic to believe that challenges experienced by the consumers may be opportunities for the suppliers, and vice versa. In addition, such evaluation, based on three different empirical perspectives on external data incorporation, would probably give a rather complete picture of where the standards are today and where the area is heading.

Acknowledgements

This work is founded by the Swedish Knowledge Foundation (Sv. KK-Stiftelsen). Thank you for your support.

References

[1] Agosta, L. (2000) "*The essential guide to data warehousing*", Prentice Hall PTR, New Jersey.

[2] Akinde, M. O., Böhlen, M. H., Johnson, T., Lakshmanan, L. V. S. and Srivastava, D. (2003) "Efficient OLAP query processing in distributed data warehouses", *Information Systems 28*, pp. 111-135, Elsevier.

[3] Anahory, S. and Murray, D. (1997) "*Data warehousing in the real world: a practical guide for building decision support systems*", Addison-Wesley Longman, Harlow.

[4] Connolly, T. and Begg, C. (2002) *"Database Systems: a Practical Approach to Design, Implementation and Management"*, 3rd Edition., Addison-Wesley Longman, Harlow.

[5] Damato, G. M. (1999) "*Strategic Information from External Sources – a Broader Picture of Business Reality for the Data Warehouse*", [online], http://www.dwway.com

[6] Devlin, B. (1997) "*Data warehouse: from architecture to implementation*", Addison Wesley Longman, Harlow.

[7] Hammer, K. (1997) "Migrating data from legacy systems", in *Building, using, and managing the data warehouse*, in Ramon Barquin & Herb Edelstein (Eds), Prentice Hall PTR, New Jersey, pp27-40.

[8] Hessinger, P. (1997) "A renaissance for information technology" in *Data warehouse practical advice from the experts*, Joyce Bischoff and Ted Alexander (Eds), Prentice Hall PTR, New Jersey, pp16-29.

[9] Inmon, W.H. (1996) "*Building the data warehouse*", 2nd Edition, John Wiley & Sons, New York.

[10] Kelly, S. (1996) "*Data Warehousing, the Route to Mass Customization – Updated and Expanded*", John Wiley & Sons, Chichester.

[11] Kruchten, P. (2000) "*The rational unified process - an introduction*", 2nd Edition, Addison-Wesley Longman, Reading.

[12] Kimball, R. (1996) "*The Data Warehouse Toolkit*", John Wiley & Sons, New York.

[13] Oglesby, W. E. (1999) "*Using External Data Sources and Warehouses to Enhance Your Direct Marketing Efforts*", [online], http://www.dmreview.com.

[14] Schubart, J. R. and Einbinder, J. S. (2000) "Evaluation of a data warehouse in an academic health sciences center", *International Journal of Medical Informatics* 60, pp319-333, Elsevier.

[15] Singh, H. (1998) "*Data warehousing: concepts, technologies, implementations, and management*", Prentice Hall PTR, New Jersey.

[16] Strand, M. and 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*, Olivier Camp, Joaquim Filipe, Slimane Hammoudi and Mario Piattini (Eds.), April 23-26, Angers, France, pp.570-573.

[17] Strand, M., Wangler, B. and 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)*, August 4-6, Tampa, Florida, USA.

[18] Theodoratos, D., Ligoudistianos, S. and Sellis, T. (2001) "View selection for designing the global data warehouse", *Data and Knowledge Engineering* 39, pp219-240, Elsevier.

[19] Weaver, P. L., Lamvrou, N. and Walkey, M. (1998) "*Practical SSADM – a complete tutorial guide*", 2nd Edition, Financial Time's Pitman Publishing, London.