The Work Process of Syndicate Data Suppliers – Actors Related and Common Problems Identified

Mattias Strand, Benkt Wangler, Björn Lundell

School of Humanities and Informatics, University of Skövde, Sweden
Box 408, SE-541 28 Skövde,
{mattias.strand,benkt.wangler,bjorn.lundell}@his.se
Fax: +46 500 44 83 99

Abstract. The research reported in this paper extends current knowledge of syndicate data incorporation into data warehouses (DWs). It does so by employing an interview study towards syndicate data suppliers (SDSs). The results describe, in detail, a generic work process, with related problems, for how the SDSs work with the data. In addition, the paper also relates the actors that regulate or collaborate with the SDSs, to the process activities. The results show that the SDSs encounter problems like transcending data errors, non-interoperating systems, and a non-tailoring attitude among their sources.

Keywords. Data Warehousing, Syndicate data, Syndicate data supplier

1 Introduction

In this paper we present the results from an interview study targeted towards syndicate data suppliers (SDSs), with a particular focus on data or services aimed at DW incorporation. For clarification, a DW is a: "*subject-oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions*" ([4], p.33). The paper gives three different results. Firstly, it describes a generic work process on how the SDSs work with their data identification, acquisition, integration and product development, and how they sell and distribute their data. Secondly, it outlines problems the SDSs encounter during this work process. Finally, the paper also relates important actors in the business environment of the SDSs to the process activities.

The results are important for several reasons. First of all, these details are important for understanding the supplier-consumer constellation of syndicate data incorporation, as the user organizations[1] acquire most of their external data from the SDSs [8]. Secondly, the results extend the current body of knowledge on how the SDSs work, who they collaborate with, and which problems they encounter. The extended body of knowledge is important, for being able to understand syndicate data incorporation into

[1] For the rest of this paper, *user organization* and *consumer* will be used interchangeably to denote an organization that buys syndicate data from one or several SDSs.

Proceedings of the CAiSE'05 Forum - O. Belo, J. Eder, J. Falcão e Cunha, O. Pastor (Eds.)

DWs and for being able to contextualize the support needed by the consumer organizations [9]. Current literature only addresses the existence of such organizations in very fragmented ways without going into any details. Kimball [6], Damato [1], and Oglesby [7] claim that these organizations deliver data for DW incorporation, but do not provide any detail on e.g. how they work and which problems they experience.

2 Background

Kimball [6] is the first to use the term *syndicate data* and *syndicate data supplier*. Unfortunately, he does not define term "syndicate data". Therefore, we have adopted the following definition: "*Business data (and its associated metadata) purchased from an organization specialized in collecting, compiling, and selling data, targeted towards the strategic and/or the tactical decision making processes of the incorporating organization" ([11], p.2).*

The consumers process of incorporating external data comprises the following four activities; 1) *identification*, 2) *acquisition*, 3) *integration*, and 4) *usage* [9]. From now on, this process will be referred to as the *external data incorporation process* (EDIP). Briefly, the activities are: *Identification* – the activity of finding and evaluating available sources to acquire external data from, *Acquisition* – the activity of acquiring the data into the own organization. *Integration* – the activity of integrating and storing the data into the DW. *Usage* – the activity of mapping and conceptually interpreting the data. (For a more detailed description of the process activities we refer to Strand and Wangler, [9]). The activities are not unique for syndicate or even external data. The same activities are found in general data warehouse development processes (e.g. [2], [3], and [1]). Still, the process of incorporating syndicate data differs from the process of incorporating internal data in two ways: 1) the data is acquired from outside the organization and 2) syndicate data is bought from special SDSs and therefore have a monetary cost associated with it [5].

3 Research Approach

The research approach taken in this work was to conduct in-depth interviews with a number of SDSs. The reason for conducting a qualitative rather than a quantitative study is that the field is rather unexplored and that we, hence, wanted to acquire deep knowledge from a fair amount of the rather few SDSs that exist in Sweden. The interview study covered 10 respondents working for 8 different SDSs. The questions were structured according to the four activities of the EDIP. Follow-up questions were asked in order to e.g. acquire illustrative examples, or if important details were left out. Hence, the interview study is to be considered as semi-structured [10]. The interviews were conducted via telephone, taped and transcribed. They lasted on average for 90 minutes. The transcripts ranged from 3560 to 7334 words (5356 words in average) and were later returned to the respondents for validation, allowing them to correct misunderstandings or complement their answers. When the transcripts had been validated, they were included in the analysis.

4 Analysis and Results

First of all, the analysis shows that the work process of the SDSs comprises the following four activities: *identifying novel data services*, *acquiring data from data sources*, *integrating data and develop services*, and *selling and delivering data*. As may be found, the SDSs work process is very alike the consumers EDIP, although the SDSs sell the data in their turn and therefore their process has a selling and delivering activity instead of the consumers usage activity. Below, the process activities of the SDSs' work process will be elaborated upon in detail.

Identifying new data sources
The analysis shows that the SDSs have three major influences for identifying data sources: their own search initiatives, influences from the user organizations, and data or services offered by their competitors. A majority of the SDSs claim that they are very mature in their search routines, making the amount of unexplored, available data very small. The lack of unexplored data sources was also considered as the main problem with respect to the identification of data sources. The SDSs further claimed that you very rarely find new data that may result in a new product or service and therefore they spend a lot of resources on trying to redevelop and refine already existing services or products.

Acquiring data from data sources
The SDSs have three main types of actors, from which they buy their data i.e. bi-product data suppliers, governmental agencies, and user organizations. In addition, the SDSs also buy data from other SDSs, both generic suppliers and suppliers in a monopoly situation. The analysis also shows that the sales of data between SDSs and data acquisition from user organizations and bi-product data suppliers is conducted without any problems, whereas a majority of the respondents claim that they encounter problems when buying or acquiring data from governmental agencies. In detail, the SDSs experience the following problems: 1) *Incompatible file types* – some agencies deliver data from legacy systems, via file types that are not supported by the SDSs' systems, 2) *Bridging systems problems* – some of the SDSs claim that it is often problematic to establish a communication with the agencies' systems, resulting in ad-hoc or home-made solutions, and 3) *Non-tailoring attitude* – some of the SDSs claim that the agencies are restricted on what data they deliver and they do not tailor the content of the data.

Integrating data and developing services
The analysis shows that all respondents account for high data quality awareness and that they claim that it is a prerequisite for being able to compete. Most SDSs also conducted tool-based data verification, e.g. verifying that every record identifier had a corresponding record or that the data sets are complete. A majority of the SDSs also conducted manual data verification, e.g. by contacting organizations if important data is missing, and verified e.g. delivery addresses or board members. Furthermore, a few SDSs also indicated that they procure external support, from data quality verifiers to verify the quality of the data. In addition, most respondents pointed to the importance of refining and adding value to the data they sell. Hence, SDSs constantly strive towards developing new services, based upon various refinements, which may

contribute to the customers. The analysis also shows two common approaches for developing these services. Firstly, data warehouse consultants identify new data or combinations of data that have not previously been exploited, but which may have a contribution to the user organizations. Based upon this new data or data combinations, they develop services which they try to sell to their customer. Secondly, the SDSs receive requests from user organizations, for data or services which they may not deliver. Based upon these needs, they try to develop services and extend them with further beneficial features or data, in order to increase the support offered to the user organizations.

Our analysis also shows that the SDSs encounter problems related to data integration and service development. Some of the respondents mentioned that they encounter data problems which originate from the governmental agencies, as these agencies have problems with controlling their data quality, and that they are rather varying in maturity and carefulness regarding how they conduct quality control, how reliable their systems are, and the coverage of their data. In detail, the SDSs indicated the following problems, caused by governmental agencies: 1) *Contradictory data* – the agencies may send data starting and terminating a business in the same delivery, making it impossible for the SDSs to know if the company is up and running or not. 2) I*ncomplete data sets* – identifiers or important data may be missing and is complemented in the next delivery. 3) *Outdated data* – the data delivered by the agency may be older than the last update on the SDSs side, making them change e.g. an updated address back to its old, outdated value. 4) *Inaccurate data* – the name of a city or a product may be misspelled in the data acquired from the agencies.

The SDSs also indicated the following problems, which are not considered as originating from the governmental agencies: 5) *Data quality verification of soft data* – the spelling of names is almost impossible to check in an automatic manner. Instead SDSs have to allocate a lot of resources to manually solve these data quality problems by contacting every single instance in the data set and verify the exact spelling of e.g. a surname. 6) *Legislation and regulation* – SDSs have to obey legislation and to be very careful when integrating data or when developing new data and services, so that they do not violate any laws or regulations.

Selling and delivering data

Most SDSs had internal resources for identifying prospects and selling data or services. However, a few SDSs also outsourced these initiatives to organizations specialized in marketing and sales. In addition, the analysis shows that the SDSs also collaborate with DW consultants for identifying prospect and establish business relations. If analyzing the collaboration with hardware vendors, two types of collaboration appear. Firstly, the SDSs and the hardware vendors collaborate in DW projects, in which the SDSs are taking an active part and populates the customers' DWs with combinations of internal and syndicate data. Secondly, many hardware vendors have informal collaborations with SDSs, suggesting a specific SDS for their customer organization. With respect to the software vendors, the analysis shows on a formal collaboration. A few respondents indicated that they cooperate, or planned to do so, with software vendors on special certificates. The underlying idea is that the SDSs and the software vendors agree upon representation of data and that they thereafter certify these representations, meaning that a user organization following the

certificate, i.e. to procure the software from the particular vendor and the data from the particular SDSs, does not have to transform the syndicate data being incorporated. Thereby, user organizations drastically reduce the resource they have to spend on data transformation and integration.

A majority of the SDSs apply a traditional approach to data delivery, i.e. they distribute the data to the user organization, which itself has to cater for the integration efforts. However, in order to lessen the data distribution and transformation problems for the user organizations, some of the SDSs have taken a different approach, in that they get the data from the user organization and integrate and refine it on the SDS side, before returning it back to the user organization.

Furthermore, the SDSs also sell their data and services via other SDSs, acting as retailers. Only a few problems arose in relation to this activity and they only have to do with data distribution: 1) *Undertaking too large projects* – a few respondents claimed that these projects tend to be rather large and are undertaken during a long time-period, which occasionally may have the effect that the user organizations do not have the strength to go through the whole project and therefore terminate it before completion. 2) *Non-interoperable systems* – a few respondents claimed that interoperability between the systems could cause problems, but also that these problems normally were solved rather easily.

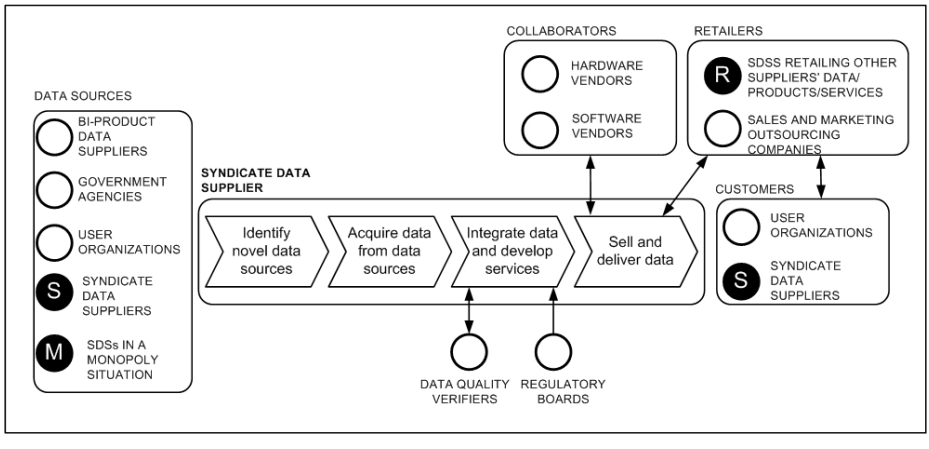

Fig. 1. A generic work process of syndicate data suppliers, comprising data identification, data acquisition, data integration and service development, and data sales and delivery.

In summary, Fig. 1 shows the different activities of a generic work process, describing how the SDSs work with the data. Arrows indicate the particular activities in which other actors in the domain influence the work of the SDSs. Bi-directional arrows show relationships which are initiated and may be terminated by the SDSs whereas the one-directional arrow shows an external influence which the data SDSs must follow and respond to.

5 Discussion and Future Work

Many of the problems encountered by the SDSs are also experienced by user organizations, e.g. contradictory data, acquiring out-dated data, and legislation and regulation [11]. Therefore, one may claim that the problems are generic for all types of organizations taking a user role in the value chain of syndicate data incorporation. However, to collect further evidence to verify how generic the problems are, it would also be interesting to conduct a study towards governmental agencies, as they share the duality of the SDSs, i.e. being both suppliers and consumers of syndicate data.

Finally, as indicated in the analysis, the SDSs anticipate that the incorporation of syndicate data into DWs will increase, as organizations are maturing and starting to understand the benefits of such incorporation. Still, user organizations experience many problems and therefore they need hands-on support for being able to fully exploit the potential of syndicate data incorporated into DWs [9].

References

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

[2] Hammer, K. (1997) "Migrating data from legacy systems", in *Building, using, and managing the data warehouse*, in Ramon Barquin & Herb Edelstein (Eds), New Jersey: Prentice Hall PTR, pp. 27-40.

[3] Hessinger, P. (1997) "A renaissance for information technology" in *Data warehouse practical advice from the experts*, Joyce Bischoff and Ted Alexander (Eds), New Jersey: Prentice Hall PTR, pp. 16-29.

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

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

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

[7] Oglesby, W. E. (1999) *Using external data sources and warehouses to enhance your direct marketing effort*. Available at Internet: http://www.dmreview.com/ [Accessed 03.02.21].

[8] 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.

[9] Strand, M. & Wangler, B. (2004) Incorporating external data into data warehouses - problems identified and contextualized. In *Proceedings of the 7th International conference on information fusion (Fusion'04)*, June 28- July 1, Stockholm, Sweden.

[10] Williamson, K. (2002) *Research methods for students, academics and professionals*. 2nd Edition, Thousand Oaks: Sage Publications, Inc.

[11] XXXXXX (the author(s) made anonymous) (2005) *Syndicate data incorporation into data warehouses: a categorization and verification of problems*. Submitted to an international conference.